



Spectrographic Speech Mask Estimation Using the Time-Frequency Correlation of Speech Presence

Ge Zhan, Zhaoqiong Huang, Dongwen Ying, Jielin Pan, and Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences

zhange@hccl.ioa.ac.cn

Abstract

This paper proposes a method to estimate the spectrographic speech mask based on a two-dimensional (2-D) correlation model. The proposed method is motivated by a fact that the time and frequency correlations of speech presence are interwoven with each other in the time-frequency (TF) domain. Conventional Markov chain is incapable of simultaneously modeling the time and frequency correlations in an adaptive way. The 2-D correlation model is presented to describe the correlation of speech presence in the TF domain, where the speech presence and absence are taken as two states of the model. The time correlation is modeled by the time state-transition probability and the forward factor, while the frequency state-transition probability and the corresponding neighbor factor are defined to describe the frequency correlation. The time and frequency correlations are incorporated into the model by maximizing the Q-function. A sequential scheme is presented to online estimate the parameter set. Given the observed spectrum and the parameter set, the state matrix that maximizes the posteriori probability is regarded as the optimal estimate of the speech mask. The proposed method was compared with some well-established methods. The experimental results confirmed its superiority.

Index Terms: Spectrographic speech mask, speech presence probability, time-frequency correlation, neighbor factor.

1. Introduction

Spectrographic speech masks have been widely used in many speech processing systems such as missing feature reconstruction [1]–[6], speech separation [7]–[9], speech perception [10], [11], and noise estimation [12], [13], where the speech mask is an essential prerequisite for their desirable performances. The speech mask can be regarded as a state matrix that represents speech presence/absence in the time-frequency (TF) domain.

Generally speaking, there exist two types of speech masks. One is the binary mask, which forces a hard decision to be made about whether speech is present or absent at each TF bin. A straightforward way for the binary mask estimation is to employ the local signal-to-noise ratio (SNR), where the noise power is estimated from the non-speech segments and the local SNR is compared with a threshold to determine the speech presence/absence [1], [2], [11]. The drawback is the empirically determined threshold that is sensitive to the performance. The priori-knowledge of clean speech signal is frequently used to estimate the Mel-scale mask [2], [4], [5], but this technique was seldom reported to be applied to linear spectral mask. The other is the soft mask, where the element of the state matrix takes on a continuum of value between 0 and 1, which is often referred to as the speech presence probability (SPP). One popular soft mask was presented based on a sigmoid function,

where the parameters were empirically determined [1], [3]. Improved minima controlled recursive averaging (IMCRA) [12] models the speech power and non-speech power by using a two-component Gaussian mixture model, and SPP is derived during the online estimation. Although IMCRA is adaptive to transitions between speech presence/absence, it is still heuristic somewhat since some parameters are optimally set. Constrained sequential hidden Markov model (CSHMM) [13] regards a time sequence of presence/absence of speech signals as a dynamic process of the transition between speech presence and absence. But the frequency correlation is simply considered by a hanning window smoothing [12], [13].

Conventional methods performed well in modeling the time correlation of speech signals on a subband. However, the frequency correlation was disregarded or not paid enough attention to. In fact, the frequency correlation plays the same important role as the time correlation in mask estimation. This paper proposes a two-dimensional (2-D) correlation model that describes the time and frequency correlations of speech presence/absence. The frequency state-transition probability is defined to describe the frequency transition, and accordingly, the neighbor factor is presented to model the correlation along the frequency. Meanwhile, the time correlation is elaborated by the time state-transition probability and the forward factor along the time. Both the pre-smoothing and post-smoothing are unnecessary since the time-frequency correlation of speech presence/absence is intrinsically considered in the model.

Some methods are utilized to model the 2-D correlation of speech signals [6], [9], [14], [15]. Markov random field (MRF) equally treats the transitions along the time and frequency [14], but MRF ignores the fact that the spectrum along the frequency is not present in chronological order. Recently, deep neural networks (DNNs) have been applied to mask estimation [6], [9], [15]. However, DNNs deeply relies on the large amount of pre-training data. If testing conditions mismatch pre-training conditions, the performance will substantially degrade.

