[10] D. L. Jones and T. W. Parks, "A resolution of several time frequency representations," in *Proc. IEEE ICASSP'89*, Glasgow, U.K., pp. 2222–2225.
[11] K. Kodera, R. Gendrin, and C. de Villedary, "Analysis of time-varying signals with small BT values," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp. 64–76, 1978.
[12] R. Koening, H. K. Dunn, and L. Y. Lacy, "The sound spectrograph," *J. Acoust. Soc. Amer.*, vol. 18, pp. 19–49, 1946.
[13] F. Plante, G. Meyer, and W. A. Ainsworth, "Speech signal analysis with reassigned spectrogram," in *Proc. IEEE Symp. Time–Frequency and Time–Scale Analysis*, Philadelphia, PA, 1994, pp. 640–643.
[14] F. Plante and W. A. Ainsworth, "Formant tracking using reassigned spectrum," in *Proc. Eurospeech 95*, Madrid, Spain, 1995, vol. 1, pp. 741–744.
[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C.* Cambridge, U.K.: Cambridge Univ. Press, 1992.
[16] E. F. Velez and R. G. Absher, "Transient analysis of speech signals using the Wigner time-frequency representations," in *IEEE ICASSP'89*, vol. 4, pp. 2242–2244.
[17] B. Zhang and S. Sato, "A time–frequency distribution of Cohen's class with a compound kernel and its application to speech signal processing," *IEEE Trans. Signal Processing*, vol. 42, pp. 54–64, 1994.
[18] Y. Zhao, L. E. Atlas, and R. J. Marks, "The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1084–1091, 1990.

Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction

Zenton Goh, Kah-Chye Tan, and B. T. G. Tan

***Abstract*—In this correspondence, we investigate whether musical noise, which often exists in speech enhanced using spectral subtraction, can be suppressed. Via exploiting some specific characteristics of human speech, we propose a method that can effectively suppress musical noise without noticeable effect on speech intelligibility. Performance assessments confirm that our method is effective.**

***Index Terms*—Musical noise, spectral subtraction, speech enhancement.**

I. Introduction

In many practical situations, speech has to be recorded in the presence of undesirable background noise. As noise often degrades the quality/intelligibility of recorded speech, it is beneficial to carry out noise suppression. In the literature, a variety of speech enhancement methods capable of suppressing noise has been proposed. Spectral subtraction [1], [2] is among the traditional methods that have been extensively studied.

Spectral subtraction is popular because it can suppress noise effectively, even in some real-life scenarios. In addition, the underlying concept is relatively straightforward, and this leads to simplicity in implementation. (There is a commercial product that employs spectral subtraction.) However, spectral subtraction tends to introduce a specific disturbance, commonly referred to as *musical noise*. Contrary to its name, musical noise is not necessarily pleasing, and can be annoying. In fact, the unacceptability of musical noise has motivated invention of enhancement methods based on considerations different from those of spectral subtraction. Some of such methods, for example those proposed by Lim and Oppenheim [3], and Ephraim and Malah [4], are very promising and have received considerable research attention.

Here, we investigate whether musical noise introduced by spectral subtraction can be suppressed without noticeable effect on speech intelligibility. In this connection, it is worthwhile mentioning the relevant methods proposed by Boll [2] and Whipple [5]. Basically, both methods are developed based on the assumption that the spectral components of musical noise usually appear as isolated peaks in the spectrogram of enhanced speech. However, in practice, musical noise manifests itself as not only isolated peaks, but also "short ridges" in the spectrogram, and therefore will not be effectively suppressed by these methods. Although it is possible to suppress those "short ridges" by modifying some parameters associated with the methods, the intelligibility of speech would usually have to be compromised.

In this work, we first identify the trade-offs among suppression of unwanted noise, generation of musical noise, and preservation of the intelligibility of desired speech. Subsequently, we propose a postprocessing method capable of suppressing musical noise effectively without noticeable effect on speech intelligibility, via exploiting some specific characteristics of human speech. Finally, we subject our method to tests with speech sentences obtained from the TIMIT speech data base [6], to which we add white Gaussian noise and computer-fan noise (separately) amounting to various values of signal-to-noise ratios (SNR's). Performance assessments based on spectrogram plots, objective measures, and informal subjective listening tests show consistently good results.

