INTEGRATING SPECTROTEMPORAL CONTEXT INTO FEATURES BASED ON AUDITORY PERCEPTION FOR CLASSIFICATION-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, which has been a challenging task for decades, especially at low signal-to-noise ratios (SNRs), can be cast as a classification problem. In such adverse acoustic environment, extracting robust features from noisy mixtures is crucial for successful classification. In the past studies, features representing temporal dynamics, known as delta features, have been widely used. Combining basic features with their deltas yields better speech separation results than using basic features alone. In this study, the commonly used delta feature was modified according to the characteristics of auditory perception, which included auditory processing on spectral change and spectral contrast. Therefore, we proposed a feature which integrated spectrotemporal context via replacing the commonly used delta feature by spectral change feature and spectral contrast feature. Experimental results showed that the proposed feature could produce better speech segregation performance than the common delta feature.

Index Terms— Speech separation, spectral change feature, spectral contrast feature

1. INTRODUCTION

Monaural speech separation is regarded as a challenging task, especially in low signal-to-noise (SNR) conditions. In recent studies, the ideal binary mask (IBM) was suggested as the computational objective of the speech separation task [1]. The IBM is a time-frequency (T-F) binary mask, constructed from premixed target and noise. A mask value 1 for a T-F unit indicates that the SNR within the unit exceeds a threshold (target dominant), and 0 otherwise (noise dominant). Therefore, the speech separation can be formulated as a classification problem [2]. The evaluation on this approach showed it could improve speech intelligibility for human listeners, including hearing-impaired people [3,4].

The design of classifier and robustness of features extracted from mixtures mainly determine the performance of such classification-based speech separation. This work focused on robust features for classification. In the speech separation community, many acoustic features have been explored, such as amplitude modulation spectrogram (AMS), perceptual linear prediction (PLP), gammatone feature (GF) and more recently multi-resolution cochleagram feature (MRCG), each having its own advantages. In this study, we aimed to explore a robust feature according to characteristics of auditory perception due to the excellent performance on speech recognition in noise for human beings.

The auditory system, like other perceptual systems, is especially sensitive to abrupt changes in stimuli [5,6,7,8]. Several findings have suggested that most important information in speech is carried in spectral changes over time, rather than in static spectral per se. It has been reported that stimuli with dynamic spectral changes at their onsets leads to better identification of articulation place [9]. Hence, the ability to detect spectral changes over time may be beneficial for speech separation from noisy mixtures.

In addition, the ability for humankind to recognize speech in noise almost depends partly on auditory functions such as frequency selective, which refers to the ability of auditory system to resolve a complex sound into its frequency components [10,11]. Some perceptual-important components, like the formant frequencies, will be smeared and not sufficiently prominent in the presence of background noise since the noise fills in the valley between spectral peaks of target speech, resulting in reduced spectral contrasts and further reduced frequency selective.

Due to the importance of temporal changes and spectral contrasts for human speech recognition in background noise, we proposed a feature that integrated the above spectrotemporal context via combining basic features with their spectral change and spectral contrast. In previous studies, spectrotemporal context has already been considered and used in the speech separation task. Kim et al. proposed a combined feature where in addition to the basic AMS feature, delta features were also included to capture feature variations across time and frequency, leading to a better speech separation result [3]. MRCG feature, proposed by Chen et al., was meant to embody the spectrotemporal context systematically and resulted in the best separation performance among many popular features [12]. However, the operations like difference and average were performed on existing features in the above works, which were not linked with the characteristics of auditory perception, including auditory processing on spectral changes and spectral contrasts. Therefore, spectral change feature and spectral contrast feature were extracted and added in our proposed feature as spectrotemporal context in order to facilitate

speech separation in consistence with mechanism of auditory processing on speech.

2. FEATURE DESCRIPTION

2.1. Spectral change feature

Temporal delta feature was first extracted to capture the basic feature variations across time, which was denoted by,

$$\Delta g_N(n,k) = g(n,k) - g(n-1,k), \quad n = 2, \dots, N, \quad (1)$$

where $g(n,k)$ was the feature extracted from a specific T-F unit with a time index *n* and a sub-band index *k*. *N* was the total number of frames. Two typical acoustic features, MRCG and GF were used here separately to produce $g(n,k)$ as they showed promising advantages for DNN-based speech separation when comparing with other features [12].

Then, our previously proposed spectral change evaluation (SCE) method [13] was applied to $\Delta g_N(n,k)$ to derive the final spectral change feature $change(n,k)$. Specifically, temporal delta feature $\Delta g_N(n,k)$ was next convoluted with a Difference of Gaussian (DoG) function to produce a spectral change function (SCF), in order to remove minor irregularities in the delta feature as well as to emphasize the difference between feature peaks and valleys.