2. Problem Formulation

The time correlation is well-known and widely used in many works. Actually, there exists high correlation along the frequency just like the time correlation. Fig. 1 demonstrates that high correlation exists along the frequency, where the correlation is expressed by the speech presence-to-presence probability. The state of speech presence/absence was hand-labeled on the clean spectrum. Then, the presence-to-presence probabilities were obtained by counting the percentage of the two frequency-adjacent bins with the same speech-presence state. From this figure, one can see the high correlation along the frequency, even for the bins with 5-unit frequency space. Since the time and frequency correlations are interwoven, a 2-D model is

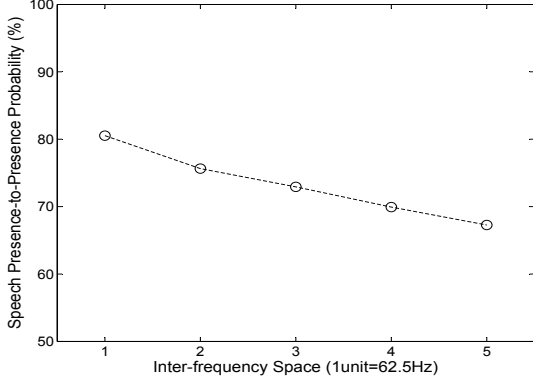


Figure 1: Statistic on frequency correlation of speech presence.

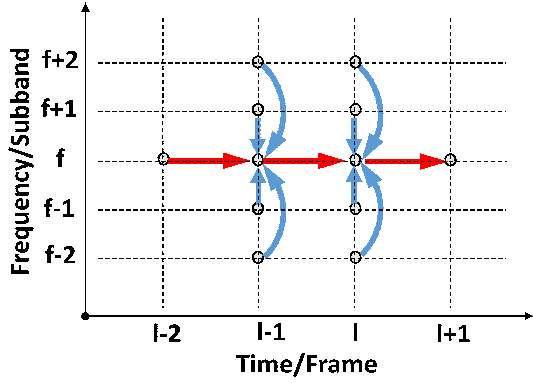


Figure 2: The integration of the time and frequency state transitions: red arrows for the time state transition, blue arrows for the frequency state transition.

proposed to elaborate the interwoven correlation.

This model considers a log-power spectrum at the time ℓ , $\mathbf{X}_\ell \triangleq [\mathbf{x}_{\ell-L+1}, \dots, \mathbf{x}_\ell]$, where $\mathbf{x}_t \triangleq [x_{1,t}, \dots, x_{F,t}]^T$ is the logarithmic power of the t -th frame ($\ell - L + 1 \leq t \leq \ell$). The state matrix corresponding to \mathbf{X}_ℓ is denoted as \mathbf{S}_ℓ , where $s_{f,t} = 0$ and $s_{f,t} = 1$ respectively denote the state of speech absence and presence at the (f, t) -th bin. On the frequency f , the time correlation is modeled by a Markov chain, in which the emission probability at the time t is represented by a Gaussian function as

$$b(x_{f,t}|s_{f,t} = j, \lambda_{f,\ell}) = \frac{1}{\sqrt{2\pi\kappa_{f,\ell}(j)}} \exp\left\{-\frac{(x_{f,t} - \mu_{f,\ell}(j))^2}{2\kappa_{f,\ell}(j)}\right\}, \quad (1)$$

where $\mu_{f,\ell}(j)$ and $\kappa_{f,\ell}(j)$ are respectively the mean and variance of the Gaussian distribution of the given state $s_{f,t} = j$. The time state-transition probability $\mathbf{a}_{f,\ell}$ in the Markov chain is a 2×2 matrix, expressed as

$$a_{f,\ell}(i, j) = p(s_{f,t} = j | s_{f,t-1} = i, \lambda_{f,\ell}). \quad (2)$$

Here, $\lambda_{f,\ell} \triangleq \{\mu_{f,\ell}, \kappa_{f,\ell}, \mathbf{a}_{f,\ell}\}$ denotes the parameter set of the Markov chain. The frequency state correlation is modeled by the frequency state-transition probability $\mathbf{c}_{d,\ell}$, denoted as

$$c_{d,\ell}(h, j) = p(s_{f,t} = j | s_{f+d,t} = h, \Lambda_\ell), \quad (3)$$

where d is the frequency space. Fig. 2 illustrates the integration of the time and frequency state transitions. The time correlation considers only one-unit time state transition, whereas the frequency correlation considers at most $2 \times D$ frequency-adjacent bins. The bins with more than D -unit frequency space are assumed to be uncorrelated. The whole parameter set $\Lambda_\ell \triangleq \{\lambda_{1,\ell}, \dots, \lambda_{F,\ell}, \mathbf{c}_{1,\ell}, \dots, \mathbf{c}_{D,\ell}\}$. It should be mentioned that all frequencies share the frequency state-transition probability.