II. Preliminary Discussion and Motivation

Since the first step to our method will be spectral subtraction, we shall present a brief discussion on spectral subtraction. We shall adopt an additive noise model

$$y[n] = s[n] + d[n] \tag{1}$$

where $y[n]$, $s[n]$ and $d[n]$ denote discrete-time samples of noisy speech, clean speech, and noise, respectively. Subjecting the samples to sampled short-time Fourier transform (SSTFT), we obtain

$$Y_r[k] = S_r[k] + D_r[k] \tag{2}$$

where $Y_r[k]$, $S_r[k]$, and $D_r[k]$ denote, respectively, the SSTFT's of $y[n]$, $s[n]$, and $d[n]$ for Frame r.

The most general version of spectral subtraction is known as *generalized spectral subtraction* (GSS). GSS first obtains the magnitudes of the SSTFT of the enhanced speech $|\hat{S}_r[k]|$'s, which we shall refer hereafter as *spectral magnitudes*, and the phases $\arg(\hat{S}_r[k])$'s with the following recipe:

$$|\hat{S}_r[k]| = \begin{cases} (|Y_r[k]|^\alpha - \beta|\hat{D}_r[k]|^\alpha)^{1/\alpha}, & \text{if } |Y_r[k]|^\alpha > \beta|\hat{D}_r[k]|^\alpha \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\arg(\hat{S}_r[k]) = \arg(Y_r[k]) \tag{4}$$

where α and β are positive constants, and $|\hat{D}_r[k]|$ is an estimate

Manuscript received February 3, 1996; revised July 9, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. John H. L. Hansen.

Z. Goh and K.-C. Tan are with the Centre for Signal Processing, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Republic of Singapore (e-mail: ezgoh@ntu.edu.sg).

B. T. G. Tan is with the Faculty of Science, National University of Singapore, Singapore 119260, Republic of Singapore.

Publisher Item Identifier S 1063-6676(98)02895-8.