To take the influence of preceding frames into account, a Gain function, $Gain(n,k)$ for a certain frame *n* and sub-band *k,* was defined by a weighted average of the SCF over several preceding frames *m* with a weight ($\xi < 1$) that progressively declined for frames that were earlier in time than the current frame,

$$Gain(n,k) = \frac{SCF_{n,k} + \xi SCF_{n-1,k} + \cdots + \xi^m SCF_{n-m,k}}{1 + \xi + \cdots + \xi^m}, \quad (2)$$

Then, the Gain function was scaled by a factor *S*, which was used to produce a controllable spectral change feature $change(n,k)$.

2.2. Spectral contrast feature

spectral contrast feature was derived by applying a typical spectral contrast evaluation (SE) method proposed by Baer et al. to $g(n,k)$ [14]. Until producing the Gain function, the processing steps were the same as the SCE except that the input was $g(n,k)$ here, rather than $\Delta g_N(n,k)$. Gain function here was derived according to the idea that for a given frame and sub-band where the spectral contrast function (SF) was positive, $g(n,k)$ was increased in value; where the SF was negative, $g(n,k)$ was decreased in value. Hence, the Gain function was denoted by,

$$Gain(n,k) = log\{abs(SF) + 1\} \times sign(SF), \quad (3)$$

The value of $Gain(n,k)$ was then scale by a certain factor *E* to produce a controllable spectral contrast feature $contrast(n,k)$ here.

The overall proposed feature vector was given by,

$$[g(n,k), change(n,k), contrast(n,k)], \quad (4)$$

we named the proposed feature as GF_proposed or MRCG_proposed according to different $g(n,k)$.

2.3. Analysis of the proposed feature

In the proposed feature, there were three components. $g(n,k)$ contained the local information embedded in each T-F unit while spectral change feature captured dynamic cues over time and spectral contrast feature highlighted the important components across frequency. The latter two features indicated the spectrotemporal context.

A visualization of each feature component extracted from premixed clean speech was given in Fig.1. Spectral change feature in panel (a) encoded not only the abrupt changes at onsets of syllables, but also the successive changes over time, such as the perceptual-important formant transitions with relatively low energy, which were marked with black rectangular boxes. However, temporal delta feature in panel (b) mainly revealed the abrupt changes at onsets almost without successive changes, which were actually the important cues for auditory perception on speech temporal patterns [8]. The successive spectral changes could be captured in (a) because of the Gain function in SCE processing taking the influence of spectral changes in preceding frames into consideration. Spectral contrast feature in panel (c) especially highlighted the feature contrasts between the formant frequencies and other sub-bands, and made the feature within the high frequency range more concentrated in certain bands, such as the third and the fourth formant frequencies, while spectral delta feature in panel (d) remained "noise" within high frequency bands and the formant frequencies which played an important role in spectral processing for auditory system was not sufficiently prominent. To depict the SE processing across frequency in details, panel (f) showed the spectral contrast feature for a certain frame on frequency scale. With GF as reference, SE processing mainly increased the feature values within the formant frequencies and decreased the feature values within their neighboring bands through Gain function, meanwhile it smoothed the subtle spectral dynamics to make the contrasts more prominent through SF operation.

3. EXPERIMENTS

3.1. Speech separation system

In the classification-based speech separation, the ideal binary mask (IBM) is used to guide the neural network training. The IBM was calculated by a 64-channel gammatone filterbank with 20 ms frame length and 10 ms frame shift. The local SNR threshold was set to -10 dB. As for evaluation, our goal was to reveal the relative performance of various spectrotemporal features, hence the classifier was fixed to a deep neural network (DNN) to simplify and speedup training [15]. The feature evaluation framework was the same as that used for MRCG evaluation in Chen et al. [12]. Acoustic features for each frame were extracted from a mixture and were feed into a DNN to estimate each frame of the IBM. HIT-FA rate and the short-time objective intelligibility (STOI)

