Speech Communication 94 (2017) 30–41

Contents lists available at ScienceDirect

Speech Communication

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

Spectral-domain speech enhancement for speech recognition

Chang Huai YOU*, Bin MA

*Institute for Infocomm Research, A*STAR, Singapore*

ARTICLE INFO

Article history:
Received 1 March 2017
Revised 11 July 2017
Accepted 24 August 2017
Available online 26 August 2017

Keywords:
Speech enhancement
Speech recognition
A priori SNR

ABSTRACT

Speech recognition performance deteriorates in face of unknown noise. Speech enhancement offers a solution by reducing the noise in speech at runtime. However, it also introduces artificial distortion to the speech signal. In this paper, we aim at reducing the artifacts that have adverse effects on speech recognition. With this motivation, we propose a modification scheme including a smoothing adaptation to frame signal-to-noise ratio (SNR) and a reestimation of *a priori* SNR for spectral-domain speech enhancement. The experiment shows that the proposed scheme of enhancement significantly improves the performance of the state-of-the-art speech recognition over the baseline speech enhancement techniques.

1. Introduction

State-of-the-art automatic speech recognition (ASR) system works well under clean environmental situation. However, it has been observed that the performance of the speech recognition system degrades rapidly in the presence of noise or other distortions (Li et al., 2014). Over the past few decades much research has been devoted to improving the robustness of speech recognition in noisy environment (Guan et al., 1993; Breithaupt and Martin, 2006; O'Shaughnessy, 2008).

The presence of noise at runtime introduces a mismatch between the training condition and test condition.

In practice, one of the solutions is the multi-conditional modeling which trains the acoustic model with various noisy databases to cover different kinds of noise environment. Multi-conditional training has proven the great advanced for noisy situation. It is a straightforward way to achieving noise robustness, but it is also known to suffer from a lack of generalisability to unseen conditions and a reduced performance of recognition on high-SNR speech. In other words, such technique fails in face of unknown noise condition. In particular, the multi-condition cannot improve the recognition accuracy while the trained model encounters unseen noise. An alternative to overcome unknown noise condition is to train the acoustic models on clean speech data and apply speech enhancement techniques to improve the runtime speech quality under noise condition (Flynn and Jones, 2008; Tchorz and Kollmeier, 1999). Moreover, the clean speech model always gives better performance to clean speech recognition than multi-conditional model does. With the speech enhancement solution, one can focus on developing a high quality clean acoustic model, a sharper model than a multi-condition acoustic model.

To understand the artifacts introduced by speech enhancement, and their effects on speech recognition system, we are interested in looking into various speech enhancement methods. In practice, it is always difficult to reduce noise without introducing the speech distortion due to the random nature of noise and the inherent complexity of speech signal. It has been a fact that the artifacts will be introduced into the speech signal as the noise is reduced in the speech signal. Thus, it is necessary to consider the tradeoff between noise reduction and speech distortion in speech enhancement (Lu et al., 2013b).

Among the most effective enhancement techniques in the past decades, the popular ones include spectral-domain denoising (McAulay and Malpass, 1980a; Stahl et al., 2000; Ephraim and Malah, 1984, 1985; You et al., 2005, 2006), speech production modeling (Gannot et al., 1998; You et al., 2004a), human auditory perceptual criterion (Hellman, 1972; Johnston, 1988; Tsoukalas et al., 1993), the probability of speech presence uncertainty (Cohen, 2002; Fodor and Fingscheidt, 2012), subspace decomposition (Ephraim and Van, 1995), and the combinations of the above techniques (You et al., 2008).

ASR speech enhancement aims to improve the quality of noisy speech input at runtime to reduce the mismatch with the trained acoustic model. In 1991, Hanson and Clements introduced a constrained iterative enhancement for speech recognition (Hansen and Clements, 1991), where an iterative Wiener filtering with vocal tract spectral constraints was formulated using interframe and intraframe constraints based on line spectral pair transformation. The enhancement approach with interframe constraints ensures more speech-like formant trajectories than those found in the uncon-

* Corresponding author.
E-mail addresses: echyou@i2r.a-star.edu.sg (C.H. YOU), mabin@i2r.a-star.edu.sg (B. MA).

http://dx.doi.org/10.1016/j.specom.2017.08.007
0167-6393/

strained approach while the intraframe constraints ensure overall maximization of the speech quality across all classes of speech. The performance was evaluated using a standard, isolated-word recognition system. In 2006, Gemello et al.proposed a modification of Ephraim-Malah log-spectral amplitude method by introducing an overestimation of noise power and an adjustment of spectral floor into *a priori* SNR and *a posteriori* SNR with respect to frame SNR (Gemello et al., 2006). Significant improvement was reported for Aurora speech recognition system. In 2008, Breithaupt et al.proposed a cepstral-domain smoothing method for estimation of *a priori* SNR (Breithaupt et al., 2008), and the experiment that was done with Wiener filter shows improvement over conventional decision-directed approach. However, the effectiveness of the *a priori* SNR estimation method was only proven in terms of speech enhancement objective measurement but not proven in terms of speech recognition performance. In the same year, Yu et al. applied the Ephraim-Malah minimum mean square error (MMSE) criterion into speech feature domain (Yu et al., 2008) instead of the discrete Fourier transform (DFT) domain for noisy speech recognition. The performance was investigated on the standard Aurora speech recognition platform (Hirsch and Pearce, 2000). In 2010, Paliwal et al.investigated the role of speech enhancement in speech recognition (Paliwal et al., 2010) where the experiments were conducted on the TIMIT speech corpus, however, there was no any solution provided for the artificial distortion caused by the investigated speech estimators against the speech recognition; and also the speech recognition decoder was only based on small Gaussian mixture model-hidden Markov model (GMM-HMM) where only eight-Gaussian mixtures per state were applied and a bigram language model was used. All the above studies could not make a very clear impression on the goodness and effectiveness of the various enhancement methods in modern speech recognition system, since we observed a fact that the performance of speech enhancer also depends on the particular speech recognition models. In other words, the enhancer may be helpful for certain speech decoder but not always contribute to another speech decoder. For this reason, it is necessary to investigate the performance with typical state-of-the-art decoding platform.

