Integrating Binary Mask Estimation With MRF Priors of Cochleagram for Speech Separation

Shan Liang, Wenju Liu, and Wei Jiang

***Abstract*—In present binary masking based speech separation systems, it is almost impossible to obtain the ideal binary mask (IBM). The error in IBM estimation usually results in energy absence in many speech-dominated time-frequency (T-F) units. It violates smooth evolution nature of the speech signal and creates great artefacts. Markov random field (MRF) is one of the promising approaches to model smooth evolution nature which has been extensively applied to image smoothing applications. In this letter, an MRF prior for modeling the spatial dependencies in audio cochleagram is introduced. With this prior model, we further smooth the binary mask based cochleagram and generalize binary mask to ratio mask via a Bayesian framework. Our algorithm is systematically evaluated and compared with other counterpart methods, and it yields substantially better performance, especially on suppressing artefacts.**

***Index Terms*—Ideal binary mask, ideal ratio mask, iterated conditional modes (ICM), Markov random field.**

I. Introduction

MONAURAL speech separation from interference is one of the key problems in speech processing. Researches on human auditory perception inspire one promising approach which is called *Computational auditory scene analysis* (CASA) [1]–[6]. The main computational goal of CASA has been set as the ideal binary mask (IBM) estimation [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. Reliable class consists of the units in which speech energy exceeds the interference, while unreliable class consists of the rest. To synthesize the waveform signal, the energy in reliable units is retained and the energy in unreliable units is rejected totally. This means that IBM transforms the complex noise spectrum estimation problem into a binary classification problem which is simpler to achieve. Since most of the speech energy is contained in a very small amount of units, IBM approximates to the ideal spectrum estimation closely.

Meanwhile, the T-F representations of speech and many real world noises show high temporal correlation and evolve smoothly. However, there are many abrupt changes in the binary mask based speech spectrum. Besides, it's almost impossible to estimate the IBM with one-hundred-percent accuracy in practice. The error in IBM estimation may greatly violate smooth evolution nature and result in huge artefacts. To the best of our knowledge, suppressing this distortion has not been received much attention up to now.

In image processing problems, MRFs have been extensively applied for smooth applications via modeling the spatial dependencies [10]. In speech separation area, MRFs have not widely used to date. Recently, Probhavalkar *et al.* introduce the theory of Discriminative Random Fields (DRFs), which are closely related to MRFs, into the IBM estimation problem for voiced speech separation [6]. The use of DRF allows them to take the spatial dependencies into account via an interaction potential function. The results suggest that CASA techniques may benefit from the DRF framework.

A major contribution of this letter is the introduction of a common MRF, Gaussian MRF, based prior model of audio cochleagram for smoothing the artefacts. To construct an MRF, a set of conditional density functions is defined for representing the correlation between neighbors. The neighborhood defines the interactions between two units in the spatial representations. Since the local temporal correlation is taken into account, a smoother speech cochleagram is obtained and artefacts are suppressed to some extent. The letter is organized as follows. An overview of the proposed framework will be presented in the next section. Follow by discussions on prior models and the smoothing algorithm in Section III. The proposed algorithm is evaluated in Section IV. The last section gives some conclusions.

II. Framework Overview

The mixture signal is decomposed into T-F domain firstly by 64-channel gammatone filters from 50 Hz to 8000 Hz [14]. 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 [2].

After that, we estimate the IBM in the continuous voiced frames with a state-of-art voiced speech separation model proposed by Hu and Wang [3]. Let Y_i, S_i and D_i denote the energy of mixture, speech and interference at the i*'th* T-F unit respectively. As in [4], binary mask M_i based estimations of speech and noise energy are given by:

$$\left[\tilde{S}_i, \tilde{D}_i\right] = \begin{cases} [Y_i, 0], & \text{if } M_i = 1 \\ [0, Y_i], & \text{if } M_i = 0 \end{cases}. \tag{1}$$

As the discussion in T-F unit level [4], accurate binary mask results in a good approximation to the true energy, while wrong mask leads to a great error.

