FREQUENCY DOMAIN BINAURAL MODEL AS THE FRONT END OF SPEECH RECOGNITION SYSTEM

Tsuyoshi Usagawa, Kenji Sakai, Masanao Ebata

Department of Computer Science, Kumamoto University, Kumamoto 860-8555, Japan

ABSTRACT

As well known as the cocktail party effect, we can communicate with others in very noisy environment such as a railway station or bus stop beside a busy street. This effect comes from many cues, but the binaural hearing takes one of the principle roles.

In this paper, the frequency domain binaural model is introduced. The proposed model is the revised one of the former time domain model which calculates the interaural crosscorrelation. The new model requires the less computational load and has the comparable performance. It is based on the FFT analysis and uses the cross-power spectrum to obtained interaural phase difference. Because of the efficiency of FFT analysis, the proposed model consumes only about from 1/10 to 1/20 of processing time of the time domain model.

The performance of models is examined not only in the isolated word speech recognition task and but also in the speech enhancement task. As the results of experiment, the improvement of robustness in speech recognition task corresponds to about 15 $\sim$ 20dB when the surrounding noise is white noise. That is a few decibell better than one obtained by the time domain model. However, when the surrounding noise is speech, the improvement decreases to 10 $\sim$ 15dB and it is almost the same as one of the time domain model. In addition, the proposed model can reproduce the signal component from the specified direction as the binaural signal; ie. it works as the speech enhancer.

1. INTRODUCTION

A small vocabulary speaker-dependent recognition system can be made with a few chips over a decade, but the actual implementations are not so many. Among several reasons for the limitation of application fields, the performance degradation due to surrounding noise is the most serious problem. For most of the systems, the performance begins to degrade as SNR (speech signal to surrounding noise ratio) decreases to 20dB or less.

There are many researches to realize a robust speech recognition system against surrounding noise. Systems proposed by those researches need some extent of *a priori* information about signal or noise, or have obvious limitations such as the physical size of sensor array. Among those, the adaptive noise canceling is very effective when it is applicable[1].

As manifested in the Cocktail-Party effect, humans have an ability to separate signals even they are embedded in high level surrounding noise. This effect is based not only on localization ability due to binaural cues, but also on the familiarity of signal characteristics and to some extend of expectation and estimation of information carried by signal. However, as the primary functionality, localization using binaural cue plays important roles. Over decades, psycho-acoustical research has been performed for both, an experimental and a theoretical point of view, and basic cues related to sound localization were revealed.

The group of Ruhr-University of Bochum led by Prof. Blauert constructs a binaural computer model, which is designed to simulate various psycho-acoustical phenomenon related to sound localization[2][3][4]. This model calculates an interaural crosscorrelation between inputs of left and right ears for each subband. Comparing to other binaural or localization models, this model has two special features. The first one is the contra-lateral inhibition which is introduced to reduce ambiguity of narrowband cross-correlation function. Secondly, the compensation of Interaural Level Difference (ILD) and Interaural Time Difference (ITD) is took into account based on psycho-acoustical evidence. This model was expanded to so-called "Cocktail-Party-Processor" by Bodden, and it is evaluated in various application fields such as a speech enhancer for normal and hearing impaired persons and a phoneme-base speech recognizer[5].

The authors have proposed the speech recognition system utilizing the time domain binaural model as the directionally selective front end[6]. Because the original binaural model aimed to simulate the binaural hearing phenomenon with high accuracy, some extent of simplification was done to adopt as a front end of the speech recognition system. Overall improvements of the robustness against the surrounding noise is 15dB or more in SNR comparing with one obtained by monaural front end. Although

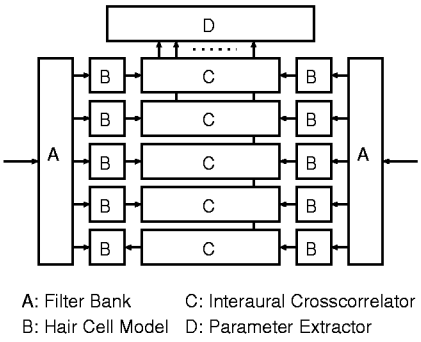


Figure 1: Time Domain Binaural Model

the obtained performance is high, the computational cost of the time domain model is very high. The time consumption of the model is an order of 100 times of one for monaural front end.