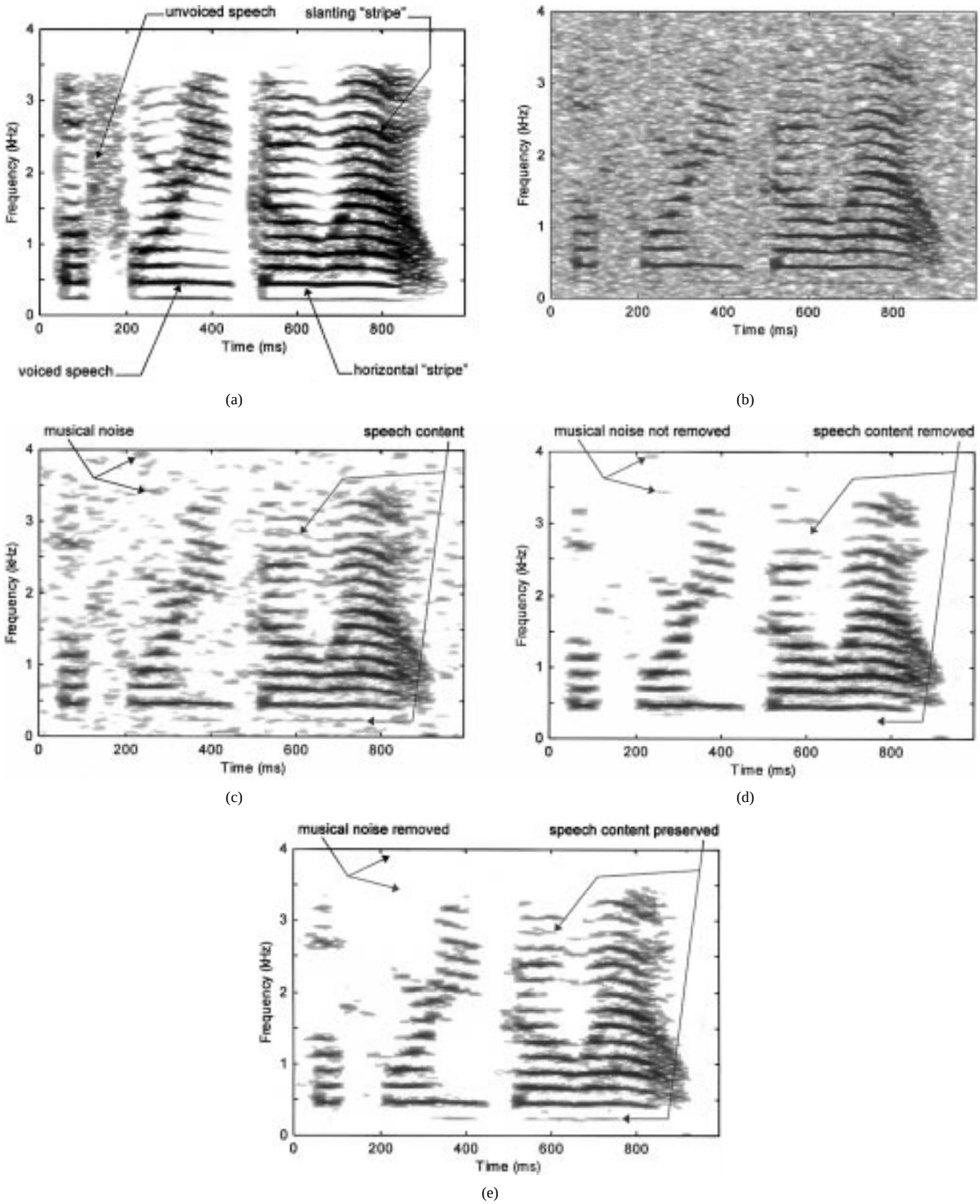


Fig. 1. (a) Spectrogram of the (clean) speech sentence "before you go out," (b) spectrogram of the speech corrupted by white Gaussian noise, (c) spectrogram of the enhanced speech obtained using GSS with a moderate β, (d) spectrogram of the enhanced speech obtained using GSS with a large β, and (e) spectrogram of the enhanced speech obtained using our method.

of $|D_r[k]|$. Usually, $|\hat{D}_r[k]|$ is obtained by averaging those $|Y_r[k]|$'s that contain only noise. With $|\hat{S}_r[k]|$ and $\arg(\hat{S}_r[k])$, the enhanced speech can be obtained via inverse discrete Fourier transform and the standard overlap-add processing.

The parameter α has some effect on speech intelligibility, whereas β controls the amount of noise suppression. For example, when α is set to two, the enhanced speech often appears to be more intelligible than that obtained with an α equals to 1 or 0.5 (see [1]). However, musical noise seems to be relatively more annoying for the case where $\alpha = 2$. In general, the value of β should be just large enough to attenuate the unwanted noise. While using a very large β could fully attenuate the unwanted noise and suppress musical noise generated in the spectral subtraction process, this leads to weakening of speech content, which in turn reduces the intelligibility.

Before we begin the discussion of our method, it is beneficial to examine the spectrograms (graphical plots of spectral magnitudes) of typical clean speech, noisy speech, and enhanced speech. Fig. 1(a) shows the spectrogram of an 8 kHz (clean) speech signal. The horizontal axis of the spectrogram denotes time, vertical axis frequency, and the spectral magnitude is shown with gray shade (darker shade indicates larger value). Observe that a large portion of the spectrogram is practically blank (i.e., unshaded) and the speech energy is concentrated in a few isolated regions. In the figure, the voiced portion of speech is characterized by dark parallel "stripes" whereas unvoiced portion is characterized by gray patches. Notice that some parallel stripes are horizontal while some are slanting up or down, indicating a change in the pitch of the speech signal.

When white Gaussian noise amounting to an SNR of 10 dB is added to the clean speech, the blank region of the spectrogram as shown in Fig. 1(a) becomes shaded, and some of the stripes corresponding to voiced speech disappear [see Fig. 1(b)]. With an appropriate spectral subtraction, we obtain an enhanced speech with spectrogram as shown in Fig. 1(c). Spectral subtraction has suppressed the noise greatly, and consequently Fig. 1(c) resembles Fig. 1(a) much more than Fig. 1(b) does. However, noise suppression is achieved at a price—many isolated short stripes which correspond to musical noise are generated in the process.

Musical noise can be easily eliminated via *oversubtraction* (i.e., GSS using a larger β), but this will be at the expense of speech intelligibility. Indeed, in Fig. 1(d) which shows the spectrogram of the enhanced speech obtained using GSS with a large β, we observe a significant reduction of the unwanted short stripes. At the same time, some stripes observed in Fig. 1(c) (the spectrogram of the enhanced speech with a smaller β), corresponding to the desired speech content, are eliminated. [Fig. 1(e) will be referred to in Section IV.]

In short, it is possible to suppress unwanted noise effectively with GSS. However, the speech quality is compromised (because of the annoying musical noise) and/or the speech intelligibility decreases. Consequently, it is a challenge to suppress unwanted noise effectively while maintaining reasonably high speech quality and intelligibility.

