INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A Mask Estimation Method Integrating Data Field Model for Speech Enhancement

Xianyun Wang[1], *Changchun Bao*[1], *Feng Bao*[2]

[1]Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing, China, 100124
[2]Department of Electrical and Computer Engineering, The University of Auckland, Auckland 1142, New Zealand

b201402001@emails.bjut.edu.cn, baochch@bjut.edu.cn, fbao026@aucklanduni.ac.nz

Abstract

In most approaches based on computational auditory scene analysis (CASA), the ideal binary mask (IBM) is often used for noise reduction. However, it is almost impossible to obtain the IBM result. The error in IBM estimation may greatly violate smooth evolution nature of speech because of the energy absence in many speech-dominated time-frequency (T-F) units. To reduce the error, the ideal ratio mask (IRM) via modeling the spatial dependencies of speech spectrum is used as an optimal target mask because the predictive ratio mask is less sensitive to the error than the predictive binary mask. In this paper, we introduce a data field (DF) to model the spatial dependencies of the cochleagram for obtaining the ratio mask. Firstly, initial T-F units of noise and speech are obtained from noisy speech. Then we can calculate the forms of the potentials of noise and speech. Subsequently, their optimal potentials which reflect their respective distribution of potential field are obtained by the optimal influence factors of speech and noise. Finally, we exploit the potentials of speech and noise to obtain the ratio mask. Experimental results show that the proposed method can obtain a better performance than the reference methods in speech quality.

Index Terms: Data field, CASA, Ratio mask, Speech enhancement

1. Introduction

In a natural environment, a target sound, such as speech, is usually deteriorated by noise. Sound separation from noise is one of the key approaches for removing or attenuating background noise in speech processing. Researches on human auditory perception inspire one promising approach which is called CASA [1-8]. The main computational goal of the CASA-based method is the estimation of the IBM [2-3]. The IBM is a two-dimension 0–1 matrix along time and frequency index which classifies all the T-F units into reliable and unreliable classes. Unreliable class consists of the units in which noise energy exceeds the speech, while reliable class consists of the rest.

Once the IBM is estimated, the clean speech can be extracted accurately from noisy speech. However, it's almost impossible to estimate the IBM with one-hundred-percent accuracy in practice. The error in IBM estimation may result in many abrupt changes, which could greatly violate smooth evolution nature of speech signal. In order to address this problem, a smooth binary mask (named as the IRM) based on the spatial dependencies of speech spectrum is used to replace the IBM for improving the spectral continuity because the predictive ratio mask with the spectral correlation is less sensitive to the error generated from IBM estimation than the predictive binary mask [7-8]. In [7], the theory of discriminative random field (DRF) was introduced into the IBM estimation for voiced speech separation and showed that the CASA-based techniques may benefit from the spatial dependencies in the DRF framework. And in [8], the high temporal correlation and smooth evolution existing in the T-F representations of speech and noise have been presented by Markov random field (MRF). In this system, it employs an MRF prior to model the spatial dependencies of speech spectrum so that the local temporal correlation between two units of speech and noise can be taken into account for generating a ratio mask (RM) based on cochleagram. In addition, the spatial dependency based on MRF model is also discussed for speech enhancement in Fast Fourier Transformation (FFT) domain [9]. These studies have suggested that the information encapsulated in the above attributes of the spatial dependencies of speech spectrum could prove a very helpful improvement in enhancing speech quality.

In this paper, we follow the framework about the spatial dependencies of speech spectrum to design a smoother speech cochleagram. A major contribution of this work is the introduction of a Gaussian DF model in the cochleagram. Comparing with the previous methods considering the spatial dependencies of speech spectrum, with the DF model, we can easily obtain the spatial dependencies with enough information and need fewer model parameters for constructing the T-F relationships of speech and noise. Thus we could provide a simple model to consider the spatial dependencies of the T-F units. Firstly, we obtain the initial estimations of noise and speech by a binary mask estimation for voiced speech portion. And we give the initial estimations of noise and speech from noisy potential in unvoiced portion. When the initial estimations of noise and speech are given, the T-F spaces of speech and noise can be constructed for obtaining the form of the potential of each T-F unit. Since each potential can be viewed as the superposition of the interaction of all cochleagrams within the space, we can obtain a local temporal correlation with more information of speech spectrum. Subsequently, each optimal potential can be determined by the influence factor optimization. Finally we could obtain a smooth mask estimation based on the optimal potentials of speech and noise.