Manuscript received May 15, 2012; revised July 01, 2012; accepted July 03, 2012. Date of publication July 19, 2012; date of current version July 31, 2012. This work was supported in part by the China National Nature Science Foundation (91120303, 90820011, and 90820303). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Constantine L. Kotropoulos.

The authors are with Institute of Automation, Beijing 100190, China (e-mail: sliang@nlpr.ia.ac.cn; lwj@nlpr.ia.ac.cn; wjiang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/LSP.2012.2209643

Since pitch is the key cue, HuWang model couldn't handle unvoiced speech. Recently, Hu and Wang combine CASA and spectral subtraction [7] for the IBM estimation in unvoiced speech frames [5]. The spectral subtraction based noise energy estimation is the main cue to generate unvoiced speech dominated units using simple thresholding or Bayesian classification. In this stage, we just aim to generate initial estimations of noise and speech energy. Therefore, no further classification of the reliable and unreliable units is required. The average value over several preceding and succeeding frames is used to approach the noise energy. The speech energy is obtained by subtracting the average value from the mixture energy. The voiced/unvoiced classification of frames is carried out with the *RAPT* pitch tracking algorithm [15].

Then, a Bayesian framework is proposed to smooth the error in cochleagram estimation:

$$\begin{aligned}\hat{S},\hat{D} &= \arg\max_{S,D} p(S,D|Y) = \arg\max_{S,D} p(Y|S,D)p(S,D)\\ &= \arg\max_{S,D} p(Y|S,D)p(S)p(D). \end{aligned} \tag{2}$$

The prior models, $p(S)$ and $p(D)$, which are modeled by two Gaussian MRFs capture the spatial dependencies in the cochleagram. The *maximum a posterior* (MAP) estimation is approached by the ICM algorithm [11].

We then generalize the IBM estimation to ideal ratio mask (IRM) estimation, R_i, which is defined in [4]:

$$R_i = \frac{\hat{S}_i}{(\hat{S}_i + \hat{D}_i)}. \tag{3}$$

Finally, the waveform signal is resynthesized by weighting the cochleagram with ratio mask and correcting phase shifts [2], [3].

III. MRF Prior Models and IRM Estimation

A. MRF Prior Models

The Gaussian MRF defines a sequence of conditional density functions:

$$p(S_i|S_{n(i)}) \propto \exp\left[-\frac{1}{\sigma_{S,i}^2}\left(S_i - \sum_{j\in N_S(i)} b_{ij}S_j\right)^2\right] \tag{4}$$

$$p(D_i|D_{n(i)}) \propto \exp\left[-\frac{1}{\sigma_{D,i}^2}\left(S_i - \sum_{j\in N_D(i)} c_{ij}D_j\right)^2\right], \tag{5}$$

where $\sigma_{S,i}$ and $\sigma_{D,i}$ control the scaling of the two densities. The parameters b_{ij} and c_{ij} determine the influence between the neighbors i and j. The neighborhoods, $N_S(i)$ and $N_D(i)$, define the interaction between the i'th unit and the others.

Suppose subscript i denotes the unit at the k'th frame and the l'th frequency channel, $i \triangleq (k,l)$. As previously mentioned, the cochleagrams of speech and noise evolve slowly along time frames. Additionally, units which are adjacent in frequency within a time frame are also highly correlated due to the fact that adjacent gammatone filters overlap heavily. Therefore, the four nearest units, $\{(k+1,l),(k-1,l),(k,l+1),(k,l-1)\}$, are selected as the neighbors for both the speech and interference. It is worth to mention that the units which lie on the edge of cochleagram have less than four neighbors. Besides, the production mechanisms of voiced and unvoiced speech are nominally very different, so some irregular and abrupt changes inherent in the border of voiced and unvoiced frames. Therefore, we add a constraint that the voiced and unvoiced units can't be neighbors for speech particularly. We assign equal-weight to all the neighbors. The mean of each conditional density is given by:

$$\begin{aligned}\mu_{S,i} &= \sum_{j\in N_S(i)} b_{i,j}S_j,\ b_{i,j} = \frac{1}{|N_S(i)|},\ \forall j \in N_S(i)\\ \mu_{D,i} &= \sum_{j\in N_D(i)} c_{i,j}D_j,\ c_{i,j} = \frac{1}{|N_D(i)|},\ \forall j \in N_D(i), \end{aligned} \tag{6}$$

where $|N_S(i)|$ and $|N_D(i)|$ represent the number of the neighbors.

