NOVEL BINAURAL SPECTRO-TEMPORAL ALGORITHM FOR SPEECH ENHANCEMENT IN LOW SNR ENVIRONMENTS

Po-Hsun Sung, Bo-Wei Chen, Ling-Sheng Jang and Jhing-Fa Wang
Department of Electrical Engineering
National Cheng Kung University
Tainan City, Taiwan, R.O.C.
phsung@gmail.com, chenbw@mail.ncku.edu.tw, lsjang@ee.ncku.edu.tw, wangjf@mail.ncku.edu.tw

Abstract—A novel Binaural Spectro-Temporal (BIST) algorithm is proposed in this paper to increase the speech intelligibility in low or negative SNR noisy environments. The BIST algorithm consists of two modules. One is the spatial mask for receiving sound from the specific direction, and the other is the spectro-temporal modulation filter for noise reduction. Most speech enhancement algorithms are not applicable in harsh environments because the energy of speech is covered by the noise. To increase the speech intelligibility in low or negative SNR noisy environments, a distinctive approach is proposed to solve this problem. First, the BIST algorithm takes binaural auditory processing as a spatial mask to separate the speech and noise according to their locations. Next, the modulation filter is applied to reduce the noise source in the scale-rate (spectro-temporal modulation) domain according to their different acoustic features. It works like the spectro-temporal receptive field (STRF) which is the perception response of human auditory cortex. The experimental results demonstrate that the proposed BIST speech enhancement algorithm can improve 20% from the noisy speech at SNR -10dB.

Keywords-binaural processing, cochleagram, spectro-temporal modulation, speech intelligibility, low SNR, noisy environments

I. INTRODUCTION

Speech enhancement used in noisy environment has been an open research problem for the past several decades. Many types of algorithms are developed to solve this problem. They can be roughly classified as follows: i) spectral subtraction techniques [1, 2], ii) statistical model-based approaches [3], and iii) subspace-based techniques [4]. Nevertheless, distortion and loss of quality, like music noise, are drawbacks of these solutions. Speech with heavy noise remains a difficult issue to solve because it is hard to separate the mixture of speech and noise efficiently. When noise is removed by those methods, the quality of speech is also reduced in the meantime. Based on human auditory perception, human can switch attention between different sound sources. Related literature showed that human can identify the location and acoustic feature of sound sources in parallel [5]. This means sound source contains cues of its location and acoustic features.

In this paper, the clean speech is extracted from noisy sound according to its location and acoustic characteristics step by step. The binaural auditory processing functions as a spatial mask which is used to separate the target source and noises based on their directions. Regarding the pattern of sound source, the transformation in the representation of sound is founded in mammalian auditory systems, called spectro-temporal receptive fields (STRFs) [6]. STRFs represent the transformation of neuron responses in the auditory cortex when the sound wave arrives. It decomposes the content of auditory spectrogram into the scale-rate domain. The scale means the spectral modulation and the unit is defined as cycles/octave (or cycles/kHz) and rate means temporal modulation and the unit is cycles/second (Hz) [7]. This multi-domain representation model of speech is proven useful in estimation of speech intelligibility [8]. Hence, speech and noise can be separated due to their different modulation characteristics in the scale-rate domain. The spectro-temporal modulation filter can be used as a mask to remove the noise pattern from noisy speech to get cleaner speech. Generally, the quality of speech is defined as the overall quality of perception measured in terms of intelligibility, clarity and naturalness. In order to enhance the perception of speech quality, this study proposes a Binaural Spectro-Temporal (BIST) algorithm, which combines the spatial mask and spectro-temporal modulation filter for enhancement. It can focus on the direction of sound source and separate the speech and noise, depending on spectro-temporal modulation characters. The feature of this approach is to improve speech intelligibility in low or negative SNR environments.

The rest of this study is organized as follows. Section II introduces the binaural auditory processing. Section III then describes spectral temporal modulation filtering. Next, Section IV summarizes the performance of the proposed method and the analysis results. Conclusions are finally drawn in Section V, along with recommendation for future research.
II. BINAURAL AUDITORY PROCESSING

The human binaural system has a number of astonishing capabilities that are effective to separate sound source in the adverse acoustic environments. Based on these observations, many theoretical models and computational algorithms have been developed by the use of interaural time difference (ITDs) / interaural phase difference (IPDs), interaural level difference (ILDs) and other cues. The transformation impulse response of a sound source to the ear is normally described by a transfer function called the head-related transfer function (HRTF). Beside the binaural cues described above, the ear canal, pinna and shoulder effects are also included in HRTF. This is the reason why the human can specify the location of sound source including the azimuth and elevation angle information. Here, we adopt this model to refine the specific direction of a sound source as if we constrain the receiving direction of binaural hearing in the spatial field.

