A SPECTRAL-CHANGE-AWARE LOSS FUNCTION FOR DNN-BASED SPEECH SEPARATION

Xiang Li, Xihong Wu, Jing Chen

Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China

ABSTRACT

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. Interference can be reduced by applying the estimated mask on a time-frequency (T-F) representation of noisy speech, resulting in improved speech intelligibility. Most of existing learning networks for speech separation aim to minimize the Mean Square Error (MSE) over the training set, where the loss from each T-F representation is equally weighted. In this paper, we proposed a spectral-change-aware loss function, where loss from the T-F units with large spectral changes over time were assigned higher weights compared to the T-F units with minor spectral changes. Such spectral-change-aware loss function was evaluated on speech separation performance in terms of mask estimation accuracy, short-time objective intelligibility (STOI) and SNR gain of unvoiced segments. The results indicated that the proposed loss function could further improve the speech intelligibility and increase SNR gain of unvoiced segments even in the cost of increased error rate of estimated mask.

Index Terms— Speech separation, loss function, spectral change, speech intelligibility

1. INTRODUCTION

Approximately 5% of the population in the world is suffering from the hearing loss, further along with reduced speech intelligibility in background noise [1, 2]. Extensive efforts have been made to improve speech intelligibility for the hearing impaired (HI) over the past several decades. Recently, methods based on supervised speech separation showed significant performance on speech enhancement, where an ideal time-frequency (T-F) mask is used as the computational objective [3]. A trained classifier, typically, a deep neural network (DNN) is used to estimate the ideal T-F mask, which indicates whether, or to what extent, each T-F unit is dominant by the target speech. A binary decision leads to the ideal binary mask (IBM) while a ratio decision leads to the ideal ratio mask (IRM) [4]. DNN-based IBM and IRM separation have been shown to improve speech intelligibility in noise for both normal hearing and HI listeners [5].

Most of existing learning networks for speech separation aim to minimize the Mean Square Error (MSE) over the training set, and the loss from each T-F representation is equally weighted. However, studies of speech perception reveal that auditory system is especially sensitive to abrupt changes in stimuli, and listeners may utilize the temporal changes to enhance perception of auditory objects [6, 7, 8]. In addition, perceptually-motivated enhancement approaches have manipulated "landmark" regions of the signal that are known to contain a high concentration of acoustic cues to phonetic identity, resulting in improved speech recognition in background noise [9, 10]. These landmark regions can be inherently transient and of low energy, e.g. the perceptually-important formant transitions following plosive release, hence they could be easily masked by background noise and difficult to be retrieved from noisy mixtures, especially when the noise level is high. Recently, a spectral-change evaluation (SCE) algorithm which aims to extract and further enhance the effective spectral changes from noisy mixtures has been demonstrated to improve speech intelligibility in noise for HI listeners [11, 12, 13]. Due to the importance of dynamic cues in the signal, separation networks could be adjusted to put more effort into training regions containing dynamic changes, in order to increase separation accuracy of these "landmark" regions and further to improve speech intelligibility. One straightforward method is to add some constrains to common loss functions to control the training effort of networks between dynamic regions and other non-dynamic regions.

In this work, the basic idea was to make DNN-based IBM or IRM system work like human auditory system, to be sensitive to those regions with dynamics and transitions. This idea was addressed by introducing a spectral-change-aware loss function, where loss from the T-F units with large spectral changes over time were assigned higher weights compared to regions with few changes. The previously proposed SCE was applied to each T-F unit to effectively extract spectral changes from the premixed clean speech. The extracted spectral changes were used as the weights assigned to the loss of corresponding T-F units. The effect of the proposed loss function was evaluated in terms of mask estimation accuracy,

short-time objective intelligibility (STOI) and SNR gain of unvoiced segments.

2. SPECTRAL-CHANGE-AWARE LOSS FUNCTION

2.1. Spectral-change-aware loss function

The weight assigned to T-F units was indicated by the spectral changes within each T-F unit, which was applied to the commonly used MSE as a constraint during the DNN training, constructing the spectral-change-aware loss function. The spectral changes of each T-F unit were extracted by a spectral change evaluation method, which was used in [11].

