Noise Perturbation Improves Supervised Speech Separation

Jitong Chen(✉), Yuxuan Wang, and DeLiang Wang

The Ohio State University, Columbus, OH 43210, USA
{chenjit,wangyuxu,dwang}@cse.ohio-state.edu

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. Improving the generalization of a classifier is challenging, especially when interfering noise is strong and nonstationary. Expansion of a noise through proper perturbation during training exposes the classifier to more noise variations, and hence may improve separation performance. In this study, we examine the effects of three noise perturbations at low signal-to-noise ratios (SNRs). We evaluate speech separation performance in terms of hit minus false-alarm rate and short-time objective intelligibility (STOI). The experimental results show that frequency perturbation performs the best among the three perturbations. In particular, we find that frequency perturbation reduces 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. 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 [6] 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) 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.

E. Vincent et al. (Eds.): LVA/ICA 2015, LNCS 9237, pp. 83–90, 2015.
DOI: 10.1007/978-3-319-22482-4_10

A binary decision on each T-F unit leads to an estimate of the ideal binary mask (IBM), which is defined as follows.

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

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. IBM separation has been shown to improve speech intelligibility in noise for both normal-hearing and hearing-impaired listeners [9,13]. Alternatively, a soft decision on each T-F unit leads to an estimate of the ideal ratio mask (IRM). The IRM is defined below [10].

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

where β is a tunable parameter. A recent study has shown that $\beta = 0.5$ is a good choice for the IRM [15]. 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 [15]. In this study, we use the IRM with $\beta = 0.5$ as the learning target.

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. In this study, we aim at expanding the noise resource using noise perturbation to improve the generalization of 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 that speech perturbation improves ASR [7]. 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.

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. As mentioned in Sect. 1, we use the IRM as the learning target. The IRM is computed from the 64-channel cochleagrams of premixed clean speech and noise. A cochleagram is a T-F representation of a signal. We use a 20 ms window and a 10 ms window shift to compute a cochleagram.

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 [1] and speech separation [16]. As shown in Fig. 1, acoustic features are extracted from a mixture sampled at 16 kHz, and then sent to a DNN for mask

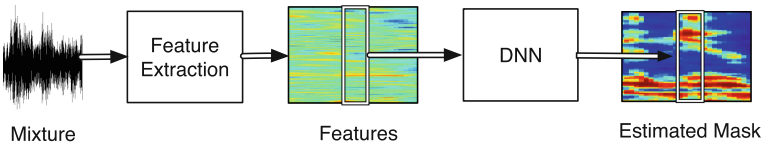

Fig. 1. Diagram of the proposed system.

prediction. To incorporate temporal context and obtain smooth mask estimation, we use 5 frames of features to estimate 5 frames of the IRM [15]. Therefore the output layer of the DNN has 64 × 5 units. Since each frame of the mask is estimated 5 times, we take the average of the 5 estimates. The acoustic features we extract from mixtures are a 59-D complementary feature set (AMS + RASTAPLP + MFCC) [14] combined with 64-D gammatone filterbank (GFB) features. To derive GFB features, an input signal is passed to a 64-channel gammatone filterbank, the response signals are decimated to 100 Hz to form 64-D GFB features.

We use hit minus false-alarm (HIT−FA) rate and short-time objective intelligibility (STOI) score [11] as two criteria for measuring the quality of the estimated IRM and the separated speech respectively. Since HIT−FA is defined for binary masks, we calculate it by binarizing a ratio mask to a binary one, following Eqs. 1 and 2. During the mask conversion, the LC is set to be 5 dB lower than the SNR of a given mixture. Both HIT−FA and STOI are well correlated with human speech intelligibity [8,11].

3 Noise Perturbation

The goal of noise perturbation is to expand noise segments to cover unseen scenarios so that the overfitting problem is mitigated in supervised speech separation. A recent study has found that three perturbations on speech samples improve ASR performance [7]. These perturbations were used to expand the speech samples by spectral perturbation. The three perturbations are introduced below. Unlike this study, we perturb noise samples instead of perturbing speech samples, as we are dealing with highly nonstationary noises.

3.1 Noise Rate (NR) Perturbation

Speech rate perturbation, a way of speeding up or slowing down speech, is used to expand training utterances during the training of an ASR system. In our study, we extend the method to vary the rate of nonstationary noises. We increase or decrease noise rate by factor γ. When a noise rate is being perturbed, the value of γ is randomly selected from an interval $[\gamma_{min}, 2 - \gamma_{min}]$. The effect of NR perturbation on a spectrogram is shown in Fig. 2a.