For speech and many real-world sounds, the energy unequally distributes in T-F representation. In order to provide an equal scaling for all units, $\sigma_{S,i}$ and $\sigma_{D,i}$ are normalized as:

$$\sigma_{S,i} = \gamma_S\mu_{S,i},\ \ \sigma_{D,i} = \gamma_D\mu_{D,i}, \tag{7}$$

where γ_S and γ_D are smoothing factors which are determined experimentally. The lower the factors are, the smoother the cochleagrams are. Given N observations, the *maximum likelihood* (ML) estimations of the two factors are given by:

$$\hat{\gamma}_S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{S_i}{\mu_{S,i}}\right)^2,\ \hat{\gamma}_D^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{D_i}{\mu_{D,i}}\right)^2. \tag{8}$$

B. Dependence Among S_i, D_i and Y_i

To model the dependence among S_i, D_i and Y_i, we define a random variable ε_i which subjects to the following constraint:

$$S_i + D_i = \varepsilon_i Y_i. \tag{9}$$

Since the speech and interference are statistically independent with each other, Y_i is approximately equal to the sum of S_i and D_i. This means that ε_i distributes in a narrow range around 1. With the short-time Fourier transform based T-F representation, Batina et.al point out that the reciprocal of obeys to an exponential distribution with mean and variance equal to 1 [12]. In each ICM iteration, however, the joint function of exponential distribution with the priors given in (4) and (5) is so complex that the MAP estimation is difficult to be solved. Therefore, it is replaced with a Gaussian distribution with mean equal to 1:

$$p(Y_i|S_i,D_i) = p(\varepsilon_i) \propto \exp\left[-\frac{1}{\sigma_\varepsilon^2}(\varepsilon_i - 1)^2\right]. \tag{10}$$

As the discussion in the following subsection, the MAP estimation can be obtained by a convex quadratic optimization problem.

C. ICM Based Optimization Algorithm

It is very difficult to solve the global optimization problem given in (2) directly due to the large scale dependencies. To simplify the computation complexity, we apply the Iterated Con-

TABLE I
PSEUDOCODE OF THE MRF BASED SMOOTHING ALGORITHM

Initial S_i and D_i for all units with Eq. (1)
For all time frames k
  For all frequency channels l
    Compute $\mu_{S,i}$ and $\mu_{D,i}$ with Eq. (6).
    Compute $\sigma_{S,i}$ and $\sigma_{D,i}$ with Eq. (7).
    Compute $\hat{S}_i$ and $\hat{D}_i$ with Eq. (13)(14).
    Normalize with -30 dB as the lower bound.
Compute the ratio mask with Eq. (3)

ditional Modes (ICM) algorithm [11] which maximizes local conditional probabilities sequentially to approach the MAP estimation. Given observation Y_i and the neighbors $S_{N(i)}, D_{N(i)}$, ICM algorithm updates S_i and D_i sequentially by maximizing the following local conditional probability:

$$\hat{S}_i, \hat{D}_i = \arg\max_{S_i, D_i} p(Y_i|S_i, D_i)p(S_i|S_{N(i)})p(D_i|D_{N(i)}). \tag{11}$$

Discard the constant components, (11) can be transformed into a convex quadratic optimization problem:

$$\hat{S}_i, \hat{D}_i = \arg\min_{S_i, D_i} \left[\frac{(\varepsilon_i - 1)^2}{\sigma_\varepsilon^2} + \frac{(S_i - \mu_{S,i})^2}{\sigma_{S,i}^2} + \frac{(D_i - \mu_{D,i})^2}{\sigma_{D,i}^2} \right]. \tag{12}$$