III. Our Method

Our strategy is to first obtain the spectral magnitudes (and hence the spectrogram) of the enhanced speech via the GSS recipe, using appropriate α and β such that the enhanced speech is of reasonably high intelligibility but with resulting appreciable musical noise. The next step involves suppressing the short stripes in the spectrogram that correspond to musical noise, without noticeable effect on the speech content. The final step requires computation of enhanced speech via inverse discrete Fourier transform and overlap-add processing, with the use of the spectral magnitudes obtainable from the processed spectrogram and the spectral phases of the noisy speech.

The effectiveness of our method depends greatly on the ability to identify which regions of the spectrogram correspond to the desired speech signal and which regions correspond to musical noise, and the processings to be carried out over these regions. Therefore, we shall focus on our classification (identification) approach and our specific treatment to the various regions of the spectrogram.

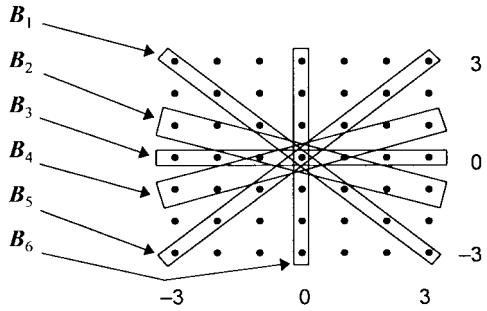


Fig. 2. Six blades over 7 × 7 spectrogram points. The point of concern coincides with (0, 0), the centroids of the blades.

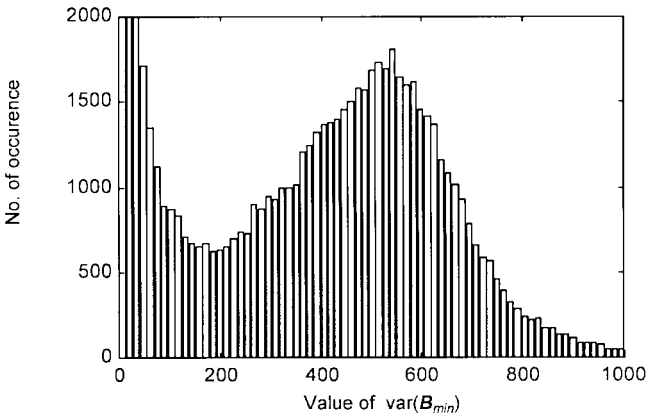


Fig. 3. Histogram of var($\boldsymbol{B}_{\min}$).

A. Classification

We shall determine which regions in the spectrogram are very likely to correspond to speech, and which regions correspond to either musical noise or speech (of low energy). For convenience, we shall refer to the regions very likely to be speech as Region 1, and the other as Region 2.

1) Stage 1: We shall exploit the fact that musical noise can be effectively reduced via oversubtraction [see Fig. 1(d) and the accompanying discussion in Section II]. Indeed, we propose first computing the spectrogram of the enhanced speech based on GSS with a large β. We then include those spectrogram points (r, k)'s that have spectral magnitudes greater than zero in Region 1:

$$(r, k) \in \text{Region 1} \qquad \text{if } |Y_r[k]|^\alpha - \beta|\hat{D}_r[k]|^\alpha > 0. \tag{5}$$

Clearly, those spectral components of speech that are strong enough will be retained.

2) Stage 2: The fact that oversubtraction leads to decrease in speech intelligibility implies that some spectral components corresponding to speech are attenuated in the oversubtraction process. It is therefore sensible to assume that there would be some additional spectrogram points that can be classified under Region 1. However, the spectral values associated with these points are low, and usually comparable to those of musical noise, meaning that any form of classification based on "energy" discrimination would be difficult. This then motivates the development of the following method which exploits specific characteristics of speech.

First define blades $\boldsymbol{B}_i$'s, for $i = 1, \cdots, p$, with different orientations over the point of concern [i.e., (0,0)], in the way shown

