

A BINAURAL MODEL AS A FRONT-END FOR ISOLATED WORD RECOGNITION

*Tsuyoshi Usagawa*¹

*Markus Bodden*²

*Klaus Rateitschek*²

¹ Kumamoto University, 2-39-1 Kurokami, Kumamoto 860, JAPAN

² Ruhr University of Bochum, 44801 Bochum, F.R.Germany

ABSTRACT

Small vocabulary isolated word speech recognition can be implemented on relative small hardware. Although the recognition problem is more or less solved in noise-free situations, the general application is hindered because of the dramatic decrease of performance in noisy environments, especially for hands-free applications. In this paper a binaural front-end for speech recognition is presented. This binaural model, which was originally developed at Ruhr-University of Bochum in Germany, allows for an effective reduction of interfering noises of any kind. Besides stationary noises also concurrent speech signals can be suppressed. The original model was designed as a precise computer model of the human binaural auditory system and can explain a variety of psycho-acoustical phenomenon. Besides those abilities the model offers sharp directional selectivity which is superior to those obtained with directional microphones. We simplified this sophisticated model by adapting it to the specific task and use the peak position and the peak level of the binaural activity pattern for each frequency band as a parameter for pattern matching. The performance was evaluated in the form of recognition rates for a variety of difference noisy environments. The results show that the binaural front-end leads to a significant improvement in recognition rates corresponding to an enhancement of over 20dB in SNR in most cases.

1. INTRODUCTION

A small vocabulary speaker-dependent recognition system can be made with a few chips over a decade. However, in contrast to enthusiastic expectations the past systems are only installed yet in just a few application fields. Among several reasons for the limitation of application fields, the performance degradation due to surrounding noise is the most serious problem, especially for hand-free applications. 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. Techniques used are divided into two categories; the first one is researches on parameter derivation and pattern matching, and the other is the signal process such as a speech enhancement or noise reduction. Most of the former works assume the stationar-

ity of the noise and are based on statistical characteristics of the noise, which are measured before recognition is performed. Beside this restriction, the performance is usually not good enough in the circumstances in which the SNR is 10dB or less. The latter type of research includes multiple microphone systems controlling the beam or dips of a directional pattern. Due to physical limitations, performance of directional selectivity at low frequencies such as the pitch of speech is not high enough, and sometimes the number of noise sources to be suppressed defines the number of microphones.

To overcome these restriction, adaptive noise canceling is very effective when it is applicable[1]. In combination with a parameter extraction method using a masking model, the resulting system works even if SNR is less than -10dB.

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 extent 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 inter-aural cross-correlation 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 narrow-band cross-correlation function. Secondly, the compensation of Interaural Level Difference (ILD) and Interaural Time Difference (ITD) is taken 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].

This paper describes the binaural front-end for an isolated-word small vocabulary speech recognizer. The front-end pre-

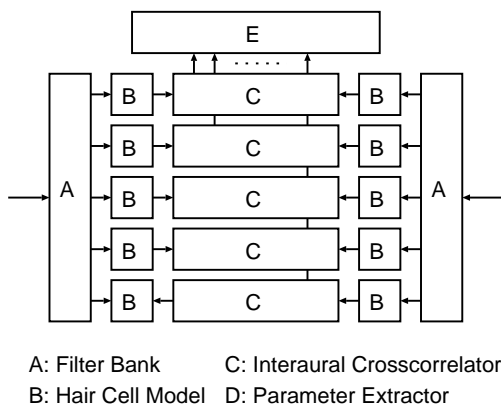


Figure 1: Configuration of Binaural Front-End

sented here is derived from the above mentioned binaural model (referred as the original model here after). Because the original model is designed to simulate a variety of psycho-acoustical phenomenon precisely, it is somewhat redundant for a front-end. For a front-end, speech parameters has to be extracted in vector form from surrounding noise, so that parameter extractor with automatic speech detection has to be introduced.

The evaluation of performance is carried out by simulating the task of a television remote control system using Japanese vocabulary.