The probability density function of the proposed model is expressed as

$$p(\mathbf{X}_\ell | \Lambda_\ell) = \sum_{\mathbf{S}_\ell} p(\mathbf{X}_\ell | \mathbf{S}_\ell, \Lambda_\ell) p(\mathbf{S}_\ell | \Lambda_\ell), \quad (4)$$

in which $p(\mathbf{X}_\ell | \mathbf{S}_\ell, \Lambda_\ell)$ is the probability density function of the given \mathbf{S}_ℓ , denoted as

$$p(\mathbf{X}_\ell | \mathbf{S}_\ell, \Lambda_\ell) = \prod_{f=1}^F \prod_{t=\ell-L+1}^{\ell} b(x_{f,t} | s_{f,t}, \lambda_{f,\ell}). \quad (5)$$

$p(\mathbf{S}_\ell | \Lambda_\ell)$ is the probability of \mathbf{S}_ℓ , given by

$$p(\mathbf{S}_\ell | \Lambda_\ell) = \prod_{f=1}^F \prod_{t=\ell-L+1}^{\ell} a_{f,\ell}(i, j) \prod_{d=1}^D c_{d,\ell}(h, j). \quad (6)$$

Hence, the modeling problem involves estimating Λ_ℓ on the basis of the maximum-likelihood criterion, given by

$$\Lambda_\ell = \arg \max_{\Lambda} p(\mathbf{X}_\ell | \Lambda_\ell), \quad (7)$$

and the estimation process serves as model-based clustering. The optimal estimate of the mask is obtained by

$$\mathbf{S}_\ell = \arg \max_{\mathbf{S}_\ell} p(\mathbf{S}_\ell | \mathbf{X}_\ell, \Lambda_\ell). \quad (8)$$

3. Sequential Scheme

The parameter set is estimated by the expectation-maximization (EM) algorithm [16], where the parameters are optimized to maximize $p(\mathbf{X}_\ell | \Lambda_\ell)$ in (4). A sequential scheme is presented to adapt the model frame by frame, and the current model Λ_ℓ is a function of the preceding model $\Lambda_{\ell-1}$ and the current observation \mathbf{x}_ℓ . The model is initialized by an offline EM algorithm, using the first M frames. Given the limited space, the initialization is omitted. The sequential scheme is presented in details.

The scheme is based on the maximum-likelihood criterion, namely

$$\Lambda_\ell = \max_{\Lambda} \log Q_{\ell|\Lambda_{\ell-1}}(\Lambda), \quad (9)$$

where the Q -function is defined as

$$Q_{\ell|\Lambda_{\ell-1}}(\Lambda) \triangleq E\{\log p(\mathbf{X}_\ell, \mathbf{s}_\ell | \Lambda_{\ell-1})\}. \quad (10)$$

The Q -function is maximized by sequentially searching for maximum likelihood based on the iterative Newton-Raphson algorithm [17], [18], given by

$$\Lambda_\ell = \Lambda_{\ell-1} + (I_\ell(\Lambda_{\ell-1}))^{-1} G_\ell(\Lambda_{\ell-1}), \quad (11)$$

where

$$I_\ell(\Lambda_{\ell-1}) = -\frac{\partial^2 Q_{\ell|\Lambda_{\ell-1}}(\Lambda)}{\partial \Lambda^2} \Big|_{\Lambda=\Lambda_{\ell-1}}, \quad (12)$$

$$G_\ell(\Lambda_{\ell-1}) = \frac{\partial Q_{\ell|\Lambda_{\ell-1}}(\Lambda)}{\partial \Lambda} \Big|_{\Lambda=\Lambda_{\ell-1}}.$$