The input mixture sampled at 16 kHz was first passed through a 64-channel gammatone filterbank. The output in each channel was then divided into 20-ms frames with 10-ms overlapping between consecutive frames. The basic spectral change over time was first derived by calculating the difference of the T-F representations across every two adjacent frames.

A spectral change function (SCF) was then derived by the convolution of the basic spectral change with a difference-of-Gaussians (DoG) function, in order to emphasize the contrast between spectral peaks and valleys as well as to remove minor irregularities in the spectrum.

To take the influence of preceding frames into account, a Gain function, $Gain(t, f)$ for a certain frame *t* and sub-band *f*, was defined by a weighted average of the SCF over several preceding frames with a weight that progressively declined for frames that were earlier in time than the current frame. Then, the Gain function $Gain(t, f)$ was scaled by a factor to produce a controllable spectral change weight $SCW(t, f)$.

Therefore, the proposed loss function was defined as,

$$\begin{aligned} loss = & \frac{1}{2T}\sum_{t}\sum_{f}\|y_{t,f}-\tilde{y}_{t,f}\|^2 + \\ & \frac{1}{2T}\sum_{t}\sum_{f}\|SCW_{t,f}\cdot(y_{t,f}-\tilde{y}_{t,f})\|^2 \end{aligned} \tag{1}$$

where $y_{t,f}$ denoted ground truth IRM at a specific T-F unit; $\tilde{y}_{t,f}$ denoted the corresponding estimated IRM; $SCW_{t,f}$ denoted the weight assigned to this T-F unit, which was derived by SCE from premixed clean speech; *T* was the total number of frames in the training data. Such loss function was addressed based on the idea that the training network was expected to give more weight to T-F units with larger spectral changes, in order to increase separation accuracy of these dynamic regions and further contribute to the speech intelligibility of the final synthetic speech. Here, the weight was indicated by the spectral changes within each T-F unit.

2.2. Analysis of the weight

The weight calculated from the spectral change evaluation (SCE) should exactly indicate the spectral changes in each T-F unit. Hence, the spectral change weight $SCW_{t,f}$ was plotted here to visually examine whether the SCE can effectively extract the spectral changes over time as a guidance to facilitate IRM estimation.

Fig. 1 showed the grayscale images of cochleagram of the premixed clean speech, the spectral changes calculated directly by SCE and the final weight used to adjust the loss of each T-F unit. Panel (b) showed the greyscale image of the spectral changes calculated by SCE which indeed captured the dynamics at the onsets of syllables and even the formant transitions with relatively low intensity. Regions with obvious spectral changes over time were marked by rectangle boxes. To highlight the regions with obvious spectral changes, the result of panel (b) was further modified by a scaled sigmoid function in order to make the values of T-F units with big spectral changes larger while the values of T-F units with minor spectral changes smaller. Panel (c) showed the modified spectral changes which were used as the final weight to produce the proposed loss function. Compared with panel (b), the modified one increased the salience of dynamics-bearing regions which were inherently transient and of low energy, such as the perceptually-important formant transitions and the energy-concentrated frequency bands of consonants.

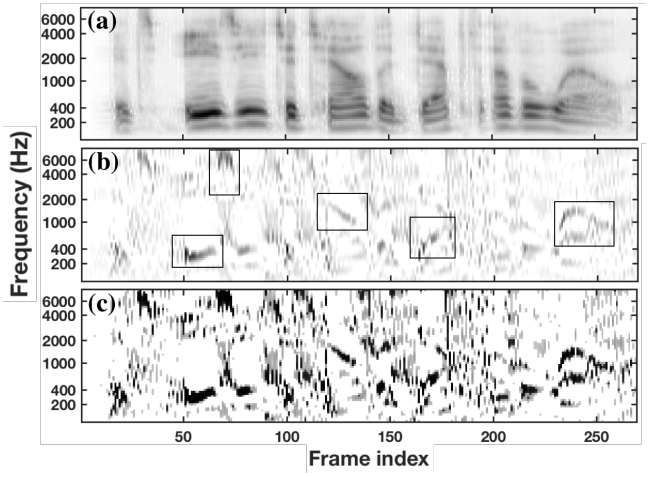

Fig. 1. The upper panel (a) showed the cochleagram of the premixed clean speech. The middle panel (b) and the bottom panel (c) showed the output of SCE and the final weight used to produce the proposed loss function, respectively.