The optimal solution of (12) is given by:

$$A\left[\hat{S}_i, \hat{D}_i\right]^T = b \Rightarrow \left[\hat{S}_i, \hat{D}_i\right]^T = A^{-1}b, \tag{13}$$

where

$$A = \begin{bmatrix} \frac{1}{\sigma_{S,i}^2} + \frac{1}{\sigma_\varepsilon^2 Y_i^2} & , & \frac{1}{\sigma_\varepsilon^2 Y_i^2} \\ \frac{1}{\sigma_\varepsilon^2 Y_i^2} & , & \frac{1}{\sigma_{D,i}^2} + \frac{1}{\sigma_\varepsilon^2 Y_i^2} \end{bmatrix}, b = \begin{bmatrix} \frac{\mu_{S,i}}{\sigma_{S,i}^2} + \frac{1}{\sigma_\varepsilon^2 Y_i} \\ \frac{\mu_{D,i}}{\sigma_{D,i}^2} + \frac{1}{\sigma_\varepsilon^2 Y_i} \end{bmatrix}. \tag{14}$$

Besides, $\hat{S}_i$ and $\hat{D}_i$ are limited to be greater than -30 dB in implementation. The algorithm is summarized in Table I.

IV. EXPERIMENTAL RESULTS

We evaluate the proposed algorithm with 40 sentences which are randomly taken from the training set in Speech Separation Challenge (SSC, 2006) [17]. All the sentences are subsequently merged into five 10 s-length signals and down-sampled to 16 kHz before noises are added. Three types of real world noises recorded in cafeteria, square and subway environments [18] are used as the interference. For each type of noise, two 10 s-length signals are selected. The mixture signals are generated with the SNR in the range of 0 to 9 dB with 3 dB steps.

According to (8), smoothing factors γ_S and γ_D are equal to 0.33 and 0.24 respectively with a quarter of all the units in the test corpus as observations. Since our main focus is smoothing the speech cochleagram, the two factors are appropriately adjusted to 0.3 and 0.3. With the same observations, the ML estimation of σ_ε is equals to 0.4 which is obtained by $\sigma_\varepsilon^2 = (1/(N-1))\sum_{i=1}^{N}(\varepsilon_i - 1)^2$.

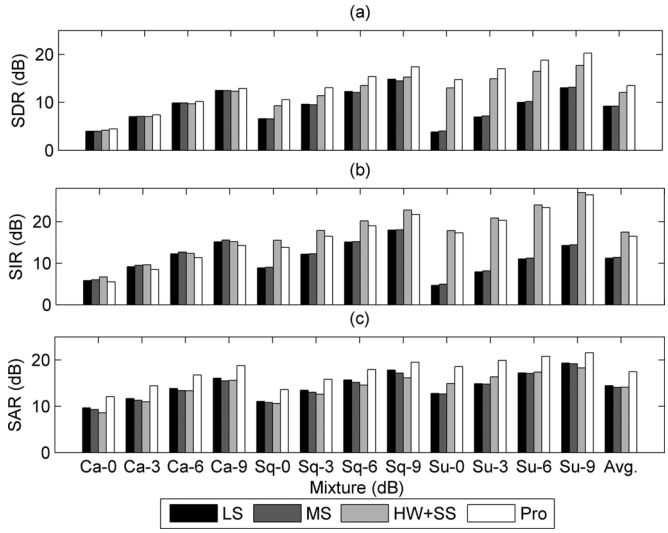

Fig. 1. SDR, SIR and SAR results for different noise types and input SNR levels. Ca: cafeteria, Sq: square, Su: subway, Avg.: the average result.

Let $s(t)$, $d(t)$ and $\hat{s}(t)$ denote the clean speech, additive interference and separated speech. Three measures, SDR(signal-to-distortion ratio), SIR(signal-to-interference ratio), and SAR(signal-to-artifact ratio), proposed by Vincent *et al.* [13] are used to evaluate the performance. The measures are calculated by projecting the separated or enhanced speech signal onto the subspaces expanded by the speech and interference. SIR and SAR measure the interferences and artifacts terms respectively, while SDR measures the total distortion. The artifact term ε_{artif} is defined as [13]:

$$\varepsilon_{artif} = \hat{s} - \frac{\langle s, \hat{s}\rangle \hat{s}}{\|s\|^2} - \frac{\langle d, \hat{s}\rangle \hat{s}}{\|d\|^2}, \tag{15}$$

where $\langle s, \hat{s}\rangle \triangleq \sum_t s(t)\hat{s}(t)$ and $\|s\|^2 \triangleq \langle s, s\rangle$. To isolate the amplification and distortion effects brought by the gammatone filters, the ground truth is resynthesized by the original speech with all-one mask [3]. We compare the performance of the proposed algorithm with HuWang model [3] combined with spectral subtraction method [7] (HW+SS) and two well-known speech enhancement methods, the Log Spectral (LS) amplitude scheme [8] and the minimum statistics noise estimator (MS) [9].

