TRAINING SUPERVISED SPEECH SEPARATION SYSTEM TO IMPROVE STOI AND PESQ DIRECTLY

Hui Zhang, Xueliang Zhang, Guanglai Gao*

College of Computer Science, Inner Mongolia University, China

alzhu.san@163.com, {cszxl, csggl}@imu.edu.cn

ABSTRACT

Supervised speech separation methods train learning machine to cast the noisy speech to the target clean speech. Most of them use mean-square error (MSE) as loss function. However, MSE is not the perfect choice because it doesn't match the human auditory perception. Short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) are closely related to the human auditory perception and widely used in speech separation research as evaluation criteria. Therefore, STOI and PESQ may be better choices for the loss function. However, they are nondifferentiable functions which cannot be optimized by the conventional gradient descent algorithm. In this work, a gradient approximation method is used to calculate the gradients of the STOI and PESQ. Then the calculated gradients are used in the gradient descent algorithm to optimize the STOI and PESQ directly. Experimental results show the speech separation performance can be improved by the proposed method.

Index Terms— Monaural speech separation, Short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ), gradient approximation

1. INTRODUCTION

Monaural speech separation separates target speech from additive noise signal by using only one microphone. It has been widely studied to improve the performance of various signal processing systems, including hearing prosthesis, mobile telecommunication, and robust automatic speech and speaker recognition [1]. For a few decades, monaural speech separation systems have achieved considerable performance improvements, especially after formalizing it as a supervised learning problem and using deep learning algorithms.

Early studies for monaural speech separation, e.g. spectral subtraction, are mostly based on the mean-square error (MSE) criterion [2] which can improve the perceptual speech quality. However, these approaches typically assume that background noise is stationary, i.e. its spectral properties do not change over time, or are stationary than the speech at least. Therefore, they have difficulties in tracking non-stationary noises, which limits its application in real-world environments.

In order to enhance the noisy speech in various noisy environments, more powerful models are involved. Deep neural networks (DNNs) and long short term memory networks (LSTMs) can model the complicated relationship between the input variables and the output targets. They were successfully introduced to the speech separation area as supervised speech separation, and obtained considerable performance improvements. In these approaches, a learning machine (DNN or LSTM) is trained to cast the acoustic features of the noisy speech to a time-frequency mask, or the spectrum of the clean speech, where these two categories methods can be generally referred as the masking-based and the mapping-based methods. Many works devoted to the supervised speech separation, which covered the most aspects of the supervised learning: Wang concluded the related works on features [3] and training targets [4], and many works studied the learning machine and its training methods [5–9]. But very few studies investigated the loss function, and most of the learning-based method employ MSE. For example, the masking-based method minimizes the MSE between the estimation and the ideal mask target, and the mapping-based method minimizes the MSE between the estimation and the target clean spectrum.

Although a lot of works show the effectiveness of the MSE. In fact, the MSE is not a perfect loss function to evaluate the estimation, because it is not closely related to the human auditory perception. The MSE has two weaknesses: it treats the estimation elements independently and equally. a) the MSE will lead to over-smooth speech trajectories and may result in muffled sound quality and decreased intelligibility [6]. Because the MSE measures are derived from each time-frequency (T-F) unit separately rather than from whole spectral trajectory. b) it treats every estimation elements with equal importance, in fact, they are not. For speech intelligibility, the distinguishable phones are more important, and for speech quality, the isolated points are more harmful which may lead to musical noise. The MSE is usually defined in the linear frequency scale, but the human auditory perception follows the Mel-frequency scale. Therefore improving the human auditory perception quality

or the speech intelligibility by minimizing the MSE is a game which is not worth the candle.