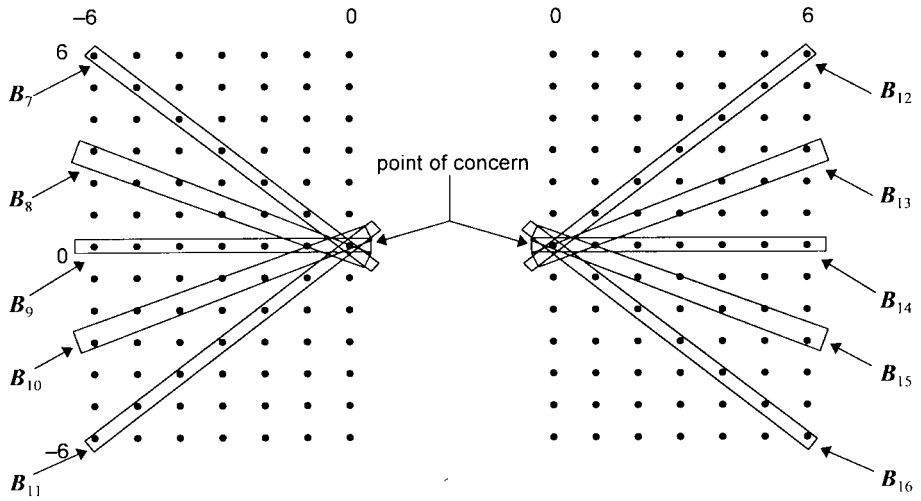


Fig. 4. (a) Five "left" blades and (b) five "right" blades.

in Fig. 2. The grid points of $\boldsymbol{B}_1, \cdots, \boldsymbol{B}_6$ as shown in Fig. 2 are $\{(-3, 3), (-2, 2), (-1, 1), (0, 0), (1, -1), (2, -2), (3, -3)\}$, $\{(-3, 1), (-2, 1), (-1, 0), (0, 0), (1, 0), (2, -1), (3, -1)\}$, $\{(-3, 0), (-2, 0), (-1, 0), (0, 0), (1, 0), (2, 0), (3, 0)\}$, $\{(-3, -1), (-2, -1), (-1, 0), (0, 0), (1, 0), (2, 1), (3, 1)\}$, $\{(-3, -3), (-2, -2), (-1, -1), (0, 0), (1, 1), (2, 2), (3, 3)\}$, and $\{(0, -3), (0, -2), (0, -1), (0, 0), (0, 1), (0, 2), (0, 3)\}$, respectively. The width of each blade should be thin enough so that the grid points being intersected form a straight line. Of course for some orientations, the width of the blade has to be somewhat larger so as to intersect a significant number of points, and the points intersected are not strictly in one straight line (see $\boldsymbol{B}_2$ and $\boldsymbol{B}_4$ of Fig. 2). In addition, the length of each blade should be longer than those of most short stripes, which correspond to musical noise, but shorter than the lengths of those stripes and large patches respectively corresponding to voiced and unvoiced speech.

To determine whether a spectrogram point belongs to Region 1, we compute $\text{var}(\boldsymbol{B}_i)$, the variance of the values of the spectrogram points which the blade intersects, for each blade $\boldsymbol{B}_i$ where $i = 1, \cdots, p$:

$$\text{var}(\boldsymbol{B}_i) = \sum_{(r,k)\in\boldsymbol{B}_i} \frac{\{20\log_{10}(|\hat{S}_r[k]|+1)\}^2}{|\boldsymbol{B}_i|} - \left\{\sum_{(r,k)\in\boldsymbol{B}_i} \frac{20\log_{10}(|\hat{S}_r[k]|+1)}{|\boldsymbol{B}_i|}\right\}^2. \quad (6)$$

We then identify $\boldsymbol{B}_{\min}$, the blade with variance being the minimum among $\text{var}(\boldsymbol{B}_i)$ for $i = 1, \cdots, p$:

$$\boldsymbol{B}_{\min} = \arg\min_{\boldsymbol{B}_i}[\text{var}(\boldsymbol{B}_i)]. \quad (7)$$

The variance associated with $\boldsymbol{B}_{\min}$, which will be denoted as $\text{var}(\boldsymbol{B}_{\min})$, would offer an indication as to whether the point concerned belongs to Region 1. Indeed, for a point belonging to those parallel stripes associated with voiced speech, $\boldsymbol{B}_{\min}$ will most likely be of the same orientation as the stripes, and $\text{var}(\boldsymbol{B}_{\min})$ will be quite small due to homogeneity in the spectral magnitude values. For a point within patches associated with unvoiced speech, all the variances will be reasonably low, especially $\text{var}(\boldsymbol{B}_{\min})$. On the other hand, the variances for a point which belongs to short stripes corresponding to musical noise will all be considerably high, because the blade length is longer than the stripe length, and $\boldsymbol{B}_{\min}$ intersects some points outside the stripes, in addition to those inside. Consequently, it is justifiable to assume that a spectrogram point (r, k) belongs to Region 1 if $\text{var}(\boldsymbol{B}_{\min})$ is considerably small:

$$(r, k) \in \text{Region 1 if var}(\boldsymbol{B}_{\min}) < \tau \quad (8)$$

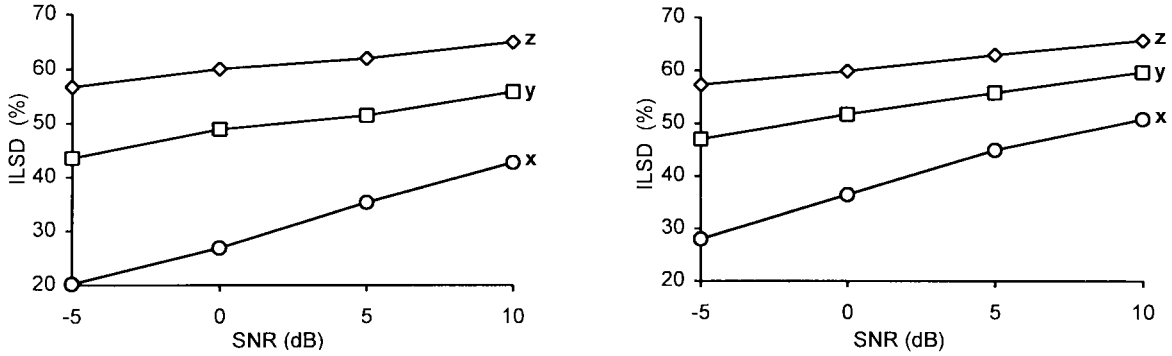
where τ is an appropriately chosen threshold. In fact, the histogram of $\text{var}(\boldsymbol{B}_{\min})$'s (such as that shown in Fig. 3) will often exhibit two peaks, of which one occurs at a large $\text{var}(\boldsymbol{B}_{\min})$ and the other small $\text{var}(\boldsymbol{B}_{\min})$. Analysis of the peaks confirms that the former correlates well with musical noise and the latter with speech signal. Therefore we recommend setting the threshold τ to be one around the valley between the two peaks.

Additional considerations were given to the points at the boundaries of stripes/patches associated with speech. In fact, for such a boundary point, almost every one of the blades will have one part which protrudes out of the stripe/patch of concern, leading to large $\text{var}(\boldsymbol{B}_{\min})$. To tackle this problem, we employ additional "left" and "right" blades as shown in Fig. 4 (the additional blades have lengths and orientation angles identical to the original ones shown in Fig. 2). For example, $\boldsymbol{B}_7 = \{(-6, 6), (-5, 5), (-4, 4), (-3, 3), (-2, 2), (-1, 1), (0, 0)\}$ and $\boldsymbol{B}_{13} = \{(0, 0), (1, 0), (2, 1), (3, 1), (4, 2), (5, 2), (6, 3)\}$ [note that the point of concern is at (0,0)]. Now for any one point of concern, one has to simply obtain $\boldsymbol{B}_{\min}$ and $\text{var}(\boldsymbol{B}_{\min})$ for the original, the left and the right blades, in exactly the way discussed in the previous paragraphs. There will thus be a single lowest $\boldsymbol{B}_{\min}$ for the three types of blades. The spectrogram points for the blade with this $\boldsymbol{B}_{\min}$ are most likely to be part of a stripe or patch associated with speech. On the other hand, for a point of concern on a short stripe associated with musical noise, the value of $\text{var}(\boldsymbol{B}_{\min})$ will remain high, as the lengths of the original, left and right blades are all longer than that of the stripe.

Remark: With the classification carried out in Stage 2, that in Stage 1 seems redundant. While this is generally true, Stage 1 is crucial when there exist sudden bursts of speech utterances which give rise to intense stripes with rapidly increasing spectral values. For these bursts, $\text{var}(\boldsymbol{B}_{\min})$ will be high and the point concerned can be confirmed to fall under Region 1 only with the energy discrimination approach that Stage 1 adopts.

B. Processing Treatment

Region 1: After the speech/musical-noise classification process, the spectrogram is divided into two regions, namely Region 1 and



x - Noisy speech

y - Enhanced speech obtained using GSS with a moderate β

z - Enhanced speech obtained using GSS with the same β and our method in cascade

(a) (b)

Fig. 5. Graphs of ILSD measurements versus SNR for (a) WGN and (b) computer fan noise.

Region 2. For those points in Region 1, we suggest leaving the spectral values untouched. For those points in Region 2, the following processing will be carried out.