In this paper, we study the spectral-domain speech enhancement and select three typical methods, including Wiener filtering (Stahl et al., 2000), log-spectral amplitude (LSA) (Ephraim and Malah, 1985), and masking-based β-order (β-masking) MMSE (You et al., 2006) algorithms. In (Paliwal et al., 2010), Paliwal et al. investigated sixteen speech enhancement methods for speech recognition, and gave a conclusion that the improvements in objective speech quality did not translate to the improvement of speech recognition; and an enhancer (with its default settings) that produced best objective speech quality gave a poor performance in speech recognition. Therefore, a speech enhancement algorithm may significantly improve human listening experience (Lim and Oppenheim, 1979; McAulay and Malpass, 1980b), direct application of the enhancement algorithm does not always work well for speech recognition system. Classical objective quality measure based on global SNR or average segmental SNR over an utterance does not, in general, provide useful estimates of the perceived speech quality as well as the quality of machine recognition. In our observation, we also noticed that the speech enhancer that has good PESQ (perceptual evaluation of speech quality) performance for human listening brings poor word-error-ratio (WER) performance of speech recognition. The reason is the improvement of PESQ cannot be directly transferred into the improvement of feature distortion that directly affects the performance of speech recognition. So far, there is no any single statistics of the speech quality measure can completely transfer the similarity between the estimated speech and the reference (or clean) speech into the ultimate performance of machine recognition. However, the distortion measure of the feature sequence which is directly used as the input of modern speech recognition system can still represent a rough estimation of the distinction between the estimated speech quality and the clean speech quality for the speech recognition. We propose to improve the ASR speech enhancement systems by alleviating the feature distortion ratio for the purpose of the speech recognition in some aspects: the noise overestimation control, weak spectral component flooring, oversuppression of unwanted residual noise, and a reestimation of *a priori* SNR (You et al., 2017). Firstly, by introducing smoothing adaptation with respect to frame SNR, we design a smoothing control of the power of the processing noise, show a way to process the weak spectral signal with a time-varying floor of spectral SNRs. Secondly, we develop an oversuppression of the residual noise with smoothing adaptation. Finally, we propose a reestimation of the *a priori* SNR and extend it to a possible iterative process. Experimental results show each and every of the modifications (i.e. the noise control, the weak spectral processing, the residual noise suppression and *a priori* SNR reestimation) are able to effectively improve the performance of the three typical speech enhancement systems in terms of WER.

In order to build up a meaningful investigation system, we setup a state-of-the-art evaluation platform which is reconstructible by open-source speech recognition tool. In particular, we use Kaldi toolkit (Povey et al., 2011) to build up a large vocabulary speech recognition system with a series of the training models that start from monophone, coarse triphone GMM-HMM to detailed triphone GMM-HMM, and then DNN-HMM which follows the pre-training of deep belief network (DBN). In this speech recognition system, cepstral mean and variance normalization (CMVN), linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), feature space maximum likelihood linear regression (fMLLR) for speaker adaptive training and state-level minimum Bayes risk (sMBR) techniques are applied. We train the models in each steps by using the labelled clean speech, and measure the performance of recognition using clean, noisy and enhanced speech.

In the remainder of the paper, we give a brief introduction of the spectral-domain speech enhancement algorithms used in this paper in Section 2. In Section 3, we propose a series of modification schemes for the speech estimators against speech recognition. We describe the speech recognition platform for the performance evaluation of speech enhancement in Section 4. The evaluation is shown in Section 5 and finally the conclusion is given in Section 6.

2. Spectral-domain speech enhancement algorithms

An observed noisy speech signal $x(t)$ is assumed to be a clean speech signal $s(t)$ degraded by uncorrelated additive noise $n(t)$, i.e.,

$$x(t) = s(t) + n(t), \quad 0 \le t \le T. \tag{1}$$

Let $S_k(l)$, $N_k(l)$, and $X_k(l)$ denote the kth spectral component of the clean speech signal $s(t)$, noise $n(t)$, and the observed noisy speech $x(t)$, respectively, where l denotes the time frame corresponding to time t in analysis interval [0, T]. The enhanced speech spectrum is given by $\hat{S}_k(l) = G_k(l)X_k(l)$, where $G_k(l)$ is the gain function of the enhancement.

2.1. Wiener filtering

With Gaussian distribution assumption of the respective complex spectra of speech and noise, we seek to minimize Bayes risk with the expectation of cost function $C(\hat{S}_k(l), S_k(l))$ given observed