Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 72–81
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Robust speech recognition by integrating speech separation and hypothesis testing

Soundararajan Srinivasan [a,*], DeLiang Wang [b,*]

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

Received 27 October 2008; received in revised form 24 August 2009; accepted 24 August 2009

Abstract

Missing-data methods attempt to improve robust speech recognition by distinguishing between reliable and unreliable data in the time–frequency (*T–F*) domain. Such methods require a binary mask to label speech-dominant *T–F* regions of a noisy speech signal as reliable and the rest as unreliable. Current methods for computing the mask are based mainly on bottom-up cues such as harmonicity and produce labeling errors that degrade recognition performance. In this paper, we propose a two-stage recognition system that combines bottom-up and top-down cues in order to simultaneously improve both mask estimation and recognition accuracy. First, an *n*-best lattice consistent with a speech separation mask is generated. The lattice is then re-scored by expanding the mask using a model-based hypothesis test to determine the reliability of individual *T–F* units. Systematic evaluations of the proposed system show significant improvement in recognition performance compared to that using speech separation alone.

Keywords: Robust speech recognition; Missing-data recognizer; Ideal binary mask; Speech segregation; Top-down processing

1. Introduction

The performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of noise and other distortions (Gong, 1995; Huang et al., 2001). Speech recognizers are typically trained on clean speech and face a problem of mismatch when used in conditions where speech occurs simultaneously with other sound sources. To mitigate the effect of this mismatch on recognition, noisy speech is typically preprocessed by speech enhancement algorithms (Loizou, 2007), such as spectral subtraction based systems (Boll, 1979; Droppo et al., 2002). If samples of the corrupting noise source are available *a priori*, a model for the noise source can additionally be trained and noisy speech may be jointly decoded using the trained models of speech and noise (Varga et al., 1990; Gales and Young, 2007) or enhanced using linear filtering methods (Ephraim, 1992). However, in many realistic applications, the performance of the above approaches to robust speech recognition is inadequate (Cooke et al., 2001).

To deal with the mismatch issue, a missing-data approach to robust speech recognition has been proposed by Cooke et al. (2001). This method distinguishes between reliable and unreliable data in the time–frequency (*T–F*) domain. When speech is contaminated by additive noise, some *T–F* regions will contain predominantly speech energy (reliable) and the rest are dominated by noise energy. The missing-data ASR treats the latter *T–F* units as missing or unreliable during recognition. The missing-data recognizer, therefore, requires a binary *T–F* mask that provides information about which *T–F* units are reliable and which are unreliable. Previous studies have shown that the missing-data recognizer performs very well

* Corresponding authors. Present address: Robert Bosch LLC, Research and Technology Center North America, Pittsburgh, PA 15212, USA (S. Srinivasan). Tel.: +1 412 325 8452 (S. Srinivasan), +1 614 292 6827 (D.L. Wang).
E-mail addresses: srinivasan.36@osu.edu, soundar.srinivasan@us.bosch.com (S. Srinivasan), dwang@cse.ohio-state.edu (D. Wang).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.08.008

when this mask is known *a priori* (Cooke et al., 2001; Roman et al., 2003; Barker et al., 2005; Srinivasan et al., 2006). Attempts to estimate such a binary mask through front-end preprocessing using speech separation techniques have been only partly successful. Spectral subtraction is frequently used to generate such binary masks in missing-data studies (Drygajlo and El-Maliki, 1998; Cooke et al., 2001). For this purpose, noise is usually assumed to be long-term stationary and its spectrum is estimated from frames that do not contain speech (speech silent frames containing just background noise). The noise spectrum is then used to estimate the SNR in each *T–F* unit. If the SNR in a *T–F* unit exceeds a threshold, it is labeled reliable; it is labeled unreliable otherwise. In the presence of non-stationary interference sources, however, the use of spectral subtraction results in a poor estimate of the mask. Methods that primarily utilize the harmonicity of voiced speech have also been proposed to estimate the mask for missing-data applications (Seltzer et al., 2000; Brown et al., 2001; van Hamme, 2004). However, these methods are unable to deal with unvoiced speech. Accurate estimation of pitch is also difficult, if not impossible, when the SNR is low. Hence, the estimated binary mask corresponding to voiced speech may not be reliable. Therefore, good estimation of the binary *T–F* mask remains a challenging problem.