The rest of this paper is organized as follows. In section 2 a brief review of the DF theory is described in speech signal.

http://dx.doi.org/10.21437/Interspeech.2017-271

The detail of the proposed method is explained in section 3. In section 4 we present our simulation examples, and we conclude the paper in section 5.

2. Data field of speech signal

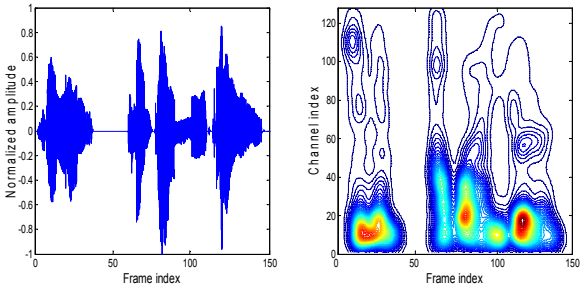
The idea of field was first proposed in 1837 by an English physicist, Michael Faraday, in which the noncontact interaction between particles was described. Inspired by the knowledge, Deyi Li [10] introduced a virtual cognitive field named "data field" to model the interaction of particles in the data space, which could describe the relationship among data points and reveal the general characteristics of the underlying data distribution.

According to the knowledge of the field theory, the Gaussian potential is the simplest way to describe a DF, so it is also used in our work. Given a space Ω produced by a set of data objects $A=\{ x_1, x_2,\ldots, x_n \}$, the Gaussian potential at any point $z \in \Omega$ can be calculated as,

$$\varphi_A(z) = \sum_{i=1}^{n} \varphi_i(z) = \sum_{i=1}^{n} m_i \exp\left(-\left(\frac{\| z - x_i \|}{\sigma} \right)^2 \right) \qquad (1)$$

where $\|z - x_i\|$ is the Euclidean distance between object x_i and point z; the strength of interaction $m_i \geq 0$ is the mass of data object x_i ($i = 1,2,\ldots,n$). $\sigma \in (0,+\infty)$ is the influence factor that indicates the range of interaction.

In this paper, the DF is employed to model the spatial dependencies of the T-F units by exploiting the interaction of all data objects through space. Figure 1 shows a transformation of speech signal to DF, where Figure 1(a) is the speech waveform containing five main "virtual energy blocks of speech". In Figure 1(b) the equipotential lines viewed as a representation of speech DF exhibit five local maximums of the associated potential field, which can be nearly considered as a DF produced by five main "virtual energy blocks". Where the number of channels is 128.



(a) Speech waveform (b) Equipotential lines

Figure 1: *An example of the equipotential lines of speech DF*

As shown in Figure 1, the distribution of DFs can better model the profile of speech energy, and the energy concentration zones could correspond to some acoustic units, such as phonemes of speech. Due to the speech potential is generated from the virtual energy blocks, all the T-F units of speech could be attracted to converge each other and exhibit certain features of self-organization. In the CASA system, a main motivation is to use an auditory filterbank to mimic cochlear filtering. The cochlea filtering introduces inherent correlation between channels. Due to the spectral smearing phenomenon which is a known effect of the windowing process, some correlation between adjacent spectral components is also introduced. So the spatial dependencies of the T-F units should be taken into account.

