Speech Communication 92 (2017) 142–151

Contents lists available at ScienceDirect

Speech Communication

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

A data-driven speech enhancement method based on A* longest segment searching technique

Yue Hao[a], Feng Bao[b], Changchun Bao[a,*]

[a] Speech and Audio Signal Processing Lab, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[b] Department of Electrical and Computer Engineering, The University of Auckland, Auckland 1010, New Zealand

ARTICLE INFO

Article history:
Received 21 September 2016
Revised 27 March 2017
Accepted 14 June 2017
Available online 23 June 2017

Keywords:
Speech enhancement
LRTDs
GMM
ALMSS
A* search technique
VTS
Modified Wiener filter

ABSTRACT

This paper proposed a data-driven speech enhancement method based on the modeled long-range temporal dynamics (LRTDs). First, by extracting the Mel-Frequency Cepstral coefficient (MFCC) features from speech and noise corpora, the Gaussian Mixture Models (GMMs) of the speech and noise were trained respectively based on the expectation-maximization (EM) algorithm. Then, the LRTDs were obtained from the GMM models. Next, based on the LRTDs, a modified maximum a posterior (MAP) based adaptive longest matching segment searching (ALMSS) method derived from A* search technique was combined with the Vector Taylor Series (VTS) approximation algorithm in order to search the longest matching speech and noise segments (LMSNS) from speech and noise corpora. Finally, using the obtained LMSNS, the estimation of speech spectrum was achieved. Furthermore, a modified Wiener filter was constructed to further eliminate residual noise. The objective and subjective test results show that the proposed method outperforms the reference methods.

1. Introduction

Enhancing speech degraded by the non-stationary noise is both an important and difficult task. The importance arises from several signal processing applications, including mobile communication, speech recognition, and hearing aids. The difficulty arises from the characteristic of the noise which is often non-stationary and speech-like. The goal of speech enhancement is to remove noise as much as possible from noisy speech for improving speech quality and intelligibility.

Currently, single-channel speech enhancement methods typically consist of two classes: unsupervised methods and supervised methods. For the unsupervised techniques, such as Wiener filter (WF) method (Lim and Oppenheim, 1979), spectral subtraction (SS) method (Boll, 1979), minimum mean-square error (MMSE) method (Ephraim and Malah, 1984), weighted Euclidean distortion measure (WEDM) method (Loizou, 2005), a common problem is that there is always a trade-off between noise suppression and speech distortion. The main reason is that they don't have specific assumption about the speech except for the assumption about independence and probability distribution of noise. By contrast, the supervised methods, such as codebook-based (CB) method (Srinivasan et al., 2007; He et al., 2017) and non-negative matrix factorization (NMF) method (Lee and Seung, 1999), by using a linear combination of multiple prototypes for the clustered features, can achieve remarkable performance in noise suppression and speech enhancement, since they can provide more priori information about speech and noise. However, these methods failed to capture the inter-frame dependency of speech signal. In recent years, researches have emphasized on cross-frame importance for improving speech quality in highly non-stationary noise conditions, such as hidden Markov model (HMM) methods (Zhao and Kleijn, 2007; Gao et al., 2014). But, their state dynamics are unrealistic for representing temporal dynamics of speech in a long-range period under the first-order Markov chain assumption. Moreover, these approaches provide specific characteristics of speech and noise, in which the priori information of the underlying speech is often the clustered features for noise removal and speech reconstruction. Thus, they often suffer poor performance under non-stationary noise environment because of their weak predictability for the fast-varying noise.

Long-range temporal dynamics and speaker characteristics are of important features for distinguishing a speech utterance from non-stationary noise environment. As a result, the HMM-based method was further developed by using a data-driven method given in Xiao and Nickel (2010), which is built on the complete training data from real speech corpus. The data-driven method provided a novel way to represent the priori information of speech to be estimated, in which a corrupted signal is reconstructed as a new "clean" signal from a large speech corpus. Although the data-driven method improved the desirable performance in producing better quality and natural sounding output, it could not

* Corresponding author.
E-mail address: baochch@bjut.edu.cn (C. Bao).

http://dx.doi.org/10.1016/j.specom.2017.06.004
0167-6393/

Fig. 1. MFCC feature extraction.

explicitly model the long-range temporal dynamics (LRTDs) of the target speech. This implies that it will be difficult to separate speech from noisy speech in a short-term period, due to the non-stationary features of speech and noise.

In this paper, we address the problem of recovering speech from non-stationary noise environment imposing priori or constraint on the speech, which is achieved by using clean speech utterance taken from large corpora of the un-clustered features of speech and noise to represent LRTDs. By maximally extracting the LRTDs and speaker characteristics, we will describe a method aiming to estimate the underlying speech accurately.

We try to extend the data-driven method used in speech segmentation and recognition (Ming, 2009) and extract LRTDs of speech and noise signals (i.e. GMMs, maximum Gaussian time sequence: MGTS) as much as possible for improving the performance of speech enhancement. The main contribution of this paper is to improve LRTDs modeling of speech and noise, and applies it to speech enhancement. The LRTDs play an important role in accurately separating out the clean speech and tracking fast-varying noise from noisy speech.