3.2 Vocal Tract Length (VTL) Perturbation

VTL perturbation has been used in ASR to cover the variation of vocal tract length among speakers. A recent study suggests that VTL perturbation improves

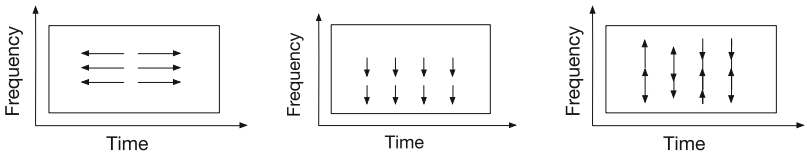

Fig. 2. (a) Illustration of noise rate perturbation. (b) Illustration of vocal tract length perturbation. (c) Illustration of frequency perturbation.

ASR performance [5]. VTL perturbation essentially compresses or stretches the medium and low frequency components of an input signal. We use VTL perturbation as a method of perturbing a noise segment. Specifically, we follow the algorithm in [5] to perturb noise signals:

$$f' = \begin{cases} f\alpha, & \text{if } f \leq F_{hi}\frac{min(\alpha,1)}{\alpha} \\ \frac{S}{2} - \frac{\frac{S}{2}-F_{hi}min(\alpha,1)}{\frac{S}{2}-F_{hi}\frac{min(\alpha,1)}{\alpha}}(\frac{S}{2}-f), & \text{otherwise} \end{cases} \tag{3}$$

where f is the original frequency, f' is the mapped frequency, α is the wrapping factor, S is the sampling rate, and F_{hi} controls the cutoff frequency. The effect of VTL perturbation is visualized in Fig. 2b.

3.3 Frequency Perturbation

When frequency perturbation is applied, frequency bands of a spectrogram are randomly shifted upward or downward. We use the method described in [7] to randomly perturb noise samples. Frequency perturbation takes three steps. First, we randomly assign a value to each T-F unit, which is drawn from a uniform distribution.

$$r(f,t) \sim U(-1,1) \tag{4}$$

Then we derive the perturbation factor $\delta(f,t)$ by averaging the assigned values of neighboring T-F units. This averaging step avoids large oscillations in spectrogram.

$$\delta(f,t) = \frac{\lambda}{(2p+1)(2q+1)} \sum_{f'=f-p}^{f+p} \sum_{t'=t-q}^{t+q} r(f',t') \tag{5}$$

where p and q control the smoothness of the perturbation, and λ controls the magnitude of the perturbation. These tunable parameters are decided experimentally. Finally the spectrogram is perturbed as follows.

$$\tilde{S}(f,t) = S(f+\delta(f,t),t) \tag{6}$$

where $S(f,t)$ represents the original spectrogram and $\tilde{S}(f,t)$ is the perturbed spectrogram. Interpolation between neighboring frequencies is used when $\delta(f,t)$ is not an integer. The effect of frequency perturbation is visualized in Fig. 2c.

4 Experimental Results

4.1 Experimental Setup

We use the IEEE corpus recorded by a male speaker [4] and six nonstationary noises from the DEMAND corpus [12] to create mixtures. All signals are sampled at 16 KHz. Note that all recordings of the DEMAND corpus are made with a 16-channel microphone array, we use only one channel of the recordings since this study is on monaural speech separation. We choose six nonstationary noises (each is five-minute long) from the DEMAND corpus, each representing distinct environment: SCAFE noise (recorded in the terrace of a cafe at a public square), DLIVING noise (recorded inside a living room), OMEETING noise (recorded in a meeting room), PCAFETER noise (recorded in a busy office cafeteria), NPARK noise (recorded in a well visited city park) and TMETRO noise (recorded in a subway).

To create a mixture, we mix one IEEE sentence and one noise type at -5 dB SNR. This low SNR is selected with the goal of improving speech intelligibility in mind where there is not much to improve at higher SNRs [3]. The training set uses 600 IEEE sentences and randomly selected segments from the first two minutes of a noise, while the test set uses another 120 IEEE sentences and randomly selected segments from the second two minutes of a noises. Therefore, the test set has different sentences and different noise segments from the training set. We create 50 mixtures for each training sentence by mixing it with 50 randomly selected segments from a given noise, which results in a training set containing 600×50 mixtures. The test set includes 120 mixtures. We train and test using the same noise type and SNR condition.

To perturb a noise segment, we first apply short-time Fourier transform (STFT) to derive noise spectrogram, where a frame length of 20 ms and a frame shift of 10 ms are used. Then we perturb the spectrogram and derive a new noise segment. The parameters of perturbations are selected by using a development set. To evaluate the three noise perturbations, we create five different training sets, each consists of 600×50 mixtures. We train a mask estimator for each training set and evaluate on a fixed test set (i.e. the 120 mixtures created from the original noises). The five training sets are described as follows.