To take over the shortnesses of the MSE, the new loss function should take the whole speech or long duration into its consideration and give different weights to the estimation elements following the human auditory perception. Some works [10, 11] have applied the element-wise weight function and added penalty terms to the MSE. Kang proposed to take the temporal and spectral variations equalization into the loss function [12, 13]. However, these works do not exactly match the requirements. Short-time objective intelligibility (STOI) [14] and perceptual evaluation of speech quality (PESQ) [15] are closely related to the human auditory perception and are widely used in the speech separation research as evaluation criteria. Because the STOI and PESQ evaluate the separated speech as a whole, and give different importance to the estimation elements according to the modeled human auditory perception systems. We think that they would be better loss functions than the MSE. If we can train the speech separation model to improve the STOI and PESQ directly, we can improve the evaluation criteria directly, too.

Taking the STOI and PESQ as the loss function is straightforward, but training the model against these loss functions is not simple because they are nondifferentiable. It is very difficult to optimize a nondifferentiable function by the conventional machine learning algorithms such as the gradient descent. In this work, we propose to use a gradient approximation method to calculate the gradient of the STOI and PESQ loss function. Experimental results show the effectiveness of the gradient approximation method. Furthermore, results show the speech separation performance can be improved by training against the STOI and PESQ loss functions directly using the proposed method.

2. SYSTEM DESCRIPTION

2.1. Loss functions

STOI is a standard objective metric for speech intelligibility. It shows a high correlation ($r > 0.9$) with speech intelligibility scores in subjective listening tests, and is widely used in the speech separation and enchantment researches. STOI computes the correlation of short-time temporal envelopes between the clean and separated speech. It varies in [0, 1], and a higher value indicates the better speech intelligibility.

PESQ is a standard objective metric for speech perceived quality recommended by ITU-T (Recommendation P.862). It was developed to predict the mean opinion scores (MOS) in subjective listening tests. It shows a high correlation ($r \approx 0.9$) with MOS on the noise-corrupted speech processed via noise suppression algorithms [16], and is widely used in the speech separation and enchantment researches. PESQ measures speech quality by computing disturbance between the clean speech and the separated speech using cognitive modeling, which ranges in [-0.5, 4.5], with high values indicating better quality.

We got the MATLAB source code for the STOI from the author's website, and the PESQ from the CD-ROM included in [17]. We can access the STOI source code, while the PESQ is provided as a set of protected function files (p-code), which is a complete black-box to us. Restricted by the accessibility and complexity of the loss functions, in order to optimize the STOI and PESQ directly, we need an algorithm which does not require the explicit expression of the loss functions.

2.2. Gradient approximation

Gradient descent is the most common choice for optimization in the machine learning research. Gradient descent takes steps proportional to the negative of the gradient of the loss function at the current point, in where the gradient defines the move direction. However, the gradient is not must be 100% accurate, it can be an approximation. Accurate gradients require the explicit expression of the loss functions, but approximate gradients not.

The gradient is a generalization of the derivative. The derivative measures the sensitivity to change of the function output value with respect to a change in its argument. The derivative of the $f(x)$ at a is defined as:

$$\left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=a} = \lim_{h\to 0}\frac{f(a+h)-f(a)}{h} \tag{1}$$

It can be approximated by calculating with a small h:

$$\left.\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right|_{x=a} \approx \frac{f(a+h)-f(a)}{h} \tag{2}$$

The gradient generalizes the derivative on multi-variable. The gradient of the $f(x)$ at a also can be approximated with a similar method:

$$\bigtriangledown_x|_{x=a} \approx \frac{f(a+\sigma\epsilon)-f(a)}{\sigma\epsilon} \tag{3}$$

Where x is a vector, ϵ is a random vector sampled from the unit Gaussian distribution, σ is a small constant, where h in the formula (2) is replaced by $\sigma\epsilon$ in multi-variable condition. To give more confidence to the gradient approximation, we sample the ϵ more time (N) and get:

$$\mathbb{E}\left(\bigtriangledown_x|_{x=a}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\frac{f(a+\sigma\epsilon_i)-f(a)}{\sigma\epsilon_i} \tag{4}$$

This gradient estimator is known as simultaneous perturbation stochastic approximation [18], parameter exploring policy gradients [19], or zero-order gradient estimation [20].