3. The proposed approach

The proposed algorithm considers all T-F units as interacting objects by means of the DF in the cochleagram. After the cochleagram of noisy speech is calculated, we attempt to estimate an ideal ratio mask with speech correlation for synthesizing the speech. Figure 2 shows a block diagram of the proposed method. Firstly, the T-F representation of the noisy signal is obtained by the peripheral analysis. Then initial energy estimations of noise and speech in each T-F unit need to be obtained. Then, the DF is used to model the dependencies of initial energy estimations of speech and noise so that the corresponding optimal potentials of speech and noise can be obtained. Here, the DF contains two stages: influence factor optimization and potential calculation. The aim of the influence factor optimization is to make the distribution of potential field as consistent with the underlying distribution of speech or noise energy as much as possible. Given the optimal factor σ, the second stage can obtain the optimal potential of each T-F unit. Finally, in order to restore the clean speech signal, the optimal potentials of speech and noise, which are also viewed as the cochleagram with the information of their neighbors, are used to generate a smoother speech cochleagram.

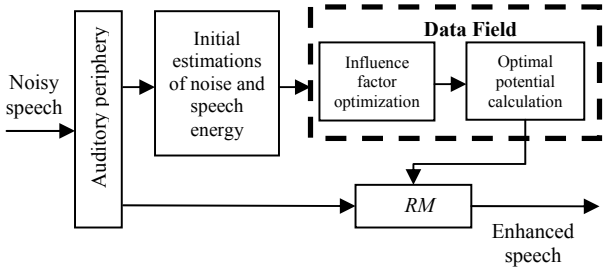


Figure 2: *Block diagram of the proposed method*

3.1. Initial estimations of noise and speech energy

The noisy speech signal $y(t)$ is decomposed into T-F domain firstly by a gammatone filter with 128 channels. Then, the response of each channel is divided into 20 *ms* time frames with 10 *ms* overlap. The resulting T-F representation is called cochleagram.

Let $Y(i,l)$, $X(i,l)$ and $D(i,l)$ denote the energy of noisy speech, speech and noise at the i^{th} T-F unit of the l^{th} frame, respectively. In this paper, we give the initial estimations of speech and noise energy based on a binary mask $M(i,l)$ in voiced portion:

$$[E_X(i,l), E_D(i,l)] = \begin{cases} [Y(i,l), 0] & M(i,l) = 1 \\ [0, Y(i,l)] & M(i,l) = 0 \end{cases} \qquad (2)$$

For the voiced speech portion, we estimate the binary mask with a typical method proposed by Hu and Wang [3].

In CASA framework, unvoiced speech has few discussions due to its relatively weak energy and lack of harmonic structure. In this work, initial T-F units of noise and speech in unvoiced portion are generated from noisy potential. In this process, firstly, we just aim to generate a simple way to obtain the noisy energy in unvoiced portion. Therefore, the unvoiced speech portion is extracted based on thresholding

used in [6] and no further classification of the reliable/unreliable units is required. Then every optimal potential in noisy T-F space constructed by the noisy energy is calculated. Finally using the optimal noisy potential, we give a method for generating initial noise and speech energies as follows.

1) By using the concept of minimum statistics exploited by R. Martin [11], given the optimal noisy potential $\varphi_Y(i,l)$, we smooth the noisy potential as follows.

$$\tilde{\varphi}_Y(i,l)=\alpha\tilde{\varphi}_Y(i,l-1)+(1-\alpha)\left|\varphi_Y(i,l)\right| \quad (3)$$

where, α=0.9 is the smoothing parameter.

2) Find the minimum in M+1 consecutive noisy potential $\tilde{\varphi}_Y(i,l)$ as the estimate of the noise potential. Here we typically use M=6.

$$\tilde{\varphi}_{Y\min}(i,l)=\left\{\tilde{\varphi}_Y(i,l-M+1),...,\tilde{\varphi}_Y(i,l),\tilde{\varphi}_Y(i,l+1)\right\} \quad (4)$$

3) Compute the final estimate of noise potential by multiplying a bias correction factor $B_{\min}$ which is typically set to 1.4, we have

$$\tilde{\varphi}_{Y\_d}(i,l)=B_{\min}\tilde{\varphi}_{Y\min}(i,l) \quad (5)$$

4) The initial estimations of speech and noise energy in unvoiced portion are represented by using binary T-F masks.

$$[E_X(i,l),E_D(i,l)]=\begin{cases}[Y(i,l),0] & \tilde{\varphi}_Y(i,l)>\tilde{\varphi}_{Y\_d}(i,l)\\ [0,Y(i,l)] & else\end{cases} \quad (6)$$

In this paper, the aim of our work is to show the important influence of the spatial dependencies of speech spectrum on the enhancing speech quality, so we only select a few constants (alpha, M, B_{min}) to get initial estimations of speech and noise for unvoiced speech.