A. Peripheral auditory system processing

The structure of the sound source separation system models the processing in the peripheral auditory system with HRTF characteristics. The acoustic inputs in this model, \( x_L(n) \) and \( x_R(n) \), are from the left and right microphones at both ears. The spatial characteristic contributed by HRTFs is included in the acoustic signals. Signals are initially processed by stimulating the periphery auditory system which is modeled by a gammatone filters with a bank of \( N = 40 \) . The center frequencies of these filters are equally spaced on the equivalent rectangular bandwidth (ERB) scale between 100 Hz and 8000 Hz. The model is implemented from Slaney’s Auditory Toolbox [9] and the flow is shown in Fig. 1.

![Figure 1. Function block of binaural spectro-temporal algorithm for enhancement of speech intelligibility](image)

B. Receiving sound from specific direction through spatial mask

The primary cues for sound directions are the difference in timing and intensity of the signals received from left-right microphones. ITDs are determined by the azimuthal angle of the source and independent of the elevation angle. The correlation processes which participate in the formation of the position of the auditory is proved. And the spatial hearing of the two ears is generally derived by cross-correlation analysis [10]. In this section, we refer the methodology of binaural sound source separation developed by Chanwoo Kim [11]. Furthermore, the received signals in our proposed method are from microphones at both ears as shown in Fig. 1. This spatial mask is realized by defining a threshold value from normalized cross-correlation. The normalized cross-correlation value with \( t = t_0 \) at fixed frequency bin can be defined as:

\[
\rho(t_0) = \frac{1}{T_0} \int_{t_0}^{t_0 + T_0} x_L(t; t_0) x_R(t; t_0) \, dt
\]

\[
\sqrt{\frac{1}{T_0} \int_{t_0}^{t_0 + T_0} (x_L(t; t_0))^2 \, dt \cdot \frac{1}{T_0} \int_{t_0}^{t_0 + T_0} (x_R(t; t_0))^2 \, dt}
\]

If the sound source is located at an angle \( \theta \) and the desired signal source is assumed as plane wave propagation, the signal with fixed frequency bin from both ears can be approximately expressed as follows:

\[
x_L(t; t_0) = A \sin(\omega_0 (t - \tau)); x_R(t; t_0) = A \sin(\omega_0 t)
\]

where \( \omega_0 \) is the center frequency and \( \tau \) is the time delay between both ears.

![Figure 2. Schematic of the spatial mask for receiving sound from the specific direction (shaded region).](image)

By substituting (2) into (1), the following relationship can be obtained:

\[
\rho(t_0) = \cos(\omega_0 \tau) = \cos(\omega_0 d \sin \theta / C_0)
\]

where \( C_0 \) is wave velocity of sound.

Thus, if \( \rho_\text{TTH} \) is defined as a threshold to constrain the direction of the received signal, we can keep a given time-frequency bin when \( \rho(t_0) \geq \rho_\text{TTH} \) and reject it when \( \rho(t_0) < \rho_\text{TTH} \), where \( \rho_\text{TTH} = \cos(\omega_0 d \sin \theta_\text{TTH} / C_0) \). By defining the threshold \( \theta_\text{TTH} \), the direction of the sound source can be constrained via spatial mask.

III. SPECTRO-TEMPORAL MODULATION FILTERING ALGORITHM

Human speech contains rich dynamic content in the spectral and temporal domain. Many features, like pitch,
formant and timbre, are used to characterize speech. Although our acoustic environment is complex, the speech can be described as an amplitude envelope fluctuating rhythmically in time and temporally in frequency with phase information. The spectrogram is the most popular expression for sound, but some important characteristics are hidden in this representation domain. The spectrogram is time-varying spectral representation and calculated through the short-time Fourier transform (STFT). In spectrogram analysis, the bandwidth of spectral is determined by the choice of window length according to Heisenberg’s uncertainty principle. If the time-length of speech with slow temporal variation is larger than window length, some temporal information is lost through this transformation. In order to describe the sound in a suitable representation and make it closer to humans’ perception in brain, the cochleagram and the spectro-temporal receptive field (STRF) have been introduced to represent the sound.