3. SPEECH SEPARATION SYSTEM

The IRM was employed as the training target for supervised speech separation [14, 15]. The IRM was defined as,

$$IRM\,(t,f) = \sqrt{\frac{S\,(t,f)}{S\,(t,f)+N\,(t,f)}} \tag{2}$$

where $S(t,f)$ and $N(t,f)$ denoted premixed speech and noise energy within a T-F unit at time t and frequency f, respectively. The IRM was computed from the 64-channel cochleagrams of the premixed speech and noise with 20 ms frame length and 10 ms frame shift. The separation system was the same as that used in Chen et al., see [16] for details. Acoustic features for each T-F unit were extracted from a mixture and were feed into a DNN to estimate the IRM of each T-F unit. Two typical acoustic features, multi-resolution cochleagram (MRCG) and gammatone feature (GF) were used, which showed promising advantages for DNN-based speech separation when comparing with other features [16]. The DNN had 64 units for GF and 256 units for MRCG in the input layer, 1024 sigmoidal units in each of the 3 hidden layers and 64 sigmoidal units in the output layer. Stochastic gradient descent with a mini-batch size of 256 and the proposed loss function was employed to train the DNN with original MSE loss function as a comparison. As we were mainly concerned with the relative performance for different loss functions, the DNN was chosen as the classifier to simplify and speed up training.

4. EXPERIMENTS

4.1. Experiment setting

The spectral-change-aware loss function was evaluated on the IEEE corpus recorded by a male speaker [17]. There are 72 phonetically balanced lists in the corpus, each with 10 sentences. Sentences from the first 48 lists were used to generate training data, 50 sentences of them set as the validation data. The proposed loss function was tested on sentences from list 49-53. The comparison condition, original MSE loss was tested on sentences from list 53-57.

Six types of nonstationary noise from the NOISEX corpus were used in this work [18]. The noise types included factory floor noise (Factory), speech babble (Babble), jet cockpit noise (Cockpit), destroyer engine room noise (Engine), military vehicle noise (Vehicle), and tank noise (Tank). Each mixture was created from one IEEE sentence and one noise type at -5 dB SNR where the recognition rate of even normal-hearing listeners was less than 50% [19]. Each noise was divided into two parts: the first half was used for training and validation sets and the second half was used for testing.

Mask estimation accuracy, short-time objective intelligibility (STOI) score and SNR gain of unvoiced segments were used as the measurements for evaluating the performance of the current method. Mask estimation accuracy was represented by the mean square error (mse) rate between the estimated IRM and the ground truth. STOI has been used widely as an objective evaluation of speech intelligibility, which is positively correlated with speech recognition scores of human listeners [20]. In addition, SNR gain corresponds to the increase of SNR values for separated speech compared comparing to the input mixtures, specifically for unvoiced segments.

4.2. Results

For the 50 test sentences, mask estimation accuracy, STOI scores and SNR gain for original MSE and proposed loss function were shown in Table 1-3, respectively.