1. Original Noise: All mixtures are created using original noises.
2. NR Perturbation: Half of the mixtures are created from NR perturbed noises, and the other half are from original noises.
3. VTL Perturbation: Half of the mixtures are created from VTL perturbed noises, and the other half are from original noises.
4. Frequency Perturbation: Half of the mixtures are created from frequency perturbed noises, and the other half are from original noises.
5. Combined: Half of the mixtures are created from applying three perturbations altogether, and the other half are from original noises.

As already mentioned, we extract a set of four complementary features (AMS + RASTAPLP + MFCC + GFB) from mixtures. Delta features are appended to

Table 1. HIT−FA rate (in %) for six noises at −5 dB, where FA is shown in parentheses.

Perturbation \ Noise	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	55 (37)	70 (23)	65 (28)	50 (40)	69 (22)	63 (32)	62 (30)
NR perturbation	64 (24)	77 (15)	72 (18)	60 (26)	77 (12)	72 (21)	70 (19)
VTL Perturbation	64 (24)	76 (16)	71 (19)	60 (27)	78 (10)	72 (21)	70 (20)
Frequency Perturbation	69 (17)	77 (14)	74 (15)	63 (21)	79 (9)	74 (18)	73 (16)
Combined	67 (21)	77 (15)	73 (16)	61 (25)	78 (10)	74 (18)	72 (18)

Table 2. STOI (in %) of separated speech for six noises at −5 dB, where STOI of unprocessed mixtures is shown in parentheses.

Perturbation \ Noise	SCAFE	DLIVING	OMEETING	PCAFETER	NPARK	TMETRO	Average
Original Noise	73.7 (64.1)	87.5 (79.3)	80.0 (67.8)	71.4 (62.5)	80.2 (67.7)	85.9 (77.5)	79.8 (69.8)
NR perturbation	76.5 (64.1)	89.2 (79.3)	82.5 (67.8)	74.1 (62.5)	83.2 (67.7)	87.4 (77.5)	82.1 (69.8)
VTL Perturbation	76.1 (64.1)	88.7 (79.3)	82.2 (67.8)	74.0 (62.5)	83.6 (67.7)	87.2 (77.5)	82.0 (69.8)
Frequency Perturbation	78.2 (64.1)	89.1 (79.3)	83.3 (67.8)	75.1 (62.5)	84.1 (67.7)	87.8 (77.5)	82.9 (69.8)
Combined	77.0 (64.1)	88.6 (79.3)	82.7 (67.8)	74.7 (62.5)	83.8 (67.7)	87.6 (77.5)	82.4 (69.8)

the feature set. A four-hidden-layer DNN is employed to learn the mapping from acoustic features to the IRM. Each hidden layer of the DNN has 1024 rectified linear units [1]. Dropout [1] and adaptive stochastic gradient descent [2] are used to train the DNN.

4.2 Evaluation Results and Comparisons

We evaluate the three perturbations with the five large training sets described in Sect. 4.1. The effects of noise perturbations on speech separation are shown in Tables 1 and 2, in terms of HIT−FA rate and STOI score respectively. The results indicate that all three perturbations lead to better speech separation than the baseline where only the original noises are used. Frequency perturbation performs better than the other two perturbations. Compared to only using the original noises, the frequency perturbed training set on average increases HIT−FA rate and STOI score by 11 % and 3 %, respectively. This indicates that noise perturbation is an effective technique for improving speech separation results. Combining three perturbations, however, does not lead to further improvement over frequency perturbation.

A closer look at Table 1 reveals that the contribution of frequency perturbation lies mainly in the large reduction in FA rate. This means that the problem of misclassifying noise-dominant T-F units as speech-dominant is mitigated. This effect can be illustrated by visualizing the masks estimated from the different training sets and the ground truth mask in Fig. 3a (e.g. around frame 150). When the mask estimator is trained with the original noises, it mistakenly retains the regions where target speech is not present, which can be seen by comparing