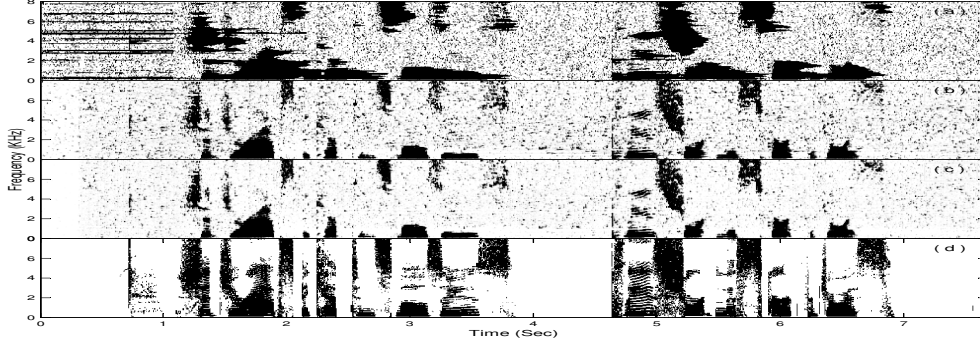


Figure 3: Comparison of the spectrographic speech masks estimated by (a) IMCRA, (b) CSHMM, (c) the proposed method; (d) the spectrogram of clean speech.

Substituting each parameter into (11) and (12) yields the recursive processes. The recursive processes for means and variances are given by

$$\mu_{f,\ell}(j) = \tilde{\alpha}_{f,\ell}(j)\mu_{f,\ell-1}(j) + [1 - \tilde{\alpha}_{f,\ell}(j)]x_{f,\ell}, \quad (13)$$

$$\kappa_{f,\ell}(j) = \tilde{\alpha}_{f,\ell}(j)\kappa_{f,\ell-1} + [1 - \tilde{\alpha}_{f,\ell}(j)][x_{f,\ell} - \mu_{f,\ell-1}(j)]^2, \quad (14)$$

where $\tilde{\alpha}_{f,\ell}(j)$ is a smoothing factor, defined as

$$\tilde{\alpha}_{f,\ell}(j) = \alpha\bar{\gamma}_{f,\ell-1}(j)/\bar{\gamma}_{f,\ell}(j), \quad (15)$$

where the smoothed SPP is given by

$$\bar{\gamma}_{f,\ell}(j) = \alpha\bar{\gamma}_{f,\ell-1}(j) + (1 - \alpha)\gamma_{f,\ell}(j), \quad (16)$$

in which α is a constant forgetting factor, $\gamma_{f,\ell}(j)$ is the conditional SPP.

The time state-transition probability is tracked through a non-linear recursive equation, given by

$$a_{f,\ell}(i, j) = a_{f,\ell-1}(i, j) + \tau_{f,\ell-1}(i, j)/\omega_{f,\ell-1}(i, j), \quad (17)$$

where

$$\tau_{f,\ell-1}(i, j) = \frac{\xi_{f,\ell}(i, j)}{a_{f,\ell-1}(i, j)} - \frac{\xi_{f,\ell}(i, 1-j)}{1 - a_{f,\ell-1}(i, j)}, \quad (18)$$

$$\omega_{f,\ell-1}(i, j) = \frac{L\bar{\xi}_{f,\ell}(i, j)}{a_{f,\ell-1}^2(i, j)} + \frac{L\bar{\xi}_{f,\ell}(i, 1-j)}{[1 - a_{f,\ell-1}(i, j)]^2}. \quad (19)$$

Here, $L = \lfloor \alpha/(1 - \alpha) \rfloor$. $1 - j$ and j denote two mutual transformation states. $\bar{\xi}_{f,\ell}(i, j)$ is the smoothed time state-transition probability, described as

$$\bar{\xi}_{f,\ell}(i, j) = \alpha\bar{\xi}_{f,\ell-1}(i, j) + (1 - \alpha)\xi_{f,\ell}(i, j), \quad (20)$$

where $\xi_{f,\ell}(i, j)$ is the conditional time state-transition probability.