2. CONFIGURATION OF 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.

The cochlear is understood as the organ to analyze frequency components of the input signal, so that it is represented as the bank of band-pass filters (A). Subband signal is fed to hair cells (B) that is a transducer of mechanical vibration of basilar membrane to electrical discharges. In the binaural model, the functionality of hair cells is simulated as a converter from signal level to the probability of firing. Both channels of converted signals are fed into interaural cross-correlator (C) which extract the coincidence of inputs.

As the final stage of the front-end, the parameter extractor (D) is positioned after the cross-correlator. He extracts speech parameters in the form of a vector for each time frame from the output of the cross-correlator.

2.1. Filter Bank

The filter bank simulates the frequency characteristics of cochlear. The proposed front-end utilized 16 channels of a gamma-tone filter bank to cover the frequency range from 50Hz to 4.5kHz. Figure 2 shows the gain characteristics of then implemented filter bank.

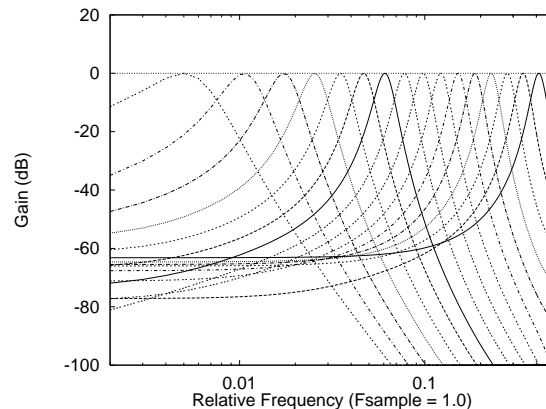


Figure 2: Gain Characteristics of Gamma-tone Filter Bank

2.2. Hair Cell Model

In the binaural front-end as well as in the original model, a very simplified hair cell model is used. The subband signals are half-wave rectified and fed into low-pass filters whose cutoff frequency is set to 800Hz. Then the square root of the low-pass filter outputs is calculated and the output of the hair-cell model is obtained after normalization to the maximal level.

2.3. Inter-Aural Cross-Correlator

The interaural cross-correlator (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.

Let us assume an index of delay line as m ($-M \leq m \leq M$) where M shows the maximum tap index of the delay lines. If the sampling interval is shown as T_s (s), the delay time between each tap $\Delta\tau$ is set as $\Delta\tau = 2T_s$ and $M = 0.001/\Delta\tau$. Because of this relationship, in order to get consistency of sampling interval, we need to insert null input between each sample. Let us assume $r(m, n)$ and $l(m, n)$ represent the signal of right and left delay lines, respectively, of tap m at time n . The input of delay line are normalized to the maximum value on the delay lines which are obtained by averaging over time with a relatively short time constant T_{intIn} .

The instantaneous cross-correlation component $k(m, n)$ is given as,

$$k(m, n) = r(m, n)l(m, n) \quad (1)$$

And the output of IACC $\Psi(m, n)$ is given as the time-averaged instantaneous cross-correlation component as follows,

$$\Psi(m, n) = \sum_{i=-\infty}^n k(m, n) e^{-(n-i)/T_{inh}} \quad (2)$$

where T_{inh} is a time constant.

In order to obtain single peaks in the cross-correlation for a band limited signal, the original model was extended by contra-lateral inhibition. The inhibition coefficients $i_r(m, n)$ and $i_l(m, n)$ are defined as the constant within the range of

[0,1], so that the update equations of is given as follows,

$$r(m+1, n+1) = r(m, n) i_r(m, n) \quad (3)$$

$$l(m-1, n+1) = l(m, n) i_l(m, n) \quad (4)$$

Contra-lateral inhibition consists of inhibition due to stationary signal, $i_{r,s}$ and $i_{l,s}$, and inhibition due to non-stationary signal, $i_{r,d}$ and $i_{l,d}$ as follows,

$$i_r(m, n) = i_{r,s}(m, n) i_{r,d}(m, n) \quad (5)$$

$$i_l(m, n) = i_{l,s}(m, n) i_{l,d}(m, n). \quad (6)$$

Contra-lateral inhibition for stationary component controls the characteristics for interaural level difference (ILD). And inhibition for non-stationary component is defined with an output of non-linear low pass filter, which has a larger time constant for onsets in order to simulate the precedence effect.