In this paper, the frequency domain binaural model is introduced. The proposed model requires the less computational load and has the improved performance comparing with the former time–domain binaural model. The proposed model is based on the FFT analysis and uses the cross-power spectrum to obtained interaural phase difference. The performance of the new model is examined not only in the isolated word recognition task as a remote control system of television set, but also in the speech enhancement task.

2. THE TIME DOMAIN BINAURAL FRONT-END

Figure 1 shows the abstract structure of the binaural front-end. The influence of the outer ear is considered by the arrangement of the recording microphones, and the middle ears are usually neglected for signal processing purposes.

- **Filter Bank**: The filter bank simulates the frequency characteristics of auditory filter in cochlea. The time domain model utilizes 16 channels of a gamma-tone filter bank to cover the frequency range from 50Hz to 4.5kHz.
- **Hair Cell Model**: Simple hair cell model consists of a half-wave rectifier and a low-pass filter whose cutoff frequency is 800Hz.
- **Interaural Crosscorrelator**: The interaural crosscorrelator (mentioned as "IACC" here after) is the main module of the binaural front-end and it is constructed mainly with two delay lines and correlator. Left and right subband signals are fed into delay lines and the outputs of IACC are calculated at every delay elements as the multiplied value of two delay line elements. To reduce the multiple peaks in IACC, the contra–lateral inhibition inhibits the further propagation in delay lines when the peak is obtained.
- **Parameter extraction**: Speech parameters for pattern matching are extracted from the IACC output.

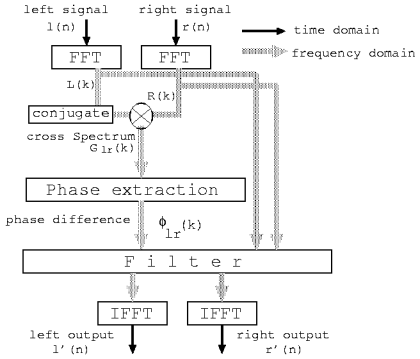


Figure 2: Frequency Domain Binaural Model

On the assumption that the direction of incidence of the target speech signal is *a priori* known, the proposed parameter extractor observes IACC components corresponding to the specified direction and detects the speech candidate time frames by itself.

3. FREQUENCY DOMAIN BINAURAL MODEL

The aim of the frequency domain binaural model is to obtain the similar property of the time domain model with sufficiently low computational cost. To reduce the computational load, the detection of interaural time difference is done in the frequency domain using the cross spectrum. Furthermore, the IACC output can be replaced with the similar output in the interaural phase difference. Figure 2 shows the block diagram of the proposed frequency domain binaural model.

3.1. FFT Analysis

The time domain signals are transformed into frequency domain using 64 taps FFT where the sampling frequency is $10kHz$. The overlap of FFT frame is set to 32 taps, $3.2ms$, and each frame of signal is fed into hanning window before transform. The number of FFT tap, 64, is selected so that the corresponding frequency band is comparable with one of gamma tone filter in the time domain binaural model around $1kHz$. Also the temporal resolution, $6.4ms$, is also concerned.

3.2. Interaural Phase Difference

The interaural phase difference is obtained through the cross spectral analysis in each frequency bin. When the spectrum of left and right signals are denoted as $S_l(k)$ and $S_r(k)$, respectively, the cross spectrum $G_{lr}(k)$ is given as,

$$G_{lr}(k) = S_l^*(k)S_r(k). \quad (1)$$

The phase difference of this frequency bin, $\phi_{lr}(k)$, is obtained from the cross spectrum as,

$$\phi_{lr}(k) = \tan^{-1}\{\frac{Im(G_{lr}(k))}{Re(G_{lr}(k))}\} \quad (2)$$

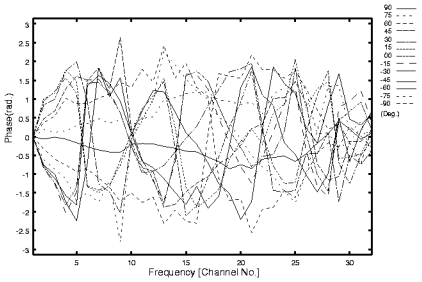


Figure 3: Characteristics of Phase Difference between Left and Right signals for specified direction

where $Im(.)$ and $Re(.)$ take the imaginary and real part of complex value.