The frequency state-transition probability is tracked in the same way with the time state-transition probability, given by

$$c_{d,\ell}(h, j) = c_{d,\ell-1}(h, j) + \eta_{d,\ell-1}(h, j)/v_{d,\ell-1}(h, j), \quad (21)$$

where

$$\eta_{f,\ell-1}(h, j) = \frac{\phi_{f,\ell}(h, j)}{c_{d,\ell-1}(h, j)} - \frac{\phi_{f,\ell}(h, 1-j)}{1 - c_{d,\ell-1}(h, j)}, \quad (22)$$

$$v_{f,\ell-1}(h, j) = \frac{L\bar{\phi}_{f,\ell}(h, j)}{c_{d,\ell-1}^2(h, j)} + \frac{L\bar{\phi}_{f,\ell}(h, 1-j)}{[1 - c_{d,\ell-1}(h, j)]^2}. \quad (23)$$

The smoothed frequency state-transition probability is

$$\bar{\phi}_{f,\ell}(h, j) = \alpha\bar{\phi}_{f,\ell-1}(h, j) + (1 - \alpha)\phi_{f,\ell}(h, j), \quad (24)$$

where $\phi_{f,\ell}(h, j)$ is the conditional frequency state-transition probability.

In fact, the recursive process (13) to (20) are similar to those in [13], the time-frequency correlation involves three conditional probabilities, namely $\gamma_{f,\ell}(j)$, $\xi_{f,\ell}(i, j)$, and $\phi_{d,\ell}(h, j)$. The three probability functions are expressed as

$$\gamma_{f,\ell}(j) = [F\Psi]_{f,\ell}(j) / \sum_j [F\Psi]_{f,\ell}(j), \quad (25)$$

$$\xi_{f,\ell}(i, j) = \frac{[F\Psi]_{f,\ell-1}(i)a_{f,\ell-1}(i, j)b_{f,\ell}(j)\Psi_{f,\ell}(j)}{\sum_{ij} [F\Psi]_{f,\ell-1}(i)a_{f,\ell-1}(i, j)b_{f,\ell}(j)\Psi_{f,\ell}(j)}, \quad (26)$$

$$\phi_{d,\ell}(h, j) = \frac{\sum_{f=1}^F [F\Psi]_{f+d,\ell}(h)c_{d,\ell-1}(h, j)F_{f,\ell}(j)}{\sum_{hj} \sum_{f=1}^F [F\Psi]_{f+d,\ell}(h)c_{d,\ell-1}(h, j)F_{f,\ell}(j)}, \quad (27)$$

where $[F\Psi]$ is the abbreviation for the product of the forward and neighbor factor. Assuming that the model gradually varies with time, the neighbor factor is defined to model the frequency correlation, given by

$$\Psi_{f,\ell}(j) = \sum_{d=1}^D \sum_h b_{f+d,\ell}(h)c_{d,\ell-1}(h, j), \quad (28)$$

where $b_{f+d,\ell}(j)$ is $b(x_{f+d,\ell}|s_{f+d,\ell} = h, \lambda_{f+d,\ell-1})$. The forward factor is defined to model the time correlation, given by

$$F_{f,\ell}(j) = \sum_i F_{f,\ell-1}(i)a_{f,\ell-1}(i, j)b_{f,\ell}(j), \quad (29)$$

where $b_{f,\ell}(j)$ is $b(x_{f,\ell}|s_{f,\ell} = j, \lambda_{f,\ell-1})$. As is proposed, the soft mask is given by the conditional SPP, and the binary mask S_ℓ can be obtained by the classic Viterbi decoding [19]. Both of them are intrinsically unified to the correlation model.

4. Evaluation

The proposed method was compared with two well-established methods, under various noisy conditions. The noise signals

Table 1: Logarithmic Spectral Distortion under various conditions (dB).

SNR (dB)	White noise			F16 cockpit noise			Babble noise		
	IMCRA	CSHMM	Proposed Method	IMCRA	CSHMM	Proposed Method	IMCRA	CSHMM	Proposed Method
-5	16.99	14.41	12.66	14.18	12.07	10.96	13.16	11.85	11.53
0	15.07	12.71	10.98	12.51	10.56	9.76	11.49	10.36	10.22
5	13.28	11.18	10.18	11.04	9.21	8.69	9.96	8.97	8.79
10	11.75	9.78	9.03	9.73	8.09	7.78	8.58	7.70	7.79

Table 2: Output SNR under various conditions (dB).

