Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 1–10

www.elsevier.com/locate/specom

Noise perturbation for supervised speech separation

Jitong Chen[a,*], Yuxuan Wang[a], DeLiang Wang[a,b]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, United States
[b] Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, United States

Received 23 June 2015; received in revised form 1 December 2015; accepted 29 December 2015
Available online 6 January 2016

Abstract

Speech separation can be treated as a mask estimation problem, where interference-dominant portions are masked in a time-frequency representation of noisy speech. In supervised speech separation, a classifier is typically trained on a mixture set of speech and noise. It is important to efficiently utilize limited training data to make the classifier generalize well. When target speech is severely interfered by a nonstationary noise, a classifier tends to mistake noise patterns for speech patterns. Expansion of a noise through proper perturbation during training helps to expose the classifier to a broader variety of noisy conditions, and hence may lead to better separation performance. This study examines three noise perturbations on supervised speech separation: noise rate, vocal tract length, and frequency perturbation at low signal-to-noise ratios (SNRs). The speech separation performance is evaluated in terms of classification accuracy, hit minus false-alarm rate and short-time objective intelligibility (STOI). The experimental results show that frequency perturbation is the best among the three perturbations in terms of speech separation. In particular, the results show that frequency perturbation is effective in reducing the error of misclassifying a noise pattern as a speech pattern.

Keywords: Speech separation; Supervised learning; Noise perturbation.

1. Introduction

Speech separation is a task of separating target speech from noise interference. The task has a wide range of applications such as hearing aid design and robust automatic speech recognition (ASR). Monaural speech separation is proven to be very challenging as it only uses single-microphone recordings, especially in low SNR conditions. One way of dealing with this problem is to apply speech enhancement (Ephraim and Malah, 1984; Erkelens et al., 2007; Jensen and Hendriks, 2012) on a noisy signal, where certain assumptions are made regarding general statistics of the background noise. The speech enhancement approach is usually limited to relatively stationary noises. Looking at the problem from another perspective, computational auditory scene analysis (CASA) (Wang and Brown, 2006), which is inspired by psychoacoustic research in auditory scene analysis (ASA) (Bregman, 1990), exploits perceptual principles to speech separation.

In CASA, interference can be reduced by applying masking on a time–frequency (T–F) representation of noisy speech. An ideal mask suppresses noise-dominant T–F units and keeps the speech-dominant T–F units. Therefore, speech separation can be treated as a mask estimation problem where supervised learning is employed to construct the mapping from acoustic features to a mask. A binary decision on each T–F unit leads to an estimate of the ideal binary mask (IBM), which is defined as follows.

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if SNR}(t,f) > \mathrm{LC} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where t denotes time and f frequency. The IBM assigns the value 1 to a T–F unit if its SNR exceeds a local criterion (LC), and 0 otherwise. Therefore, speech separation is translated into a binary classification problem. Recent studies show IBM separation improves speech intelligibility in noise for both normal-hearing and hearing-impaired listeners

* Corresponding author. Tel.: +1 6146203690.
E-mail addresses: chenjit@cse.ohio-state.edu (J. Chen), wangyuxu@cse.ohio-state.edu (Y. Wang), dwang@cse.ohio-state.edu (D. Wang).

http://dx.doi.org/10.1016/j.specom.2015.12.006
0167-6393/

(Ahmadi et al., 2013; Brungart et al., 2006; Li and Loizou, 2008; Wang et al., 2009). Alternatively, a soft decision on each T–F unit leads to an estimate of the ideal ratio mask (IRM). The IRM is defined below (Narayanan and Wang, 2013).

$$\mathrm{IRM}(t,f) = \left(\frac{10^{(\mathrm{SNR}(t,f)/10)}}{10^{(\mathrm{SNR}(t,f)/10)}+1}\right)^{\beta} \quad (2)$$

where β is a tunable parameter. A recent study has shown that $\beta = 0.5$ is a good choice for the IRM (Wang et al., 2014). In this case, mask estimation becomes a regression problem where the target is the IRM. Ratio masking is shown to lead to slightly better objective intelligibility results than binary masking (Wang et al., 2014). In this study, we use the IRM with $\beta = 0.5$ as the learning target.