The cochleagram which is also called auditory spectrogram represents the neural firing rates produced by basilar membrane and inner hair cells of the cochlea due to the sound pressure ear into the outer ear [12]. At first approach, the cochleagram is used to display the perceived energy distribution of a sound input at time-frequency space and make sound components audible. At second stage, the STRF have been used as linear approximation to the sound transform from stimulus to auditory neural response. The STRF have the properties of quadrant separability and can be measured [13]. It can be decomposed into the product of a purely temporal impulse response and a purely spectral response field [6]. In this way, the auditory spectrogram can be reconstructed by cortical representation, also called scale-rate representation [14]. It is a linear transformation and invertible. The model of this transformation is defined as follows [15]

$$z(t;f;\omega,\Omega) = y(t,f)^*h_{\text{STRF}}(t;f;\omega,\Omega)$$

where $y(t,f)$ is the cochleagram, $h_{\text{STRF}}(t;f;\omega,\Omega)$ is the spectro-temporal impulse response to represent the STRF, $z(t;f;\omega,\Omega)$ is the scale-scale representation and $*f$ denotes the convolution with respect to both $t$ and $f$ domain.

Here, the spectral-temporal impulse response can be separated as the temporal and spectral impulse responses

$$h_{\text{STRF}}(t;f;\omega,\Omega) = h_T(t;\omega,\phi) * h_S(f;\Omega,\phi)$$

where $h_T$ is the temporal impulse response and $h_S$ is the spectral impulse response.

Both can be assumed as analytical signal and be expressed by Hilbert transform.

$$h_T(t;\omega,\phi) = u(t;\omega) \cos \phi + \dot{u}(t;\omega) \sin \phi = |u(t;\omega)| e^{i\phi}$$

$$h_S(f;\Omega,\phi) = v(f;\Omega) \cos \phi + \dot{v}(f;\Omega) \sin \phi = |v(f;\Omega)| e^{i\phi}$$

where the $\omega$ is defined as rate along the time axis and measured in cycles/sec (Hz) which is characterized by the temporal phase $\phi$. The $\Omega$ is referred to as scale along the frequency axis and measured in cycles/octave which is characterized by spectral phase $\Phi$.

The Hilbert transform is defined as

$$\hat{f}(\tau) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(\tau)}{\tau - \chi} \, d\tau$$

The seed functions, $u$ and $v$, can be chosen for approximating the spectro-temporal impulse response separated from the STRF.

$$u(t;\omega) = t^3 e^{-\omega t} \cos(2\pi t); \quad v(f;\Omega) = (1-f^2)^{1/2} e^{-f^2/2}$$

where $u$ is modeled by one gamma probability density function and $v$ is modeled by a Gabor-like function.

The transformation of (4) is calculated by the NSL Matlab tool box [16]. In scale-rate representation, it is shown that characteristics of speech and noise at spectral and temporal modulation domain are different. From Fig. 3, it is shown that the speech have more energy at slow temporal modulation (rate) below 8 Hz. According to the feature, the emphasis of modulation filter can be set as low pass filter or band pass filter that passes only sound with temporal modulation characteristics of speech [17]. The filters are designed for a range, scales from 0.25 to 8 cycles per octave and rates from 2 to 32 Hz, which corresponds to the spectro-temporal response field (STRF) in the auditory cortex processing [7]. The speech can be enhanced by filtering the modulation spectrum of the noise.

Figure 3. Representations of clean speech and noise by cochleagram and scale-rate plot.
IV. EXPERIMENTAL SETUP AND RESULTS