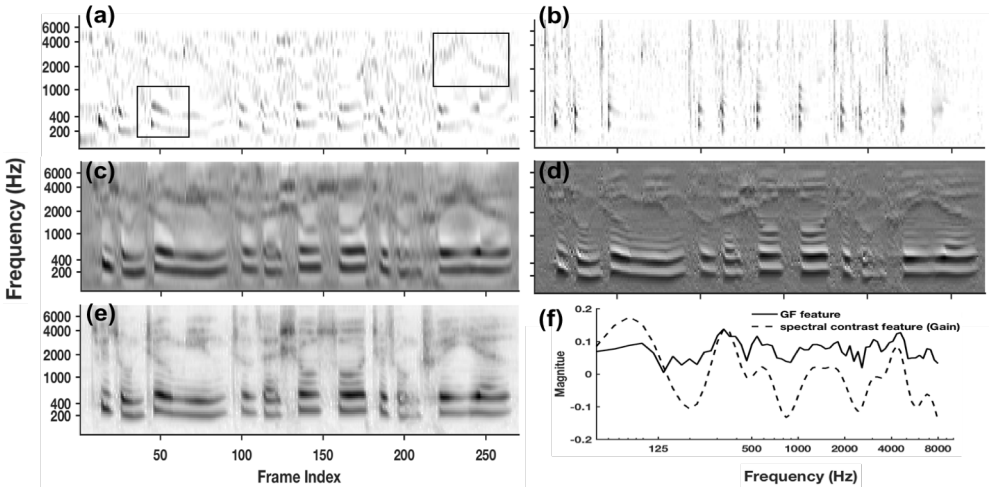

Fig. 1. Visualization of each feature component extracted from premixed clean speech. (a) spectral change feature; (b) temporal delta feature; (c) spectral contrast feature; (d) spectral delta feature; (e) GF feature; (f) spectral contrast feature relative to GF for a certain frame on frequency scale.

were included as the evaluation measurements.

3.2. Experiment setting

The experiment setting was also consistent with that in Chen et al. To exclude any mistake from network training, we used the DNN toolbox supplied by those authors ('http://web.cse.ohio-state.edu/pnl/software.html').

Mixtures were created using the IEEE corpus recorded by a male speaker [16] and six types of nonstationary noise (see Table 1) from the NOISEX corpus [17]. 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% and could be regarded as an adverse environment [3]. To create the training set, we used 480 IEEE sentences and the first half of each noise. As for the test set, we used another 50 IEEE sentences and the second half of the noises. We set aside 50 mixtures from the training set as a cross validation set for early stopping.

3.3. Comparison feature

The most commonly used spectrotemporal context was the delta across time and frequency [3], which was used as the comparison condition here. The overall comparison feature also consisted of three components: $g(n,k)$ representing GF or MRCG, temporal delta feature $\Delta T_N(n,k)$ and spectral delta feature $\Delta S_K(n,k)$, denoted as GF_delta or MRCG_delta,

$$[g(n,k), \Delta T_N(n,k), \Delta S_K(n,k)], \quad (5)$$

where

$$\Delta T_N(n,k) = g(n,k) - g(n-1,k), n = 2, \dots, N,$$
$$\Delta S_K(n,k) = g(n,k) - g(n,k-1), k = 2, \dots, K, \quad (6)$$

where N was the total number of frames. The number of total subbands K, was set to 64 in this work.

3.4. Results

For the 50 test sentences, HIT-FA rate and STOI scores of each feature were shown in Table 1 and Table 2, respectively. Boldface indicated the best result for each noise type.

As shown in Table 1, when GF and MRCG were combined with spectrotemporal context, HIT-FA rates were always higher compared with using GF or MRCG alone. The improvement was more significant for GF than for MRCG since MRCG itself had already embodied the spectrotemoral context to some extent. Generally, the proposed feature achieved better average performance compared with the delta feature for both GF and MRCG. Except for the engine noise, the proposed feature performed the best for all the noise types.

The performance situation for STOI scores in Table 2 was similar with that for HIT-FA rates, except that the performance of the proposed feature in STOI scores was worse than that of the delta feature for the tank noise while this situation appeared in HIT-FA rates for the engine noise. It was noteworthy that adding the delta feature did not always improve speech intelligibility, which was indicated for the factory for both GF and MRCG and for the babble noise and vehicle noise for MRCG. In addition, for the babble noise, GF_proposed feature even outperformed MRCG which showed the best performance for all noise types in earlier work [12].

4. DISCUSSION

Table 1 showed adding spectrotemporal context could always result in better HIT-FA performance than using basic features alone, no matter applying delta method or the proposed method. However, results for STOI scores in Table 2 was not always the case, where adding delta feature did not produce

Table 1. HIT-FA (%) for six noise types at -5 dB for IEEE sentences. Boldface indicated best result.