Supervised speech separation is a data-driven method where one expects a mask estimator to generalize from limited training data. However, training data only partially captures the true data distribution, thus a mask estimator can overfit training data and do a poor job in unseen scenarios. In supervised speech separation, a training set is typically created by mixing clean speech and noise. When we train and test on a nonstationary noise such as a cafeteria noise, there can be considerable mismatch between training noise segments and test noise segments, especially when the noise resource used for training is restricted. Similar problems can be seen in other supervised learning tasks such as image classification where the mismatch of training images and test images poses a great challenge. In image classification, a common practice is to transform training images using distortions such as rotation, translation and scaling, in order to expand the training set and improve generalization of a classifier (Ciresan et al., 2012; LeCun et al., 1998). We conjecture that supervised speech separation can also benefit from training data augmentation.

In this study, we aim at expanding the noise resource using noise perturbation to improve supervised speech separation. We treat noise expansion as a way to prevent a mask estimator from overfitting the training data. A recent study has shown speech perturbation improves ASR (Kanda et al., 2013). However, our study perturbs noise instead of speech since we focus on separating target speech from highly nonstationary noises where the mismatch among noise segments is the major problem. To our knowledge, our study is the first to introduce training data augmentation to the domain of speech separation.

This paper is organized as follows. Section 2 describes the system used for mask estimation. Noise perturbations are covered in Section 3. We present experimental results in Section 4. Section 5 concludes the paper. A preliminary version of this paper is included in Chen et al. (2015). Compared to the preliminary version, this paper has added a comparison with an alternative supervised separation method (Virtanen et al., 2013), detailed analysis of the three perturbation methods, and more evaluations in unvoiced and voiced intervals of speech, unmatched noises, expanded training and the very low SNR condition of −10 dB.

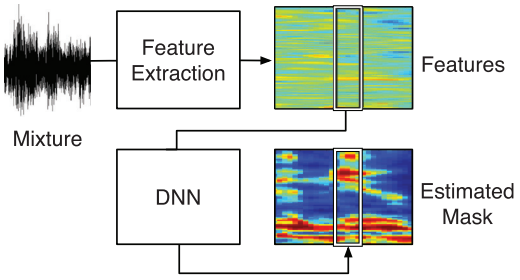

Fig. 1. Diagram of the proposed system.

2. System overview

To evaluate the effects of noise perturbation, we use a fixed system for mask estimation and compare the quality of estimated masks as well as the resynthesized speech that are derived from the masked T–F representations of noisy speech. While comparison between an estimated mask and an ideal mask reveals the spectrotemporal distribution of estimation errors, resythesized speech can be directly compared to clean speech. As mentioned in Section 1, we use the IRM as the target of supervised learning. The IRM is computed from the 64-channel cochleagrams of premixed clean speech and noise. The cochleagram is a time-frequency representation of a signal (Wang and Brown, 2006). We use a 20 ms window and a 10 ms window shift to compute cochleagram in this study. We perform IRM estimation using a deep neural network (DNN) and a set of acoustic features. Recent studies have shown that DNN is a strong classifier for ASR (Mohamed et al., 2012) and speech separation (Wang and Wang, 2013; Xu et al., 2014). As shown in Fig. 1, acoustic features are extracted from a mixture sampled at 16 kHz, and then sent to a DNN for mask prediction.

We use classification accuracy, hit minus false-alarm (HIT−FA) rate and short-time objective intelligibility (STOI) score (Taal et al., 2011) as three criteria for measuring the quality of the estimated IRM. Since the first two criteria are defined for binary masks, we calculate them by binarizing a ratio mask to a binary one. In this study, we follow Eqs. (3) and (1).

$$\mathrm{SNR}(t,f) = 10\log_{10}\left(\frac{\mathrm{IRM}(t,f)^2}{1-\mathrm{IRM}(t,f)^2}\right) \quad (3)$$

During the mask conversion, the LC is set to be 5 dB lower than the SNR of a given mixture. The three criteria evaluate the estimated IRM from three different perspectives. Classification accuracy computes the percentage of correctly labeled T–F units in a binary mask. In HIT−FA, HIT refers to the percentage of correctly classified target-dominant T–F units and FA refers to the percentage of wrongly classified interference-dominant T–F units. HIT−FA rate is well correlated with human speech intelligibility (Kim et al., 2009). In addition, STOI is computed by comparing the short-time envelopes of clean speech and resynthesized speech obtained from IRM masking, and it is a standard objective metric of speech intelligibility (Taal et al., 2011).