Because natural combinations of ILDs and ITDs can be observed for head related transfer functions, Gaik extended the original binaural model to consider these combinations[4]. The extension introduces an additional weighting to the update equation to translate ILDs into ITDs as follows,

$$r(m+1, n+1) = r(m, n) w(-m) i_r(m, n) \quad (7)$$

$$l(m-1, n+1) = l(m, n) w(m) i_l(m, n) \quad (8)$$

These extra weights satisfy the following relationship with the level difference between channels $\Delta L_w(m)$.

$$\sum_{i=-M}^m w(-i) - \sum_{i=m}^M w(i) = -\Delta L(m) \quad (9)$$

Although $\Delta L(m)$ is determined from measurements for a specific subject and tuned individually in the original model, it is parameterized with the following function in the binaural front-end.

$$dL(dT) = P_{max}(1 - e^{-P_{slope}(dT - P_{offset})}) \quad (10)$$

This function describes the relationship between time difference dT and level difference dL using three parameters; maximum level difference P_{max} (dB), slope of relationship P_{slope} (dB/ms) and offset of relationship P_{offset} (s). Based on this function, the weights are given as follows.

$$w(m) = 10^{dL(m/(2fs)) - dL((m+1)/(2fs))} \quad (11)$$

2.4. Parameter extraction

Speech parameters for pattern matching are extracted from the IACC. 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.

For each sample, the algorithm searches for a peak of the IACC which the specified range of the interaural delay, and the peak magnitude is estimated based on the quadratic interpolation technique.

Although an IACC pattern is obtained for each sampling interval, the peak magnitude is averaged over a certain time frame. In the binaural front-end, the frame length is set to 20ms with an overlap of 10ms between subsequent frames.

As a parameter vector for a frame, the binaural front-end uses the peak magnitudes for each channel and the averaged magnitude of all peaks. All of parameters are in dB scale.

To detect speech candidate frames, the average peak magnitude is compared to a fixed threshold. Because of normalization at both, the hair cell model and the input stage of the interaural cross-correlator, the fixed threshold works well for a wide range of SNRs. However, the so-called pre-triggering is necessary to pick up a frame which may contain a first consonant of a target word. In order to prevent the detector from splitting up single words, the binaural front-end continues to pick up 10 more frames after the average peak is lower than the threshold.

3. PERFORMANCE OF FRONT-END

3.1. Condition of Speech Recognition

The performance of the binaural front-end is evaluated as the speech recognition rate for a 22 isolated word speaker dependent condition at various SNRs and directions of speech and noise source. Vocabulary is selected to simulate a remote control system of television set. The SNR varies from -10dB to +20dB by 5dB step and the direction of incidence of the noise source varies from -90degree to +90degree by 15degree step. The basic conditions of the experiment are summarized in Table 1. Please note that the SNR should be 15dB or more to get 90% of correct recognition for white noise when normal LPC-cepstrum is used as speech parameter vector.

Table 1: Experimental Condition of Speech Recognition

| | |
|-------------------------|-----------------------------------|
| Matching Algorithm | DP (DTW) |
| Parameter/frame | 15 (14ch.+Overall) |
| Sampling Frequency | 10kHz |
| Analyzing Frame Length | 20ms |
| Analyzing Frame Shift | 10ms |
| Speaker | Male |
| No. of Vocabulary words | 22 Japanese words |
| Range of word length | 1 mora - 4 mora 80 ms - 420 ms |

3.2. Experimental Results

Figure 3 shows the resulting speech recognition rate as a contour map when a speech signal arrives from 0degree (in front of head) and a single 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. As shown in Fig.3, even if the directions of speech and noise are the same, the binaural front-end works well when the SNR is less than +10dB. And, as expected, when the directional difference between speech and noise is