We assume that algorithm is operated in noisy environment without reverberation, like outdoor environment. Here, the source of speech source is from Aurora 2.0 and the noise database is from NOIZEUS [18]. In order to compare the conventional spectral subtract algorithm and the proposed BIST algorithm. Two kinds of conditions are setup to simulate the received noisy speech. In condition 1, the clean speech is at front side (azimuth angle, $\theta = 0^\circ$) with noise at right side ($\theta = 90^\circ$). Both are 1m away from the single microphone. Fig. 4(a) shows the auditory spectrogram of noisy speech measured at this position with SNR-10dB. It can be found that the speech signal is almost covered by noise. The spectrum subtraction algorithm [19] is applied to enhance the speech as shown in Fig. 4(b). Clearly, parts of some important feature are loss in this process. In condition 2, the spatial sounds are synthesized with HRTF. The HRTF database is from the measurements of a KEMAR dummy-head microphone at MIT [20]. A speech at front side ($\theta = 0^\circ$) with a noise source at right side ($\theta = 90^\circ$) are 1m away from the center of artificial head (KEMAR dummy-head), where the distance between both ear is set as 15.2cm. The received signals from left and right microphones on ear carry the spatial information. The signals are processed by the BIST algorithm and the detail flow is shown in Fig. 1. And, Fig. 4(c) describes the auditory spectrogram of directional sound processed by spatial mask. Here, the threshold angle ($\theta_{TH}$) is set as 20$^\circ$ at one side and a window length of 50 ms is chosen for model calculation. It is obvious shown that part of lower frequency components are reduced at this stage. Fig. 4(d) describes the cochleagram of speech enhanced by the BIST algorithm. The spectro-temporal modulation filter is used to separate the speech and noise depended on their acoustic feature. To choose suitable parameters of scale and rate is important the performance of enhancement. According to the characteristic of noisy speech in our experiments, the low pass spectro-temporal modulation filter is used to separate speech and noise. This work is based on the STRFpak developed by T. M. Elliott and FE. Theunissen [21]. The cut-off spectral modulation (scale) is chosen from 6 to 8 cycles/ octave and temporal modulation (rate) is chosen from 8 to 16 Hz (cycles/second) to find the optimization result for speech enhancement. In our experience, the optimal value is depending on the type of noise sources and SNR conditions.

![Figure 4. Auditory spectrograms of noisy speech and processed signals through different algorithms](image)

To examine the effectiveness of the speech enhancement algorithm, the objective test is used to compare the quality with the conventional spectral subtraction algorithms and BISE algorithms. The perceptual evaluation of speech quality (PESQ) has been used successfully for objective evaluation and is one recommend by ITU-T for speech quality assessment of narrow-band speech code [22]. The range of the PESQ score is –0.5 to 4.5, although for most cases the output range will be a score between 1.0 and 4.5. Most of speech enhancement algorithm is verified from SNR-5dB to 20dB. In order to evaluate the performance of BIST algorithm in practical harsh environments, conditions between SNR 0dB and -10dB are chosen for test, as shown in Table 1. The original noisy speech means the signals measured at condition 1 without any processing. It is obvious shown that the speech quality is low at SNR-10dB and the spectral subtract is not useful to enhance it at this conditions. Nevertheless, the BIST algorithm can enhance the quality about 20% even at SNR-10dB.
Furthermore, in order to make the BIST algorithm can be realized in practical application, the microphones are integrated in the earphone with opening hole, as shown in Fig. 5(b). When a user wears this type of earphone in noisy environment, the binaural signal can be received with HRTF characteristics. In noisy environments, the speech of target source can be enhancement by BIST algorithm and replayed by earphone. Nevertheless, processing time of BIST algorithm and the acoustic leakage are still the issues for real time application.

![Earphone and Microphone](image)

Figure 5. The system setup for capturing the binaural sound in practical application

### V. CONCLUSIONS

We have demonstrated a novel approach to improve the speech intelligibility in negative SNR environments. In this method, the integration of binaural auditory processing and auditory cortical processing functions as mechanism of human auditory perception. It is proved that the adaption of binaural hearing as spatial mask at the first stage is helpful to reduce the complexity of multi sound sources. Furthermore, the speech can be extracted from noisy signal based on their distinct spectro-temporal modulation patterns. The objective tests proved the effectiveness of this method in speech enhancement. However, the effect of reverberation and the test set of noise source are still not comprehensive enough in this study. Besides, the design of an optimal modulation filter applied for any situation is an important step for the future study. In summary, the placement of sensor ends and implantation of the BIST algorithm in real-time processing are important to increase the robustness of human computer interface.

### ACKNOWLEDGEMENT

This work was supported in part by the National Science Council of the Republic of China under Grant No. NSC 100-2221-E-006-248-MY3. The authors would like to thank Dr. Frederic Theunissen, who is Professor of Psychology and Neurosciences Institute, UC Berkeley, USA, for providing the modulation toolbox of STRF Pak.

### REFERENCES