3.2. Definition of the potential field in the cochleagram

Data potential means the work done to move a unit object from some position to the reference point, it is clear that potential is only a function of the coordinates, regardless of whether the objects exist. The overall distribution of data potential field is usually represented by equipotential lines.

Given M+1 frames of noisy T-F energy $Y(i,l)$, initial T-F energies of speech and noise, $E_X(i,l)$ and $E_D(i,l)$, we can form their $C*(M+1)$ two-dimensional T-F space Ω_Y, Ω_X, and Ω_D, respectively. Where C is the number of channel, $0\le i\le C-1, 0\le l\le M$. For speech signal, we consider each T-F energy of speech as the mass $m_{i,l}^X$ of associated data object $O(i,l)$ in T-F space of the speech energy. The interaction of all the T-F points will generate a DF through speech space. The form of the Gaussian potential at any point z of the speech space Ω_X can be calculated as

$$\varphi_X(z)=\sum_{i=0}^{C-1}\sum_{l=0}^{M}m_{i,l}^X\exp\left(-\left(\frac{\|z-z_{i,l}\|}{\sigma_X}\right)^2\right) \quad (7)$$

Likewise for the noise and noisy speech signals.

3.3. Influence factor optimization for potential field

Once the form of potential is fixed, the distribution of the associated DF needs to be determined by the influence factor σ. If σ is too small, the range of interaction between the T-F points is very short, that is, each T-F point can only affect few points around it. If σ is very large, each T-F point will influence some points away from it. The extreme is that the potential at the location of each T-F unit only approximates normalization constant. Obviously, the potential obtained by an inappropriate σ cannot produce a meaningful estimate of the underlying distribution of speech energy. Thus, the choice of σ should make the distribution of potential field as consistent with the underlying distribution of original data as much as possible.

Since entropy is a useful measure to show the randomness of data in information theory, potential entropy can be used to measure the uncertainty about the distribution of potential field. Here, we can use the potential entropy in speech enhancement task based on the following aspects:

a) We know that the statistical characteristics of speech and noise are highly skewed, for instance, super-Gaussian distribution is more suitable for the statistical characteristics of speech signal, while Gaussian distribution is more suitable for noise signal. So the asymmetrical distribution could help to distinguish speech and noise.

b) Unvoiced speech has relatively the more characteristic of energy concentration than the noise, so the onset and offset cues are often used to analyze unvoiced speech. The different characteristic between noise and unvoiced speech could lead to a fact that their influence radius of T-F energy has a big gap. So we may separate noise from noisy speech in unvoiced frames by adjusting an appropriate factor σ.

If $\varphi_X(i,l)$ is the potential at the position of object $O(i,l)$, its potential entropy can be defined as

$$H=-\sum_{i=0}^{C-1}\sum_{l=0}^{M}\frac{\varphi_X(i,l)}{\Phi_X}\log\left(\frac{\varphi_X(i,l)}{\Phi_X}\right) \quad (8)$$

where $\Phi_X=\sum_{i=0}^{C-1}\sum_{l=0}^{M}\varphi_X(i,l)$ is the normalization factor. Note that the frequency interval represented by the channel is not uniform and this may affect the performance of speech restoration. So we consider the center frequency of each channel as frequency bin index through space in this paper. The optimal choice of σ is given as follows,

$$\sigma^*=\arg\min_{\sigma}H(\sigma) \quad (9)$$

In this paper, a golden section search method in [10] is adopted to obtain the optimal influence factors.