Appling the approximate gradient in the gradient descent algorithm, it results in the algorithm 1, where T is the training

epochs, N is the sample size. Specifically, we start with some initial parameters w_0. At each step we take a parameter w_t and explore its neighborhood by jittering w_t with small Gaussian noise, then get the approximate gradient of the loss function $f(w)$ at w_t. Then we move the slightly along the negative direction of the approximate gradient, which results in a new parameter w_{t+1}. This procedure is iterated until the loss function is fully optimized.

Algorithm 1 Approximate gradient descent

Require: Learning rate α, noise variance σ^2, initial parameters w_0
  for $t = 1, 2, 3 \ldots, T$ **do**
    Sample $\epsilon_1, \ldots \epsilon_n \sim \mathcal{N}(0, I)$
    Compute $F_i = f(w_t + \sigma\epsilon_i) - f(w_t)$ for $i = 1, \ldots, N$
    Set $w_{t+1} \leftarrow w_t - \alpha \frac{1}{N\sigma} \sum_{i=1}^{N} F_i/\epsilon_i$
  end for

The approximate gradient descent algorithm treats the loss function as a black-box. It does not require to access the explicit expression of the loss function and does not require the loss function is differentiable. With this algorithm, many non-trivial loss functions, such as the STOI and PESQ in this study, can be optimized directly.

3. EXPERIMENTAL RESULTS

We use the approximate gradient descent algorithm to train speech separation models to maximize the STOI and PESQ directly. Three models are involved: "STOI-model", "PESQ-model" and "combine-model", whose loss function is STOI, PESQ and the combination of them. To evaluate these proposed models, we compare them with a model trained with MSE loss function ("MSE-model").

3.1. Dateset

We use 2000 randomly chosen utterances from the TIMIT [21] training set as our training utterances, and use the TIMIT core test set as our test utterances. In the TIMIT core test set, there are 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). We use a speech shape noise (SSN) and 4 other noises from the NOISEX dataset [22]: a babble noise, a factory noise (factory1), a destroyer engine room noise (destroyerengine), and an operation room noise (destroyerops) for training. Aside from the aforementioned noises, we also use an unseen factory noise (factory2) and a tank noise (m109) from NOISEX to evaluate generalization performance. To create the training sets, we use random cuts from the first 2 minutes of each noise to mix with the training utterances at -5 and 0 dB SNR. The test mixtures are constructed by mixing random cuts from the endmost 2 minutes of each noise with the test utterances at -5, 0 and 5 dB SNR, where 5 dB is an unseen SNR condition.

3.2. Model

The speech separation model is a long short-term memory (LSTM) recurrent neural network (RNN). It has 3 layers and there are 384 memory cells in each layer. This model is trained to map the acoustic features of the noisy speech to the amplitude spectrum of the clean speech. The input features are based on the short time Fourier transform (STFT) of the mixture signal. Under the sampling frequency of 16k Hz, the STFT is obtained using the 320-point (20 ms) hamming window with 50% overlap. As the STFT is conjugate symmetric, in each frame, a preliminary feature vector is formed using the amplitude of only the first 161 STFT coefficients. Then the vector is cubic-rooted and normalized to be zero-mean and unit variance. The training target is STFT vector of the clean speech signal without cubic-rooting and normalization.

We train the model 600 epochs with the Adam optimizer [23] against the MSE loss function as the baseline, and continue train this baseline model 10 epochs against the MSE, STOI, PESQ and the combine loss function, as MSE-model, STOI-model, PESQ-model, and combine-model. The combine loss function is $5 \cdot \mathrm{STOI} + \mathrm{PESQ}$, where factor 5 is used to balance the two parts. In the approximate gradient descent algorithm, we set $\sigma = 0.01$, $N = 20$, $T = 10$.

3.3. Results

We first report the performance of the baseline model. In Fig. 1, we list the average STOI and PESQ scores on all test data after each training epoch. Its performance is improved with the training. Its performance gains a lot in the first several epochs, then the improving speed slows down. After 600 epochs training, the STOI and PESQ scores are 0.8183 and 2.3804, improve 0.1174 and 0.5826 compared to the unprocessed noisy speech. The baseline model is comparable with the state-of-the-art speech separation models.