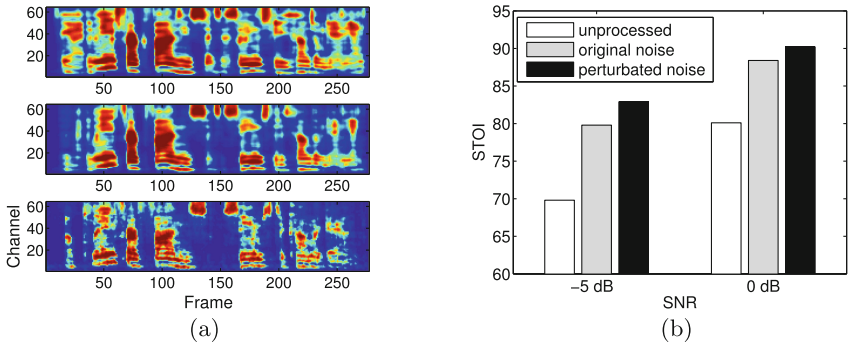

Fig. 3. (a) Mask comparison, the top shows a mask estimate using original noise, the middle shows a mask esimate using perturbed noise, the bottom shows the IRM. (b) The effect of frequency perturbation, the average STOI scores (in %) for six noises are shown for unprocessed speech, separated speech using original noise, and separated speech using frequency perturbed noise.

the top and bottom plots of Fig. 3a. Applying frequency perturbation to noises essentially exposes the mask estimator to more noise patterns and results in a more accurate mask estimator, which is shown in the middle plot of Fig. 3a. While HIT−FA rate evaluate the estimated binary masks, STOI directly compares clean speech and the resynthesized speech. As shown in Table 2, frequency perturbation yields higher average STOI scores than using original noises with no perturbation and NR and VTL perturbations.

Finally, to evaluate the effectiveness of frequency perturbation at a higher SNR, we carry out additional experiments at 0 dB input SNRs, where we use the same parameter values as for −5 dB SNR. Figure 3b shows frequency perturbation improves speech separation in terms of STOI in each SNR condition. Also, we find that frequency perturbation remains the most effective among the three perturbations at 0 dB SNR.

5 Concluding Remarks

In this study, we have explored the effects of noise perturbation on supervised monaural speech separation at low SNR levels. Noise perturbation is used to expand training noise to improve generalization of a classifier. We have evaluated three noise perturbations with six nonstationary noises recorded from daily life for speech separation. The three are noise rate, VTL, and frequency perturbations. With perturbed noises, the quality of the estimated ratio mask is improved as the classifier has been exposed to more scenarios of noise interference. In contrast, a mask estimator learned from a training set that only uses original noises tends to make more false alarm errors (i.e. higher FA rate). The experimental results show that frequency perturbation, which randomly perturbs the noise spectrogram along frequency, almost uniformly gives the best speech separation results among the three perturbations examined in terms of HIT−FA rate and STOI score.

Acknowledgments. This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

References

1. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: Proceedings of the ICASSP, pp. 8609–8613 (2013)
2. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
3. Healy, E.W., Yoho, S.E., Wang, Y., Wang, D.L.: An algorithm to improve speech recognition in noise for hearing-impaired listeners. J. Acoust. Soc. Am. **134**, 3029–3038 (2013)
4. IEEE: IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17, 225–246 (1969)
5. Jaitly, N., Hinton, G.E.: Vocal Tract Length Perturbation (VTLP) improves speech recognition. In: Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processes (2013)
6. Jensen, J., Hendriks, R.C.: Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions. IEEE Trans. Audio, Speech, Lang. Process. **20**, 92–102 (2012)
7. Kanda, N., Takeda, R., Obuchi, Y.: Elastic spectral distortion for low resource speech recognition with deep neural networks. In: Proceedings of the ASRU, pp. 309–314 (2013)
8. Kim, G., Lu, Y., Hu, Y., Loizou, P.C.: An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. **126**, 1486–1494 (2009)
9. Li, N., Loizou, P.C.: Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J. Acoust. Soc. Am. **123**, 1673–1682 (2008)
10. Narayanan, A., Wang, D.: Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: Proceedings of the ICASSP, pp. 7092–7096 (2013)
11. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio, Speech, Lang. Process. **19**, 2125–2136 (2011)
12. Thiemann, J., Ito, N., Vincent, E.: The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. J. Acoust. Soc. Am. **133**, 3591 (2013)
13. Wang, D.L., Kjems, U., Pedersen, M.S., Boldt, J.B., Lunner, T.: Speech intelligibility in background noise with ideal binary time-frequency masking. J. Acoust. Soc. Am. **125**, 2336–2347 (2009)
14. Wang, Y., Han, K., Wang, D.L.: Exploring monaural features for classification-based speech segregation. IEEE Trans. Audio, Speech, Lang. Process. **21**, 270–279 (2013)
15. Wang, Y., Narayanan, A., Wang, D.L.: On training targets for supervised speech separation. IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**, 1849–1858 (2014)
16. Wang, Y., Wang, D.L.: Towards scaling up classification-based speech separation. IEEE Trans. Audio, Speech, Lang. Process. **21**, 1381–1390 (2013)