3.4. Ideal Ratio mask estimation

As a result, given the optimal potentials of speech and noise, $\varphi_X^*(i,l)$ and $\varphi_D^*(i,l)$, the normalized ratio is used as a mask value as follows,

$$RM(i,l)=\frac{\varphi_X^*(i,l)}{\varphi_X^*(i,l)+\varphi_D^*(i,l)} \quad (10)$$

Using the estimated IRM, it is straightforward to resynthesize the enhanced speech signal from the output of the gammatone filterbank [4].

4. Experimental results

In this section, the clean speech is extracted from NTT database and the sampling rate of speech signal is 8 kHz in test set. Four different types of background noises are chosen from NOISEX-92 noise databases, including white noise

(White), babble noise (Babble), f16 noise (F16) and factory noise (Factory). The clean speech test set is degraded by adding these noise types in five input SNR levels, which are defined as -10dB, -5dB, 0dB, 5dB and 10dB. Since the main aim of this paper is to show the effectiveness of the spectral dependencies based on the DF model for improving the speech quality, we select another classical method [8] with the spatial dependencies of speech spectrum as reference method (named as MRF). Meanwhile a binary mask method with the typical Hu and Wang model [3] combined with spectral subtraction method [12] is considered as reference method (named as HW+SS) as well, and it is also used in [8]. The evaluation for speech enhancement is performed by SNR improvement, the percentage of energy loss (P_{el}) and the percentage of noise residue (P_{nr}) measure.

The SNR is often applied to evaluate the de-noising performance of speech enhancement method [4]. It is computed as

$$SNR = 10\log 10\left(\frac{\sum_n S_{allone}^2(n)}{\sum_n \left(S_{allone}(n) - \hat{S}_{out}(n)\right)^2}\right) \qquad (11)$$

$S_{allone}(n)$ and $\hat{S}_{out}(n)$ are the clean speech signal resynthesized from an all-one mask and the enhanced speech signal, respectively.

Figure 3 shows the average SNR improvement (SNRI) of the reference methods and proposed method for different noise types at different input SNRs. For the MRF-based algorithm, it is not easy to obtain the model parameters for constructing the spectral dependency. The error introduced by the model parameters may cause a result that the speech energy in some units could be more overestimated. So the method gets relatively lower SNRI.

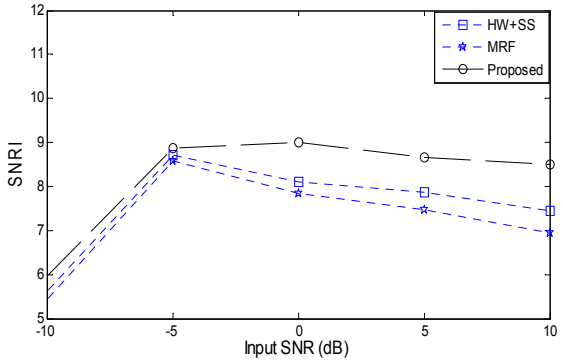


Figure 3: *Average SNR Improvement comparison*

In the proposed framework, we only need to know that the statistical characteristics of speech and noise instead of determining the specific distribution of speech and noise. In addition, for a data space, only one model parameter needs to be calculated to construct the spatial relationship and can be determined by potential entropy. As shown in Figure 3, our method with relatively simple structure can achieve consistently higher SNR results than the reference algorithms. Also, we can see that the SNRI at lower input SNR is less than one at higher input SNR, for example at -10dB SNR. The reason for this is probably that the asymmetry of speech and noise is obscured at lower input SNR. Thereby the sub-optimal σ could be generated so that the ability of separating speech and noise is weakened.

The percentage of energy loss (*Pel*) indicates the percentage of target speech excluded from enhanced speech, and the percentage of noise residue (*Pnr*) is the percentage of intrusion included. They provide complementary error measures of a segregation system and a successful system needs to achieve low errors in both measures.