The data-driven framework (Ming, 2009) has been applied to speech enhancement in Ming and Crookes (2011), which can be outlined as follows: Firstly, a multi-condition parallel speech corpus and MGTS of speech are modeled to represent the LRTDs of speech. Then, given the noisy speech segment, the longest matching speech segment is found based on MAP criterion. Next, by concatenating the corresponding speech segment, the spectrum of the estimated clean speech can be constructed. Finally, the enhanced speech can be obtained by constructing a wiener filter. In fact, this method can perform well under non-stationary noise environment. By comparing with the method given in Ming and Crookes (2011), there are several differences in this paper: 1) We introduced the noise modeling for speech estimation instead of using a combination of multi-condition speech modeling, which fully captures the LRTDs of noise as well; 2) A fast searching method was proposed to accelerate the searching procedure of the mixed MGTS of speech and noise, which decreases the complexity effectively; 3) We utilized the VTS expansion to adapt the statistics of clean speech for the calculation of the likelihood of the mixed MGTS of speech and noise; 4) The MAP-based searching method was modified by using A* search technique used in speech recognition to get a better estimation of speech, which is called as ALMSS method; 5) A modified wiener filter was introduced to further eliminate background noise, especially those noise existed in unvoiced and silence segment.

The preliminary work and rough description of this paper was first published in Interspeech 2015 (Hao et al., 2015). In this paper we additionally proposed the aforementioned ALMSS method into Hao et al. (2015). Besides, we give more detailed description to Hao et al. (2015), and re-train the algorithm with ALMSS. Moreover, the objective tests used in Hao et al. (2015) are evaluated again including a state of the art method. Meanwhile, a subjective test, namely, the MUSHRA listening test is supplemented. All of this makes our method more credible and reliable.

The remainder of this paper is organized as follows. Section 2 presents an overview of the data-driven framework proposed in Ming (2009). The proposed speech enhancement method based on data-driven framework is described in Section 3. The performance evaluation results are shown in Sections 4 and 5 gives the conclusions.

2. Overview of data-driven framework

In this section, the extraction of Mel-Frequency Cepstral coefficient (MFCC) feature is firstly described. Then, we will present the LRTDs modeling of speech used for speech enhancement. Lastly, the MAP-based searching method for the longest speech segment is introduced. Meanwhile, its drawbacks are analyzed.

2.1. MFCC feature extraction

The feature extraction of speaker is very important for the data-driven speech enhancement. Since MFCC (Davis and Mermelstein, 1980) is commonly used in modern speech recognition system, we choose it as the parameters to be modeled. The extraction procedure is illustrated in Fig. 1.

Firstly, the input speech is divided into the frames by windowing. Then, the power spectrum of speech is estimated by doing a FFT for each frame. By applying the Mel-bank frequencies and summing the energy in each Mel-bank, the logarithmic energy of all Mel-banks is given by

$$\mathbf{s}_t^{mfs} = \log(MEL(|FFT(win(\mathbf{s}_t))|)) \tag{1}$$

where $\mathbf{s}_t$ and $\mathbf{s}_t^{mfs}$ are defined as a input vector and a Mel-frequency spectral (MFS) feature vector of clean speech in the tth frame, respectively, and the lth element of $\mathbf{s}_t^{mfs}$ is $s_t^{mfs}(l), l = 0, 1, 2, \cdots, 41$. The symbols, *win*, *DFT*, $|\cdot|$, *MEL* and log denote the windowing, discrete Fourier transform, magnitude calculation, Mel-filtering, logarithmic operation, respectively. Finally, the MFCC is given as follow by taking the DCT with respect to vector $\mathbf{s}_t^{mfs}$,

$$x_i = \sum_{l=0}^{M-1} s_t^{mfs}(l) \cos\left(\frac{\pi i\left(l+\frac{1}{2}\right)}{M}\right), \quad 0 \le i \le M \tag{2}$$

In this paper, we truncate the number of DCT coefficients is to 42, i.e., $M = 42$ in (2).

2.2. LRTDs modeling

In contrast to most conventional modeling methods for speech enhancement, we directly use data-driven modeling method (Ming, 2009) to provide the LRTDs based on the pre-recorded clean speech sentences spoken by different speakers. This model is good at representing LRTDs with any segment and any length in the training corpus. Because longer speech segment has more distinct dynamics and richer speaker characteristics, it can be considered as a unit that will be identified, which has lower error rather than shorter segment. Therefore, the estimation based on the longest identified segment could increase the de-noising performance.

For each test sentence, we try to identify the longest matching segment from the trained LRTDs represented by GMM, MGTS and spectral dictionary for speech enhancement. Here, LRTDs are modeled in each training utterance. Let $\{\mathbf{x}_i: i = 1,2,\ldots,I\}$ be a complete MFCC feature sequence derived from all utterances, where I is the number of speech frames for all utterances, and $\mathbf{x}_i$ is the MFCC feature vector at frame i. We take three steps to build a model for all training utterances.

In the first step, by constructing matrix $\mathbf{X}$ from I MFCC feature vectors aforementioned, a GMM probability distribution function