Region 2: Region 2 comprises points that are associated with either musical noise or speech (of low energy). The criterion for processing should be that the spectral values of the points corresponding to musical noise are considerably attenuated, while those corresponding to speech are at most slightly altered. With this in mind, we propose replacing the spectral value of the point concerned by the median of the values of those spectrogram points which $\boldsymbol{B}_{\min}$ intersects, if the median value is not larger than the current spectral value

$$|\hat{S}_r[k]| = \underset{(r',k')\in\boldsymbol{B}_{\min}}{\text{median}}\,(|\hat{S}_{r'}[k']|) \quad \text{if} \quad \underset{(r',k')\in\boldsymbol{B}_{\min}}{\text{median}}\,(|\hat{S}_{r'}[k']|) < |\hat{S}_r[k]|. \tag{9}$$

Of course, over a spectrogram point corresponding to speech, any form of processing will likely to change the spectral value, and simply leaving the value untouched as we have suggested for Region 1 is probably a safer approach. However, the difficulty here is that we are unsure whether the point corresponds to either speech or musical noise. Fortunately, $\boldsymbol{B}_{\min}$ would be likely to coincide/overlap with the stripes/patches associated with speech. Consequently, the median value will not be too different from the spectral value of the point concerned due to the uniformity of points within such stripes. On the other hand, over a short stripe associated with musical noise, the median will take a spectral value considerably smaller than that of the point itself, since many points that the blade intersects will fall outside the stripe and will thus have much lower values.

C. The Complete Enhancement Procedure

Now we shall present the complete enhancement procedure we propose. Given a noisy speech signal, it is first buffered into overlapping frames with a frame size of 32 ms and an overlap of 24 ms. Each frame is then multiplied by a Hanning window and transformed to the frequency domain via a fast Fourier transform (FFT). Next, spectral subtraction based on (3) is employed, with $\alpha = 2$ and $\beta = 1.8$, for obtaining the SSTFT magnitudes of the enhanced speech. Subsequently, every SSTFT magnitude point is subject to classification: it will be classified to be in Region 1 (region very likely to be speech) or Region 2 (those not in Region 1) according to (5) with $\alpha = 2$ and $\beta = 16$, and also (8) with $\tau = 200$. Note that all 16 blades $\boldsymbol{B}_1, \cdots, \boldsymbol{B}_{16}$ as shown in Figs. 2 and 4 are used in the computation of $\boldsymbol{B}_{\min}$ as given by (7), which will in turn be used for classification via the recipe specified by (8). For points classified under Region 1, we leave them untouched. For points classified under Region 2, we recompute the SSTFT magnitude via (9). Finally, by combining the SSTFT magnitudes so obtained with the SSTFT phases obtained via (4) as well as applying inverse FFT, standard overlap-add, we get the desired postprocessed enhanced speech.

IV. Performance Assessment

We shall now assess the performance of our method. We used phonetically balanced speech sentences taken from the TIMIT Acoustic-Phonetic Speech Database [6] provided by the National Institute of Standards and Technology (NIST). Two male and two female sentences were chosen and down-sampled from 16 to 8 kHz. Also, 2 types of noise, namely computer-generated white Gaussian noise (WGN) and real computer-fan noise, amounting to various values of SNR (-5, 0, 5, and 10 dB) were considered.

For performance assessment, we relied on not only spectrogram plots, but also on objective measures such as segmental SNR (SEGSNR) and inverted linear spectral distance (ILSD) [7], and informal subjective listening tests. ILSD was employed because it has reasonably high correlation with diagnostic acceptability measure [7], a widely adopted *subjective* measure for overall speech quality and intelligibility. (Interested readers may refer to [7] for more details about SEGSNR and ILSD.) Note that the ILSD measure takes value between 0% and 100%, with 100% (0%) being best (worst) in overall quality and intelligibility. Note also that we removed silent intervals in the speech signals before computing SEGSNR since the silent intervals could drastically affect the value of SEGSNR.