The average SDR, SIR and SAR results of the four algorithms are shown in Fig. 1(a)–(c) respectively. We can see from Fig. 1(a) that our algorithm achieves consistently higher SDR results than HW+SS algorithm. On average, the SDR improvements are 0.43, 1.74 and 2.19 dB for the three types of noise respectively. Compared to LS and MS, the proposed and HW+SS algorithms show a great advantage on the square and subway noises. The main reason is that both of the LS and MS couldn't effectively track the highly non-stationary noise.

As is shown in Fig. 1(b) and (c), HW+SS algorithm achieves relatively high SIR results and relatively low SAR results at most conditions. By contrast, the proposed algorithm improves SAR results effectively. One minor disadvantage of the proposed smoothing algorithm is that it may decrease the SIR results slightly. This is due to the fact that the speech energy in some units may be overestimated. Overall, the improvement on

TABLE II
PERCENT PREFERENCE (%) FOR THE PROPOSED ALGORITHM COMPARED TO OTHER METHODS FOR DIFFERENT NOISES

Method	Ca	Sq	Su	Avg.
LS	69.2	83.3	82.6	78.4
MS	63.6	77.3	72.7	71.2
HW+SS	62.5	58.3	57.9	59.6

SAR results is more significant than the SIR decreasing, about 2.37 dB higher on average.

As in [16], we use formal listening tests for evaluating the quality of speech separated by the proposed algorithm. In the test, pairs of sentences, one processed by our method and one processed with one of the other methods, are presented to six normal-hearing listeners. The order of the sentences is randomized. Then, the listeners select from the pair which is more natural and includes fewer background noises. The preference results are presented in Table II. From the results, we can find that the proposed algorithm has higher preference ratio compared to HW+SS algorithm under all noise conditions. This mainly results from the better artefacts suppression of our algorithm.

V. CONCLUSION

In this letter, the Gaussian MRF prior models of cochleagram are jointed with one binary mask estimation based speech separation algorithm. As the temporal correlation is taken into account, the two prior models could smooth the binary mask based cochleagrams and suppress the artefacts effectively. Experiments on three real-world noises show that the proposed algorithm outperforms previous systems in the terms of SDR and SAR results.

REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press.
[2] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley/IEEE Press, 2005, ch. 1.
[3] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.
[4] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2009.
[5] K. Hu and D. L. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, pp. 1600–1609, 2011.
[6] R. Probhavalkar, Z. Jin, and E. Fosler-Lussier, "Monaural segregation of voiced speech using discriminative random fields," *Interspeech*, pp. 856–859, 2009.
[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 27, pp. 113–120, Apr. 1979.
[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Signal Audio Process*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[10] R. Kinderman and J. L. Snell, *Markov Random Fields and Their Applications*. Providence, RI: AMS, 1980.
[11] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Statist. Soc. Ser. B*, vol. 48, pp. 259–302, 1986.
[12] I. Batina, J. Jensen, and R. Heusdens, "Kalman filtering based noise power spectral density estimation for speech enhancement," in *Proc. 13th European Signal Processing Conf.*, Sep. 2005.
[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio., Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[14] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, An Efficient Auditory Filterbank based on the Gammatone Function MRC Applied Psychology Unit, 1988, Rep. 2341.
[15] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. New York: Elsevier, 1995, pp. 495–518.
[16] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220–231, 2006.
[17] M. Cooke and T. Lee, Signal Separation Challenge 2006 [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm
[18] *Signal Separation Evaluation Campaign*, , 2011 [Online]. Available: http://sisec.wiki.irisa.fr/tiki-index.php?page=Two-channel+mixtures+of+speech+and+real-world+background+noise