On the other hand, the human auditory system exhibits a remarkable ability to segregate a target speech source from various interferences (Darwin, 2008). According to Bregman (Bregman, 1990), this is accomplished via a process termed auditory scene analysis (ASA). ASA involves two types of organization, primitive and schema-driven. Primitive ASA is based on bottom-up cues such as pitch and spatial location of a sound source. Schema-based ASA is based on top-down use of stored knowledge about auditory inputs, e.g. speech patterns, and supplements primitive analysis. Top-down information has also been used successfully in computational ASA studies previously (Barker et al., 2005; Srinivasan and Wang, 2005b). In particular, Barker et al. (2005) have proposed a top-down approach to identify *T–F* units that are dominated by speech in a noisy mixture. We believe that a top-down approach, using speech models, can be used to refine the mask generated by bottom-up processing to achieve improved recognition results.

In this paper, we present a two-pass missing-data recognition system that estimates an ideal binary *T–F* mask and improves recognition results at the same time. A *T–F* unit in the ideal binary mask is 1 if in the corresponding *T–F* unit the noisy speech contains more speech energy than interference energy; it is 0 otherwise. The ideal binary mask is obtained a priori from premixed speech and noise. In the first pass, a mask produced by a speech separation system is used to generate an *n*-best lattice using a missing-data recognizer. This corresponds to bottom-up processing. This lattice is then re-scored, to produce the final recognition results by augmenting the initial mask using the information contained in states along individual paths. Specifically, we propose a state-based hypothesis test to determine the reliability of each *T–F* unit. This corresponds to top-down analysis. The resulting recognition accuracy is substantially better than that of the conventional ASR as well as the missing-data recognizer using the mask produced by speech separation alone.

The rest of the paper is organized as follows. The next section contains a detailed presentation of the system. The proposed system has been systematically evaluated on a noisy connected digit recognition task and the evaluation results are presented in Section 3. Section 4 concludes the paper.

2. System description

The proposed system is a two-pass recognition system as shown in Fig. 1. In the first pass, we use an initial, conservative mask generated through bottom-up separation as input to a missing-data recognizer. The output of the missing-data ASR is a lattice containing *n*-best hypotheses. The initial mask is then augmented by another mask generated through spectral subtraction to result in a three-way mask. In the second pass, we use a state-based hypothesis test to refine this three-way mask and improve recognition results at the same time.

2.1. Bottom-up speech separation

The input to the system is a mixture of speech and interference, sampled at 20 kHz. Following the original study of Cooke et al. (2001), we use an auditory filterbank decomposition (Patterson et al., 1988) of the input signal to generate feature vectors for recognition. Specifically, the input is first analyzed using a 128 channel gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 80 Hz to 5 kHz (see (Wang and Brown, 2006) Chapter 1). Our previous studies (Srinivasan, 2006) have shown that this frequency range is adequate for recognition of male speech considered in this study (see Section 3). The instantaneous Hilbert envelope at the output of each gammatone filter is then downsampled to a frame rate of 100 Hz and finally cube-root compressed (Cooke et al., 2001). As a result, the input signal is decomposed into a two-dimensional matrix of *T–F* units.

The missing-data recognizer (Cooke et al., 2001) makes use of spectro-temporal redundancy in speech to recognize a noisy signal based on its speech-dominant *T–F* units. Specifically, it modifies the computation of the observation probability in a state of an HMM-based ASR to handle missing or unreliable data. The observation density in a conventional ASR is typically modeled using a mixture of Gaussians as shown below:

$$p(x|q) = \sum_{k=1}^{M} p(k|q)p(x|k,q), \tag{1}$$