Table 1: *Pel and Pnr results*

Input SNR (dB)			-10	-5	0	5	10
Average values from different noise types	HW+SS	P_{el} (%)	24.9	17.8	14.7	12.1	9.7
		P_{nr} (%)	36.0	13.9	6.9	2.5	0.8
		$P_{el}+P_{nr}$	60.9	31.7	21.6	14.6	10.5
	MRF	P_{el} (%)	20.7	13.8	10.8	9.9	8.4
		P_{nr} (%)	39.9	15.6	8.0	2.9	0.9
		$P_{el}+P_{nr}$	60.6	29.4	18.8	12.8	9.3
	Proposed	P_{el} (%)	18.2	11.4	9.8	9.1	8.3
		P_{nr} (%)	37.1	15.0	7.3	2.3	0.8
		$P_{el}+P_{nr}$	55.3	26.4	17.1	11.4	9.1
	Noisy	P_{nr} (%)	91.3	75.8	52.3	27.1	10.6

Table 1 gives the average results of the P_{el} and P_{nr} tests at different input SNRs, which are often used in CASA-based methods [2-3]. Compared with HW+SS method, the MRF and proposed methods with the spatial dependency show an advantage in restoring the speech energy. And because the method based on MRF does not consider enough information for establishing T-F dependency, it gets a relatively weak result. As a comparison, although the proposed smoothing algorithm does not perform well enough for the P_{nr} results in all cases, our system can produce the best performance for the sum of the P_{el} and P_{nr}.

5. Conclusions

In this paper, we present a new speech enhancement algorithm based on the spatial dependencies of the T-F units to improve the evolution nature of speech exciting in IBM prediction. In order to consider the spatial correlation, we introduce the DF prior model in the cochleagram. First, we exploit the concept of minimum statistics to obtain the initial T-F units of speech and noise from the energy potential of noisy speech in unvoiced portion, and we obtain their initial T-F units from a classical CASA-based method in voiced portion. Then given these initial value, we can construct the data spaces of speech and noise, and obtain the forms of the potentials of speech and noise. Subsequently the potential entropy is used to measure the uncertainty about the distributions of speech and noise for obtaining their respective optimal influence factors. Finally, given the optimal factors, the appropriate potentials of speech and noise are determined so that we can employ the potentials to generate the ratio mask for synthesizing enhanced speech. Experiments show that the proposed method can achieve a good performance compared to the reference methods.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61471014 and 61231015).

7. References

[1] A. S. Bregman, Auditory Scene Analysis. Cambridge, MA: MIT Press.

[2] G. Hu and D. L. Wang, "An Auditory Scene Analysis Approach to Monaural Speech Segregation," *Topics in Acoustic Echo and Noise Control.* Springer Berlin Heidelberg, pp. 485–515, 2010.

[3] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[4] M. Geravanchizadeh and R. Ahmadnia. "Monaural speech enhancement based on multi-threshold masking," *In blind source separation, G. R. Naik, W. Wang,* Springer Berlin Heidelberg, pp.369-393, 2014.

[5] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communications*, vol. 51, pp. 230–239, 2009.

[6] K. Hu and D. L. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 1600–1609, 2011.

[7] R. Probhavalkar, Z. Jin, and E. Fosler-Lussier, "Monaural segregation of voiced speech using discriminative random fields," *In Interspeech proceedings*, pp. 856–859, 2009.

[8] S. Liang, W. J. Liu, and W. Jiang. "Integrating binary mask estimation with MRF priors of cochleagram for speech separation." *IEEE Signal Processing Letters*, vol. 19, no.10, pp. 627-630, 2012.

[9] Y. Andrianakis and P. R. White. "A speech enhancement algorithm based on a Chi MRF model of the speech STFT amplitudes." *IEEE Transactions on Audio, Speech and Language Processing*, vol.17, no. 8, pp. 1508-1517, 2009.

[10] D. Y. Li and Y. Du, "Artificial Intelligence with Uncertainty," National Defense Press, 2005.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol.9, no.5, pp. 504-512, Jul. 2001.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.