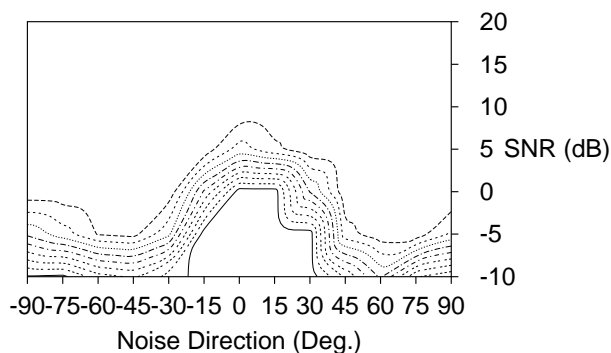


Figure 3: Results of recognition(White noise source at 0 degree)

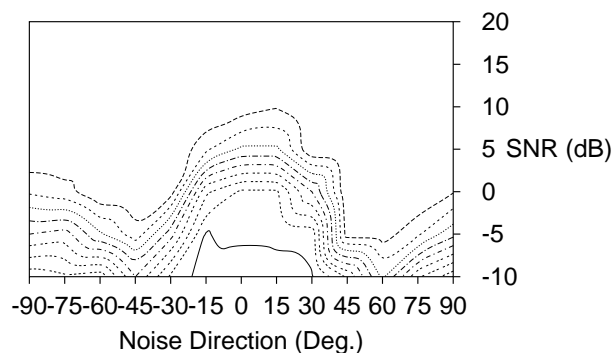


Figure 5: Results of recognition (Male voice as noise at 0 degree)

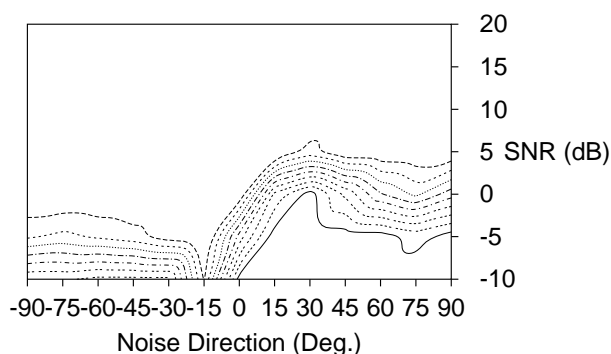


Figure 4: Results of recognition (White noise source at 30 degree)

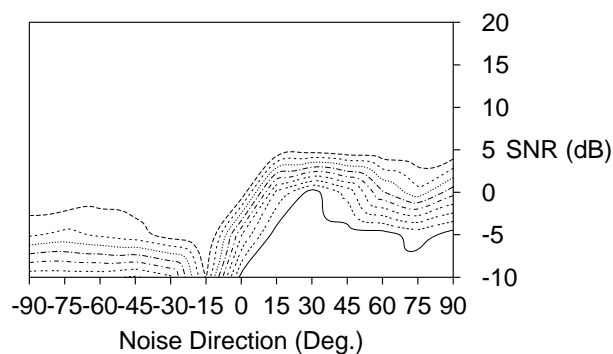


Figure 6: Results of recognition (White noise at 0 degree using template obtained 30degree)

more than 45degree, the 90% contour is around -5 dB of SNR, that is, the front-end recovers about 20dB of SNR.

The results obtained when the signal source is located at $+30$ degree are shown in Fig.4. The shape of contour is different from the one in Fig.3, however, the improvement in SNR stays almost the same.

Figure 5 shows the results obtained with noise of male speech. Comparing the results shown in Fig. 5 with results obtained by using an LPC-cepstrum, the binaural front-end improves recognition rates by an equivalent in SNR of more than 5dB if speech and noise arrive from the same direction and by more than 15dB if the directional difference between speech and noise is more than 45degree.

To confirm the robustness of the front-end, the recognition performance is measured using the template which has been extracted from a different direction. Figure 6 shows the recognition results when the speech signal arrives from 30degree, but the template has been obtained for a signal from 0 degree. Comparing to Fig.3 only very small difference can be observed.

4. CONCLUSION

In this paper we described a binaural front-end for speech recognition in noisy environment. Comparing to the results obtained with conventional recognition systems, we found

out that the binaural front-end recovers more than 5dB in SNR when noise arrives from the same direction as the speech, and that it recovers 20dB at its best. Although the computational load of the proposed binaural front-end is not small, the front-end is very effective for noisy environment.

5. 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* MIT Press, 1983
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)