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
GF	52.61	46.06	69.88	67.96	70.43	62.5	61.57
GF_delta	55.30	47.42	**72.45**	70.89	72.66	65.46	64.03
GF_proposed	**55.36**	**48.53**	72.23	**71.76**	**73.01**	**66.14**	**64.59**
MRCG	59.35	49.82	74.96	75.01	75.03	69.38	67.26
MRCG_delta	60.87	50.77	**76.04**	75.98	75.49	71.15	68.38
MRCG_proposed	**60.89**	**51.53**	76.03	**76.26**	**75.71**	**71.86**	**68.71**

Table 2. STOI (%) score for six noise types at -5dB for IEEE sentences. Boldface indicated best result.

Feature \ Noise	Factory	Babble	Engine	Cockpit	Vehicle	Tank	Average
GF	64.46	65.47	68.96	67.98	77.15	70.84	69.14
GF_delta	64.00	64.51	68.99	68.66	77.37	**71.93**	69.24
GF_proposed	**64.89**	**65.95**	**69.30**	**68.95**	**77.64**	71.85	**69.76**
MRCG	64.88	65.13	70.26	70.22	78.33	72.65	70.25
MRCG_delta	64.34	64.72	70.64	70.64	78.11	**73.36**	70.30
MRCG_proposed	**65.19**	**65.18**	**70.78**	**70.88**	**78.87**	73.26	**70.69**

higher STOI scores for some conditions, while adding the proposed feature still always improved the STOI performance. The possible reason might be that, HIT-FA is more associated with classification accuracy which can be theoretically increased when extra useful spectrotemporal information is added to input features. While STOI is an objective measurement directly correlated with human speech intelligibility which is expected to be improved when spectrotemporal context exactly includes crucial cues that have been proven to be helpful for human speech perception in noise. Compared with the delta feature which was derived by just simply calculating the difference of features across time and frequency, the proposed one considered the characteristics of human auditory perception on speech, to focus on perceptual-important components, e.g., formant transitions and formant frequencies.

Table 3. Separation performance for cockpit noise at 0 dB and 5 dB, respectively.

Feature	HIT-FA		STOI	
	0 dB	5 dB	0 dB	5 dB
GF	66.08	76.76	77.69	86.85
GF_delta	68.99	78.92	77.91	86.79
GF_proposed	69.37	79.05	77.98	86.94
MRCG	76.54	84.12	78.84	87.72
MRCG_delta	76.90	84.37	78.92	87.55
MRCG_proposed	77.18	84.54	79.60	88.08

To evaluate the generalization of the proposed feature to other SNR conditions, separation performance for additional 0 dB and 5 dB was shown in Table 3 only for cockpit noise where the proposed feature produced the most benefit due to the space limitation. However, benefit for the proposed feature decreased with the improvement of SNR values. Since the basic GF or MRCG could capture enough information when SNR was high, less benefit might be provided by extra spectrotemporal information from both the proposed and delta features, compared with low SNR conditions. The results indicated that the more adverse the acoustic environment was, the more important spectrotemporal context turned to be in speech separation task. Meanwhile, the proposed feature showed the most benefit at -5 dB which was a relatively adverse noisy condition, indicating its promising advantage on robustness in resistance to background noise.

5. CONCLUSION

In this study, we explored a feature that integrated spectrotemporal context inspired by characteristics of auditory perception. The proposed feature was evaluated relative to other features which also included the spectrotemporal context for classification-based speech separation at -5 dB SNR. Experimental results showed that the proposed feature outperformed the delta feature for most noise types in terms of HIT-FA and STOI and revealed the promising advantage on robustness in resistance to background noise.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

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

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

[3] 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.

[4] E. W. Healy, S. E. Yoho, Y. Wang, and D. 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.

[5] 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.

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

[7] A. J. Watkins and S. J. Makin, "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," *Journal of the Acoustical Society of America,* vol. 99, no. 6, pp. 3749-3757, 1996.

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

[9] D. Kewley-Port, D. B. Pisoni, and M. Studdert-Kennedy, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *The Journal of the Acoustical Society of America,* vol. 73, no. 5, pp. 1779-1793, 1983.

[10] J. M. Festen and R. Plomp, "Relations between auditory functions in impaired hearing," *Journal of the Acoustical Society of America,* vol. 73, no. 2, pp. 652-662, 1983.

[11] B. C. Moore, "Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids," *Ear Hear,* vol. 17, no. 2, pp. 133-161, 1996.

[12] J. Chen, Y. Wang, and D. 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.

[13] 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," *Journal of the Acoustical Society of America,* vol. 131, no. 4, pp. 2987-2998, 2012.

[14] T. Baer, B. C. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," *Journal of Rehabilitation Research & Development,* vol. 30, no. 1, pp. 49-72, 1993.

[15] Z. Jin and D. 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.

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

[17] A. Varga and H. J. 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.