The Overall mse rates were calculated for all T-F units while the $\mathrm{TFunit_{sc}}$ mse rates were calculated only for the T-F units with spectral changes (weight>0), both of which were calculated separately for original MSE and proposed loss function, with GF and MRCG as the input features, respectively. Compared with original MSE loss function, error rates for $\mathrm{TFunit_{sc}}$ decreased for both GF and MRCG at all noise types when the proposed loss function was used, indicating that the estimated IRM for the T-F units weighted by the spectral changes was actually more accurate than before. It was noteworthy that the benefit of proposed loss function was especially significant for GF at the babble noise. However, error rates for Overall were increased for GF feature at some noise types when using the proposed loss function, such as the factory, babble and engine noises, which might be brought from the less accuracy of estimated IRM for other unweighted (weight=0) T-F units. The Overall result for the proposed loss function were decreased for MRCG at all noise types as expectations. Therefore, the weight should be appropriately and precisely determined to achieve a balance between the estimation accuracy of weighted T-F units and other unweighted T-F units.

Table 2 showed that the performance of speech separation using the proposed loss function always revealed higher STOI scores for both GF and MRCG at all noise types. Proposed loss function still performed well for GF at the factory, babble and engine noise types where even at the cost of increased Overall mse rates (see Table 1). However, the benefit for the proposed loss function was small, 0.57% for GF and 0.46% for MRCG on average. Though the masks of T-F units with spectral changes were estimated more accurately, their low percentage in total units resulted in few improvement of sentence STOI scores.

In addition, the performance of separation for unvoiced segments was examined. As a subset of consonants, unvoiced speech consists of unvoiced fricatives, stops and affricates, which show obvious dynamics in spectrum over time and is more susceptible to background noise due to relatively weak energy [21, 22]. Hence, unvoiced speech should be dominant by the T-F units with large spectral changes here. Table 3 showed the SNR gains of unvoiced segments for original MSE and proposed loss function respectively. Compared with original MSE, proposed loss function always performed better on unvoiced segments for both GF and MRCG at all noise types. The average benefit of the proposed loss function was 0.8 dB for GF and 0.71 dB for MRCG for SNR gain.

Table 1. Mean square error (in %) calculated for overall T-F units (Overall) and the T-F units with spectral changes ($\mathrm{TFunit_{sc}}$) separately for six noise types at -5 dB. Boldface indicated best result.

Feature	Loss function	Portion	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
GF	MSE	Overall	**2.83**	**4.93**	**1.46**	1.51	1.45	2.09	**2.38**
		$\mathrm{TFunit_{sc}}$	1.87	6.68	1.15	1.68	1.35	1.69	2.40
	Proposed	Overall	2.87	5.04	1.48	**1.50**	**1.44**	2.09	2.40
		$\mathrm{TFunit_{sc}}$	**1.73**	**4.53**	**1.00**	**1.27**	**1.25**	**1.54**	**1.89**
MRCG	MSE	Overall	2.53	4.79	1.12	1.14	1.12	1.61	2.05
		$\mathrm{TFunit_{sc}}$	3.20	6.89	1.18	1.34	1.70	1.84	2.69
	Proposed	Overall	**2.47**	**4.78**	**1.11**	1.14	**1.11**	**1.60**	**2.04**
		$\mathrm{TFunit_{sc}}$	**2.47**	**6.45**	**1.04**	**0.77**	**1.61**	**1.65**	**2.33**

Table 2. STOI scores (in %) for six noise types at -5 dB. Boldface indicated best result.

Feature	Loss function	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
GF	MSE	65.68	65.74	72.22	70.50	78.14	71.92	70.70
	Proposed	**66.29**	**66.27**	**72.86**	**71.14**	**78.73**	**72.33**	**71.27**
MRCG	MSE	67.03	64.96	74.11	73.15	79.84	73.68	72.13
	Proposed	**67.57**	**65.38**	**74.64**	**73.64**	**80.31**	**73.98**	**72.59**

Table 3. SNR gain (in dB) of unvoiced segments for six noise types at -5 dB. Boldface indicated best result.