We train the baseline model another 10 epochs with different loss functions. The average STOI and PESQ scores are shown in Fig. 2. The MSE-model continues the changing trend of the baseline model. Finally, the MSE-model improves neither the STOI or the PESQ scores. The STOI-model improves the STOI score much but the PESQ score few. The PESQ-model improves the PESQ score much but the STOI score few. By combining the STOI and PESQ, the combine-model improves both the STOI and PESQ scores. As our expected, the combine-model cannot achieve high STOI scores as the STOI-model, while interestingly, the combine-model obtains a higher PESQ score than the PESQ-model. It indicates the PESQ scores can be improved with the help of the STOI loss function. All of these three models improve the STOI and PESQ scores. Compared with the MSE-mode, the STOI-model, PESQ-model, and combine-model improve the STOI scores with 0.0129, -0.0007 and 0.0048, and improve the PESQ scores with 0.0053, 0.0520 and 0.0843, respectively.

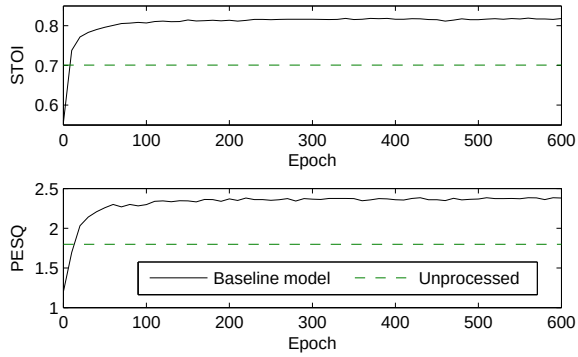

Fig. 1. Average performance of the baseline model.

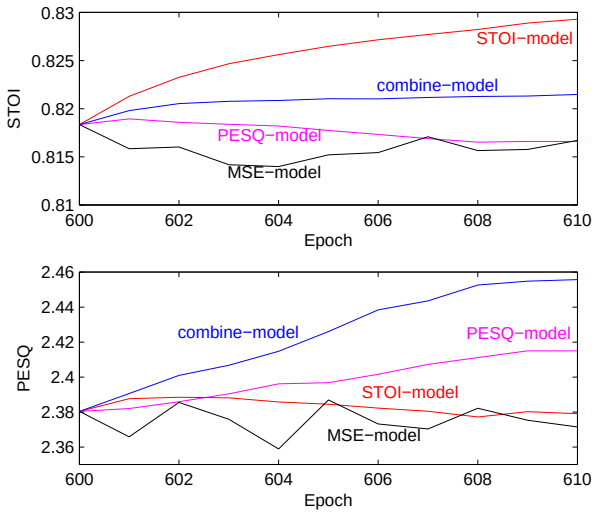

Fig. 2. Average performance of the models trained with different loss function.

3.4. Loss Function Visualizing

With the approximate gradient descent algorithm, we can find out what the STOI and PESQ actually care, by maximizing the STOI and PESQ, directly.

We select a clean speech and use its waveform as our target. Its amplitude spectrum is shown in Fig. 3(a). We start from a random vector (pure white noise, no information about the target are included), then optimize it to maximize the STOI, PESQ and their combination. In this experiment, we set $\sigma = 0.01$, $N = 20$, $T = 10000$. The optimization results are shown in Fig. 3 (b)-(d). Form Fig. 3, we can see that the approximate gradient descent algorithm can maximize the STOI, PESQ and their combination. The viewpoints of the STOI and PESQ are different. They are located on the different side of a seesaw, one rises and the other falls. The combination loss function gets a balance between them.

By comparing the optimized and the target spectrum, we can discover the important and unimportant parts under different loss functions. We can see that both of the STOI and PESQ give more value to the low-frequency parts as the human auditory system. The STOI gives more importance to the speech parts but few to the silence part, since STOI removes silence before its calculation. Compared with the STOI, the PESQ take more attention on the inactive T-F units in the speech parts, because that these T-F units most likely to damage the speech quality. Their combination pays attention to both of their more valued parts.