SNR (dB)	White noise			F16 cockpit noise			Babble noise		
	IMCRA	CSHMM	Proposed Method	IMCRA	CSHMM	Proposed Method	IMCRA	CSHMM	Proposed Method
-5	-0.41	3.08	4.92	-1.77	1.65	2.98	-3.36	-1.04	0.05
0	4.14	7.31	8.49	3.02	5.99	6.88	1.56	3.63	4.42
5	8.66	11.39	12.05	7.56	10.22	10.58	6.24	7.92	8.63
10	12.96	15.48	15.74	12.00	14.43	14.36	11.02	12.64	12.82

Algorithm 1 : Sequential scheme

- 1: For the incoming (ℓ)-th frame
- 2: Perform FFT and extract the log power \mathbf{x}_ℓ .
- 3: For each frequency f and each state j
- 4: Calculate $\Psi_{f,\ell}(j)$ via (28).
- 5: Calculate $F_{f,\ell}(j)$ via (29).
- 6: End
- 7: For each frequency space d and each state $[h, j]$
- 8: Calculate $\phi_{d,\ell}(h, j)$ via (27).
- 9: Update $\mathbf{c}_{d,\ell}$ via (21) to (24).
- 10: End
- 11: For each frequency f and each state $[i, j]$
- 12: Calculate $\gamma_{f,\ell}(j)$ via (25).
- 13: Calculate $\xi_{f,\ell}(i, j)$ via (26).
- 14: Update $\lambda_{f,\ell}$ via (13) to (20).
- 15: End
- 16: End

used for evaluation, such as white Gaussian, F16 cockpit, and babble noises, were obtained from the NOISEX-92 database [20]. Twenty short clean utterances were taken from the TIMIT database [21]. Ten long clean utterances were constructed by connecting every two short utterances. Then, the noise signals were artificially added to the long utterances at SNRs of -5 , 0 , 5 , and 10 dB. The sampling rate of all the signals is 16 kHz. The testing dataset consisted of 120 groups (10 utterances \times 3 noises \times 4 noise levels). Each group comprised three utterances that were respectively processed by IMCRA [12], CSHMM [13], and the proposed method. In the proposed method, the parameters were empirically set as $D = 5$, $M = 40$, and $\alpha = 0.97$.

A long noisy utterance was used to give an informal test. The soft masks consisting of the SPP were estimated by the three methods at white Gaussian noise conditions with 5 dB SNR, in Fig. 3. The transition from white to black corresponds to probability changing from 0 to 1 . The spectrogram of SPP reflects a trade-off between restoring speech signals and eliminating noise signals. The pseudo speech presence is significantly suppressed by the proposed method. Then, two objective evaluation methods were used for comparison. In each noisy speech spectrum, the FFT coefficients at the bins within the speech-presence state were conserved, but the others were set as zero. The logarithmic spectral distortion [22] was obtained by comparing the binary-masked spectrum and the correspond-

ing clean speech spectrum. A smaller distortion suggests better performance. Then, the binary-masked spectrum was inversely transformed into the separated speech signals. The output SNR was obtained by comparing the separated speech signals and the corresponding clean speech signals, given by

$$\epsilon = 10 \log_{10} \left[\frac{\sum_{n=1}^N z_n^2}{\sum_{n=1}^N (z_n - \hat{z}_n)^2} \right], \quad (30)$$

where N denotes the number of samples, \hat{z} represents the separated speech signals, and z is the clean speech signals. A higher score suggests better performance. Tables 1 and 2 present the evaluation results. The performance difference is owing to the different methods that model the time and frequency correlations. Both IMCRA and CSHMM utilize the frequency correlation through a Hanning-based averaging over neighboring frequencies. However, the state of the current frequency will be substantially influenced if there exists an amplitude outlier on the neighboring frequency. On contrast, the proposed method models the time-frequency correlation in an adaptive way. The influence of amplitude outliers is controlled at a low level. Therefore, the proposed method yields the clearer mask and better experimental results than CSHMM and IMCRA do.

5. Conclusions

This paper presents a 2-D correlation model to estimate the spectrographic speech mask. A sequential scheme is proposed to adapt the model frame by frame. The conditional SPP is considered as the soft mask, and the \mathbf{S}_ℓ given by (8) is considered as the optimal estimate of the binary mask. The experimental results have confirmed that the proposed method performs better than CSHMM and IMCRA since it fully takes advantage of the time-frequency correlation of speech presence/absence. The false alarms of speech presence and absence are efficiently suppressed.