Note that the range of obtained phase difference is rounded within $-\pi < \phi_{lr}(k) \leq +\pi$ while the binaural time differences are in the range of $-1ms \sim +1ms$ for all frequency range. The limited range in phase difference means the aliases when phase difference is converted into time difference in the frequency range higher than about $1000Hz = 1/0.001s$. Because of this alias phenomenon, the signal components arrived from the specified direction are selected based on the binaural phase difference in the proposed frequency model while the time domain model uses the binaural time difference.

Figure 3 shows the plot of the phase difference between left and right measured head related transfer functions (HRTF) of head torso (B&K 0000) in anechoic room. The abscissa shows the channel number corresponding to the frequency, the ordinate shows the phase difference in degree, and the lines correspond for the direction of arrival for -90 degree (left side) to $+90$ degree (right side) while the front of the torso is defined as 0 degree. Note that in the higher frequency range, there are many closings of plots which are aliasing in phase difference.

3.3. Paramter Extraction

Parameters used for speech recognition task are obtained as follows. At first, the standard phase difference $\hat{\phi}_{lr}(k)$ for specific direction in database is retrieved. The observed phase difference, $\phi_{lr}(k)$, is compared with $\hat{\phi}_{lr}(k)$ for each channel k. When $\phi_{lr}(k)$ satisfies the following condition,

$$|\phi_{lr}(k) - \hat{\phi}_{lr}(k)| \leq \theta, \qquad (3)$$

the component of cross-spectrum $G_{lr}(k)$ is assigned to the one from the specified direction. Note that θ is the threshold for detection in phase difference. Otherwise, it is neglected. Parameters are given as the 32 element vector, which consist of magnitudes of obtained cross–spectrum $|G_{lr}(k)|$ where channel k satisfies Eq.(3), or zeros for other cases.

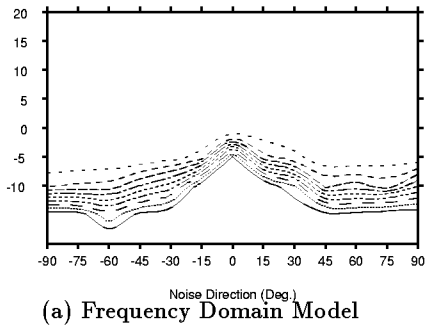
3.4. Signal Extraction

Signal components assigned from the specific direction are extracted and reproduced as a binaural signal as follows. As the same as the parameter extraction, the signal components, $S_l(k)$ and $S_r(k)$, arrived from the specified direction are collected based on the Eq.(3) in frequency domain. And the obtained spectra are transformed into time domain separately using IFFT. Note that concatenated time domain frames are added with appropriate windowing because each of analyzing frame, which is $6.4ms$ in length, is overlapped $3.2ms$ with the neighbor frames.

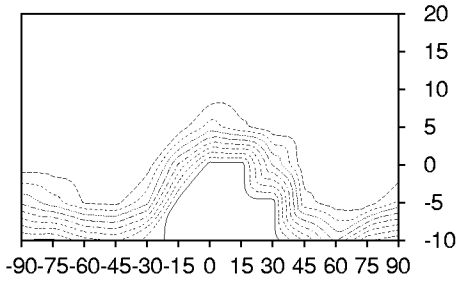
4. SIMULATIONS

4.1. Speech Recognition

The performance as a binaural front-end is evaluated as the speech recognition rate for an isolated word speaker dependent condition at various SNRs and directions of speech and noise source. The DP (DTW) matching is used to recognize 22 Japanese words which are selected to simulate a remote control system of television set. The SNR varies from -10dB to $+20$dB by 5dB step and the direction of incidence of the noise source varies from -90degree to $+90$degree by 15degree step. The SNR should be $+15$dB or more to get 90% of correct recognition for white noise when LPC-cepstrum obtained by monaural front-end is used as speech parameters for the task.



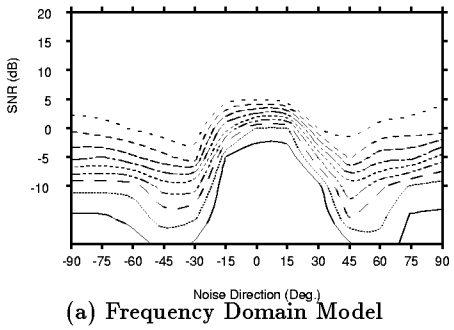
(a) Frequency Domain Model



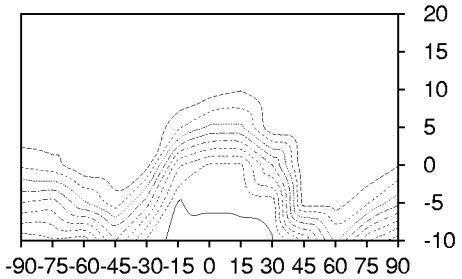
(b) Time Domain Model

Figure 4: Results of recognition (White noise source at 0 degree)