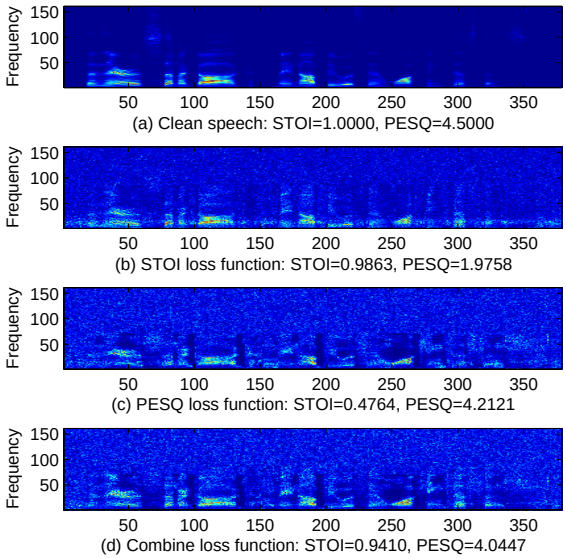

Fig. 3. Optimizing against different loss functions.

4. CONCLUSION

The STOI and PESQ evaluate the speech intelligibility and quality similar to the human audition system. These two criteria may be better choices for the loss function in the supervised speech separation system. But, they are nondifferentiable functions which cannot be optimized by the conventional machine learning algorithm. In this study, an approximate gradient descent algorithm is proposed to optimize them directly. Experimental results show the speech separation performance can be improved by the proposed method. The optimization algorithm treats the loss function as a black-box, which does not require it is differentiable. In this way, many nondifferentiable loss functions can be used. The system with nondifferentiable modules can be actually trained jointly. For example, in the robust speech recognition system, the speech separation model can be trained to maximize the recognition accuracy, directly.

5. ACKNOWLEDGMENTS

This research was supported in part by the China national nature science foundation (No. 61773224, No. 61365006, No. 61263037).

6. REFERENCES

[1] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *arXiv preprint arXiv:1708.07524*, 2017.

[2] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[3] Yuxuan Wang, Kun Han, and DeLiang Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[4] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *Audio Speech & Language Processing IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1849–1858, 2014.

[5] Zhaozhang Jin and DeLiang Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[6] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[7] Rupesh Kumar Srivastava, Klaus Greff, and Jrgen Schmidhuber, "Highway networks," *Computer Science*, 2015.

[8] Xueliang Zhang, Hui Zhang, Shuai Nie, Guanglai Gao, and Wenju Liu, "A pairwise algorithm using the deep stacking network for speech separation and pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1066–1078, June 2016.

[9] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[10] Bingyin Xia and Changchun Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[11] Prashanth Gurunath Shivakumar and Panayiotis G Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement.," in *INTERSPEECH*, 2016, pp. 3743–3747.

[12] Tae Gyoon Kang, *Deep Learning Approach for Robust Voice Activity Detection and Speech Enhancement*, Ph.D. thesis, Seoul National University, Seoul, Korea, 2017.

[13] Tae Gyoon Kang, Jong Won Shin, and Nam Soo Kim, "Dnn-based monaural speech enhancement with temporal and spectral variations equalization," *Digital Signal Processing*, vol. 74, pp. 102–110, 2018.

[14] Cees Taal, Richard Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *icassp*, 2001, pp. 749–752.

[16] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[17] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[18] J. C Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *Automatic Control IEEE Transactions on*, vol. 37, no. 3, pp. 332–341, 1992.

[19] F Sehnke, C Osendorfer, T Rckstiess, A Graves, J Peters, and J Schmidhuber, "Parameter-exploring policy gradients," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 23, no. 4, pp. 551, 2010.

[20] Yurii Nesterov and Vladimir Spokoiny, "Random gradient-free minimization of convex functions," *Core Discussion Papers*, vol. 17, no. 2, pp. 527–566, 2017.

[21] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[22] Andrew Varga and Herman J.M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[23] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.