6. Acknowledgement

This work was supported by the National Program on Key Basic Research Project (2013CB329302), the National Natural Science Foundation of China (Nos. 61271426, 11461141004, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), and by the CAS Priority Deployment Project (KGZD-EW-103-2).

7. References

- [1] J. Barker, L. Josifovski, M. Cooke, P. Green “Soft decisions in missing data techniques for robust automatic speech recognition,” *INTERSPEECH 2000*, pp 373–376.
- [2] W. Kim, J.H. Hansen, “A novel mask estimation method employing posterior-based representative mean estimate for missing-feature speech recognition,” *IEEE Trans. Speech, Audio, and Language Process.*, 19(5):1434–1443, 2011.
- [3] J. Hout, A. Alwan, “A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition,” *ICASSP 2012*, pp. 4105–4108.
- [4] M.L. Seltzer, B. Raj, R.M. Stern, “A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition,” *Speech Commun.*, 43, 379–393, 2004.
- [5] B.J. Borgstrom, A. Alwan, “A statistical approach to Mel-domain mask estimation for missing-feature ASR,” *IEEE Signal Process. Letters*, 17(11):941–944, 2010.
- [6] A. Narayanan, D.L. Wang, “Investigation of speech separation as a front-end for noise robust speech recognition,” *IEEE/ACM Trans. on Speech, Audio, and Language Process.*, 22(4):826–835, 2014.
- [7] M. Cobos, J.J. Lopez, “Maximum a posteriori binary mask estimation for underdetermined source separation using smoothed posteriors,” *IEEE Trans. on Speech, Audio, and Language Process.*, 20(7):2059–2064, 2012.
- [8] A.M. Reddy, B. Raj, “Soft mask methods for single-channel speaker separation,” *IEEE Trans. on Audio, Speech and Language Process.*, 15(6):1766–1776, 2007.
- [9] P.S. Huang, M. Kim, M.H. Johnson, P.Smaragdis “Deep learning for monaural speech separation,” *ICASSP 2014*, pp. 1562–1566, 2014.
- [10] U. Kjems, J.B. Boldt, M.S. Pedersen, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *J. Acoust. Soc. Am.* 126(3):1415–1426, 2009.
- [11] A. Narayanan, D.L. Wang, “The role of binary mask patterns in automatic speech in background noise,” *J. Acoust. Soc. Am.*, 133(5):3083–3093, 2013.
- [12] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.* 11(5):466–475, 2003.
- [13] D. Ying, Y. Yan, “Noise estimation using a constrained sequential hidden markov model in the log-spectral domain,” *IEEE Trans. Speech, Audio, and Language Process.*, 21(6):1145–1157, 2013.
- [14] Y. Andrianakis, P.R. White, “A speech enhancement algorithm based on a chi MRF model of the speech STFT amplitudes,” *IEEE Trans. Speech, Audio, and Language Process.*, 17(8):1508–1517, 2009.
- [15] B. Li, K.C. Sim, “A spectral masking approach to noise-robust speech recognition using deep neural networks,” *IEEE/ACM Trans. on Speech, Audio, and Language Process.*, 22(8):1296–1305, 2014.
- [16] L. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” *Ann. Math. Statist.*, 41:164–171, 1970.
- [17] E. Weinstein, M. Feder, A. Oppenheim, “Sequential algorithm for parameter estimation based on the Kullback-Leibler information measure,” *IEEE Trans. Acoust., Speech, Signal Process.*, 38(9):1652–1654, 1990.
- [18] V. Krishnamurthy, J. Moore, “On-line estimation of hidden Markov model parameters based on the Kullback - Leibler information measure,” *IEEE Trans. Signal Process.*, 41(8):2557–2573, 1993.
- [19] A. J. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. Inf. Theory.*, 13(2):260–269, 1967.
- [20] A. Varga, H. 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 Commun.*, 12(3):247–251, 1993.
- [21] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,” *Nat. Inst. Standards Technol. (NIST)*, Gaithersburg, MD, prototype as of Dec. 1988.
- [22] L. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.