We first evaluated the performance through a visual inspection of the spectrograms. The evaluation was carried out on the speech sentence "before you go out" corrupted by white Gaussian noise (WGN). Now recall that Fig. 1(c) is the spectrogram of the enhanced speech obtained using GSS with a moderate β, which exhibits a few isolated short stripes corresponding to musical noise. Subjecting the spectrogram to our postprocessing method, we obtained the spectrogram shown in Fig. 1(e). A comparison of the two spectro-

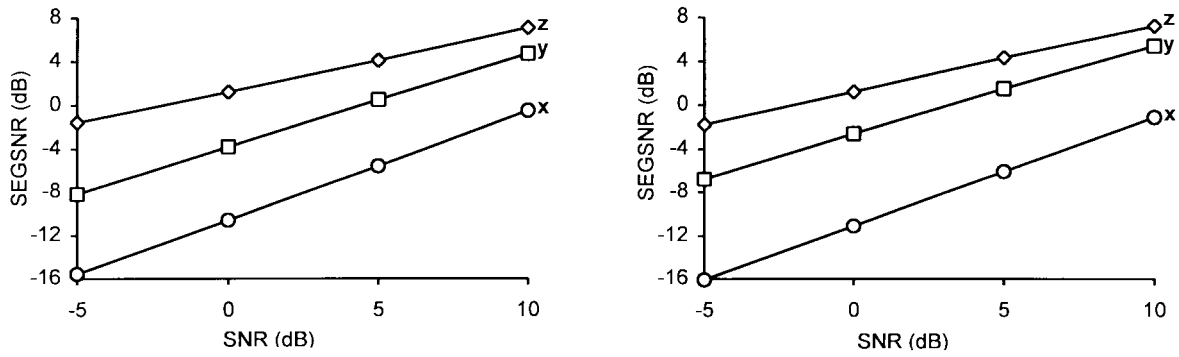


Fig. 6. Same as Fig. 5 except that SEGSNR instead of ILSD is used.

grams shows that most unwanted short stripes are eliminated while the parallel stripes and large patches corresponding to voiced and unvoiced speech respectively remains practically untouched. Fig. 1(d) shows the spectrogram of the enhanced speech obtained using GSS with a large β. The value of β is indeed large since some stripes associated with speech in Fig. 1(a) (the spectrogram of the clean speech) have disappeared. Unfortunately, even with such a large β, some musical noise still survives. On the other hand, it is encouraging to see that with our method, not only can the musical noise be almost completely removed, but also that the speech content is better preserved [see Fig. 1(e)]. Indeed, some musical noise observed in Fig. 1(d) is absent in Fig. 1(e), while some stripes corresponding to voiced speech are retained in Fig. 1(e) but not in Fig. 1(d) (see the relevant labels on both figures).

Next we compute the objective measures ILSD and SEGSNR with the use of all the four sentences mentioned. Figs. 5 and 6 shows ILSD and SEGSNR, respectively, for noisy speech, enhanced speech obtained using GSS with a moderate β, as well as enhanced speech obtained using GSS with the same β and our method in cascade. Clearly, our method offers significant improvements consistently in the presence of both WGN and computer-fan noise at various SNR's.

We also performed informal subjective listening tests on the four speech sentences. It was clear that the enhanced speech obtained using GSS with our method was much more pleasant than that with GSS alone (in the sense that musical noise could hardly be heard). Moreover, it was found that the intelligibility of the former was comparable to, if not higher than, the latter.

V. Conclusion

We have developed a postprocessing method for suppressing musical noise generated by spectral subtraction. Visual evaluation based on spectrograms, objective assessment based on ILSD and SEGSNR, and informal subjective listening tests all indicated that our method is reasonably effective.

Our method involves two crucial steps, speech/musical-noise classification and processing of spectral values with incorporation of the classification results. For classification, we carry out oversubtraction as the first stage to single out those regions clearly corresponding to speech. Then we adopt the multiblade approach to effect a finer classification, with the objective of identifying speech components that are of comparable energies to musical noise. On processing, we propose leaving the spectral values untouched for those regions classified to be speech. For other regions, we propose a treatment based on median filtering, which has a tendency to suppress musical noise while altering the speech content only slightly.

For further work, we propose to explore the use of our method for removing unwanted noise inevitably generated in other enhancement process. Indeed, although our method was applied only in conjunction with spectral subtraction, it may be appropriate to employ our method to suppress the enhancement noise associated with other methods such as Wiener filtering [4], signal subspace decomposition and filtering [8], etc. This issue will be addressed in our future work.

References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[3] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197–210, June 1978.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[5] G. Whipple, "Low residual noise speech enhancement utilizing time-frequency filtering," in *Proc. ICASSP'94*, pp. I-5/I-8.

[6] "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Nat. Inst. Standards Technol., Gaithersburg, MD, 1988.

[7] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[8] Y. Ephraim and H. L. Van Trees, "A spectrally-based signal subspace approach for speech enhancement," in *Proc. ICASSP'95*, pp. 804–807.