Figure 4 shows the obtained speech recognition rate as a contour map when a speech signal arrives from 0degree



(a) Frequency Domain Model



(b) Time Domain Model

Figure 5: Results of recognition (Male voice as noise at 0 degree)

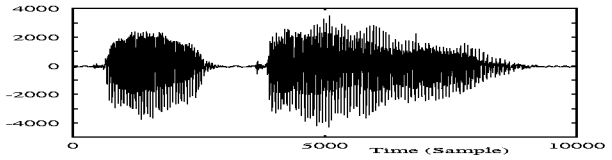
and a white noise source emits at various directions. The ordinate and the abscissa represent the SNR and the direction of incidence of the noise source, respectively. The contour lines are drawn for each 10% of recognition ratio, and the top most line corresponds to 90% correct recognition. As shown in Fig.4, even if the directions of speech and noise are the same, the frequency domain binaural front-end works well when the SNR is more than 0dB. Comparing with Fig. 4(a) and (b), which show the result obtained by the frequency domain and time domain models, respectively, there is up $8dB$ improvement where angle between -45degree and $+45$degree. Figure 5 shows the results obtained with noise of male speech. Comparing the results for white noise case, the improvement against the time domain model is decreased, but the overall performance both models are almost the same.

4.2. Speech Enhancement

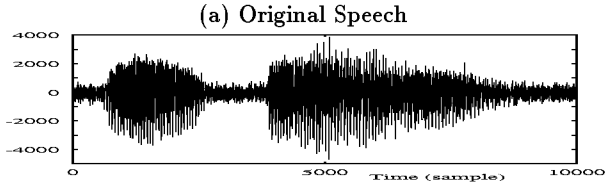
Figure 6 shows a demonstration of speech enhancement. The speech uttered /tiikeiyu/ "TKU" is arrived from the front and the white noise is arrived from $+45$degree and SNR is set to $+5dB$. Figures (a), (b) and (c) show the waveforms of source signal, observed signal at right ear, and the enhanced signal, respectively. As shown in Figure (c), reduction of white noise is obvious.

5. CONCLUSION

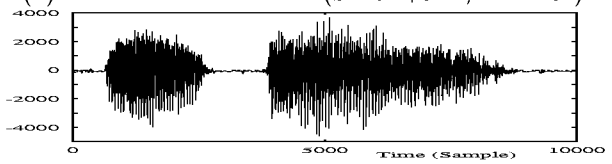
In this paper the frequency domain binaural model is proposed and it works both as the front-end of speech recognizer and as the speech enhancer. Comparing to the results obtained with conventional recognition systems, we found out that the frequency domain binaural front-end recovers more than $0dB$ in SNR when noise arrives from the same direction as the speech, and that it recovers $20dB$ at its best. Also in some case the performance is better than that of time domain model. The speech enhancer is also configured using the proposed binaural model and it shows the sufficient noise reduction.



(a) Original Speech



(b) Mixed with white noise (SNR=+5dB, from 45°)



(c) Enhanced Speech

Figure 6: Demonstration of speech enhancement. (a) Speech (0degree) signal, (b) mixed signal with white noise (SNR=0dB, +45degree) (c) enhanced speech.

6. REFERENCES

1. T. Usagawa, Y. Morita and M. Ebata, "A configuration of remote control system using speech within a priori known noise", J. Acoust. Soc. Jpn. (E), Vol.13, No.5, 295–300 (1992)
2. J. Blauert, *Spatial Hearing – The Psychophysics of Human Sound Localization, Revised Edition*, MIT Press, LONDON, 1996
3. W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals", J.Acoust.Soc.Am., 80(6), 1608–1622 (1986)
4. W. Gaik,"Combined evaluation of interaural time and intensity differences : Psychoacoustic results and computer modeling", J.Acoust.Soc.Am., 94(1), 98–110 (1993)
5. M. Bodden,"Modeling human sound-source localization and the cocktail-party-effect", Acta Acoustica, 1, 43–55 (1993)
6. T. Usagawa, M. Bodden, K. Rateitschek, "A binaural model as a front-end for isolated word recognition," Proc. ICSLP'96, 2352-2355 (1996)