Feature	Loss function	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
GF	MSE	15.23	8.59	16.91	16.57	17.24	16.72	15.21
	Proposed	**16.01**	**16.01**	**17.34**	**17.43**	**18.07**	**17.46**	**16.01**
MRCG	MSE	15.94	8.97	17.03	16.79	17.96	16.83	15.59
	Proposed	**16.67**	**9.68**	**17.69**	**17.69**	**18.57**	**17.58**	**16.30**

5. DISCUSSION

In this study, we proposed a spectral-change-aware loss function during the training of DNN-based speech separation in order to improve speech intelligibility. The initial results suggested this method could improve the SNR gain for unvoiced segments and the STOI scores. It should be noticed that enhancement of unvoiced segments could achieve more benefit for HI listeners than normal hearing listeners [23], hence, it could be expected that the performance improvement could be increased if subjective speech test is conducted for HI listeners.

When the weights were assigned to T-F units, they were derived by spectral change evaluation. In the panel (c) of Fig. 1, although some prominent dynamic features were captured with the SCE processing, some "noise" still existed. It remains unclear to what extent the SCE could be manipulated, and how this manipulation could impact the final performance of speech segregation. It is worthy to adjust the SCE processing for assigning more accurate and effective weights in future work.

The proposed loss function was an early attempt to realize a discriminative training since it was just simply addressed by adding a constraint to the original MSE loss function using the spectral changes evaluation. The experimental results showed that the performance of speech separation could be further improved by making training networks to put more efforts into some important regions. Therefore, it is possible to introduce attention mechanism for DNN, which has been applied a lot in many intelligent systems, e.g. image captioning and machine translation [24], to simulate the human beings perceptual property on the sensitivity of spectral changes.

6. CONCLUSION

In this study, we proposed a spectral-change-aware loss function in order to realize a discriminative training for DNN-based speech separation. The experimental results showed speech intelligibility could be improved in adverse background environments when using the proposed loss function even at the cost of less accuracy of estimated IRM for some conditions. In addition, the proposed loss function produced a better separation performance on unvoiced segments.

7. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 11590773, 61771023, and 61473008), and and a research grant by SONOVA Shanghai, China.

8. REFERENCES

[1] B. C. J. Moore, *Cochlear hearing loss: physiological, psychological and technical issues*, John Wiley & Sons, 2007.

[2] H. Dillon, *Hearing aids*, Hodder Arnold, 2008.

[3] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[4] K. Han and D. L. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[5] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[6] Q. Summerfield, M. Haggard, J. Foster, and S. Gray, "Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect," *Perception & Psychophysics*, vol. 35, no. 3, pp. 203–213, 1984.

[7] A. J. Watkins, "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortiona," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2942–2955, 1991.

[8] B. C. J. Moore, "Temporal integration and context effects in hearing," *Journal of Phonetics*, vol. 31, no. 3-4, pp. 563–574, 2003.

[9] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," *The Journal of the Acoustical Society of America*, vol. 69, no. S1, pp. S116–S116, 1981.

[10] D. G. Jamieson and D. E. Morosan, "Training non-native speech contrasts in adults: Acquisition of the english//-/θ/contrast by francophones," *Perception & psychophysics*, vol. 40, no. 4, pp. 205–215, 1986.

[11] J. Chen, T. Baer, and B. C. J. Moore, "Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2987–2998, 2012.

[12] J. Chen, T. Baer, and B. C. J. Moore, "Effect of spectral change enhancement for the hearing impaired using parameter values selected with a genetic algorithm," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2910–2920, 2013.

[13] J. Chen, B. C. J. Moore, T. Baer, and X. Wu, "Individually tailored spectral-change enhancement for the hearing impaired," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. 1128–1137, 2018.

[14] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[15] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind source separation*, pp. 349–368. Springer, 2014.

[16] J. Chen, Y. Wang, and D. L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[17] E. H. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[18] A. Varga and H. 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.

[19] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[21] P. Ladefoged and S. F. Disner, *Vowels and consonants*, John Wiley & Sons, 2012.

[22] K. N. Stevens, *Acoustic phonetics*, vol. 30, MIT press, 2000.

[23] J. Chen, T. Baer, and B. C. J. Moore, "Effects of enhancement of spectral changes on speech quality and subjective speech intelligibility," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Lukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.