LATERAL INHIBITION MECHANISM IN COMPUTATIONAL AUDITORY MODEL AND IT'S APPLICATION IN ROBUST SPEECH RECOGNITION

Lu Xugang Li Gang Wang Lipo
Nanyang Technological University, School of EEE, Workstation Resource Lab,
Nanyang Avenue, 639798, Singapore
Email: EXGLU@ntu.edu.sg

Abstract: In auditory neural system, lateral inhibition mechanism is very common, such as in cochlear nucleus, auditory cortex, etc. The function of this lateral inhibition is to sharpen the contrast of the temporal and spatial structures, thus prominent features of stimulation in spatial and temporal domains can be enhanced. In this paper, a new mathematical model based on lateral inhibition is proposed. Traditional feature MFCC and auditory feature AFCC (Auditory Frequency Cepstral Coefficient) are processed by this model, new features can be gotten as MFCCI and AFCCI, experiments of the new features with HMM show that new features can improve the robustness of the recognition rate in noisy condition.

1. INTRODUCTION

Lateral inhibition is a very general phenomenon in neural system, the most famous theory of lateral inhibition is discussed by Hartline in visual system, the function of lateral inhibition can sharpen edges of the detected objects, thus the contrast of the profile of objects can be enhanced. In auditory system, lateral inhibition effect is also very general, such as in the inferior colliculus, auditory cortex, especially in the cochlear nucleus. When one neuron is excited by a stimulation, the firing rate can be suppressed by the neighboring neuron's excitation. Seeing from neural tuning curve, it is very clear that the excited region is always circled by inhibition region, such as in the

0-7803-6278-0/00$10.00 (C) IEEE

following fig.1

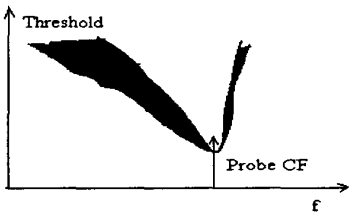

Fig.1 Neural tuning curve, an excited region is circled by a inhibition region(black region)

Of course, this lateral inhibition can sharpen the contrast of changing part of the spectral structure. This is a very complex nonlinear phenomenon which relates to other nonlinear mechanisms, such as two tone suppression, masking effect, etc. But general speaking, the function of this nonlinear mechanism is to enhance the prominent parts of the coming stimulation. As to speech signal, this nonlinear mechanism can sharpen the changing parts in spatial and temporal domains, this function is discussed very detail in papers[1][2](Shamma). In speech recognition task, prominent features should be enhanced to get much more robustness in noisy condition. Of course, this lateral inhibition can be regarded as a differential operator in mathematical description, so it is sensitive to noise. Since noise can add meaningless small ripples to spectral structure, in real implementation, Mel or Bark scale spectrum is always used for the spectral representation, thus the small ripples of noise can be reduced. In this paper, lateral inhibition mechanism of auditory neural system is discussed in part 2. Based on this discussion, a mathematical model is given in part 3 , with this model, recognition experiments are given in part 4. At last, a discussion for further implementation is given.

2. THE IMPORTANTANCE OF LATERAL INHIBITION

In this paper, only the function of lateral inhibition in spatial domain is discussed, that is the function of sharpening the changing parts of spectral structure. Experiments from psychology show, in speech recognition, spectral peaks are much more important than valleys in a spectral structure[3]. How can these larger peaks come into being? It is very

clear that these larger spectral peaks are powerful harmonics of speech stimulation, such as fundamental frequency, formants, etc.. Auditory neural fibers can lock to those larger periodical parts of stimulation, the temporal structure can be controlled by the firing timbre of these frequency parts, also the average firing rate is controlled by the intensity. In noisy condition, noise always can be masked by this mechanism. What's the effects of noise to the formation of auditory spectrum? First, noise can mask the speech signal if the noise intensity is large enough, then in spectral structure, it can bring small ripples in spectrum, second, noise can disturb the temporal structure of the neural firing, thus it is difficult for neural fibers to code the speech signal in temporal mechanism, so in real auditory information processing, a certain or some certain spatial scales are used to smooth or to reduce these effects of noise, such as following auditory spectrum in Fig.2:

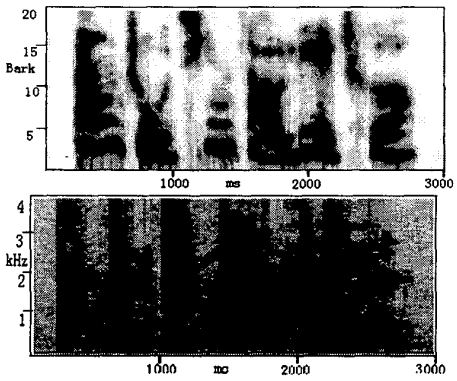

Fig.2 Top is auditory spectrum in Bark scale, bottom is FFT power spectrum

It is very clear that the auditory spectrum can be regarded as a smoothed representation of the FFT power spectrum in Bark scale, thus noise can be smoothed by this larger scale representation. Apparently, this kind of smoothness can also bring drawbacks to speech representation, such as the peaks in spectral structure is also smoothed, the contrast of the spectral structure will be degraded. So a sharpen mechanism should be used to compensate this degradation, lateral inhibition can give help for this function.

3. MODEL FOR LATERAL INHIBITION IN SPEECH RECOGNITION

Traditional simulation of lateral inhibition is always using weights to weight the spectral structure, as the following form:

$$y[k] = w[k].x[k] \qquad (1)$$

where $x[k]$ is original spectrum, $w[k]$ is the lateral inhibition weight, $y[k]$ is the output of the lateral inhibition. We hope, after processed by lateral inhibition, peaks of spectrum can be reserved, and the contrast of the local peaks can be enhanced, also the neighboring parts of local peaks should be suppressed, that is to say, the band width of local peaks should be narrowed, such as in Fig.3:

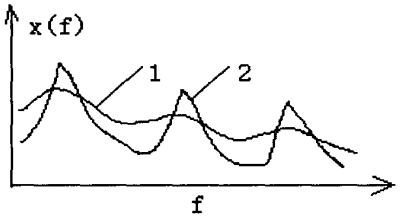

Fig.3 Spectral structure ,curve 1 is original spectral profile, curve 2 is the spectral profile processed by lateral inhibition

From Fig.3, the processed spectral structure is a sharpened form of the original spectral profile, so the lateral inhibition function should be chosen properly to adapt to this property. The following kinds of lateral inhibition functions are defined, look at Fig.4:

Fig.4 Four kinds of lateral inhibition situations

Suppose the original spectral profile is $f(x)$ (always in log scale), the processed spectral structure by lateral inhibition function is $y(x)$, where x is topological frequency axis, in Fig.4, the weights can be chosen as four kinds of lateral inhibition functions' , and the middle point is the currently processed point, for 1:

$$\frac{\partial f(x)}{\partial x} > 0 \text{,if } f(x) >= 0, \quad y(x) = \alpha 1 f(x), 0 < \alpha 1 < 1$$
$$\text{else } y(x) = \beta 1 f(x), 1 < \beta 1 < A \quad (2)$$

for 2:

$$\frac{\partial f(x)}{\partial x} = 0 \text{ and } f(x) \text{ is the local maximum,}$$
$$\text{if } f(x) >= 0, \ y(x) = \alpha 2 f(x), 1 < \alpha 2 < A$$
$$\text{else } \ y(x) = \beta 2 f(x), 0 < \beta 2 < 1 \quad (3)$$

for 3:

$$\frac{\partial f(x)}{\partial x} < 0 \text{,if } f(x) >= 0, \ y(x) = \alpha 3 f(x), 0 < \alpha 3 < 1$$
$$\text{else } y(x) = \beta 3 f(x), 1 < \beta 3 < A \quad (4)$$

for 4:

$$\frac{\partial f(x)}{\partial x} = 0 \text{ and } f(x) \text{ is the local minimum,}$$
$$\text{if } f(x) >= 0, \ y(x) = \alpha 4 f(x), 1 < \alpha 4 < A$$
$$\text{else } y(x) = \beta 4 f(x), 0 < \beta 4 < 1 \quad (5)$$

Of course, other lateral inhibition functions can be chosen, the weights can be chosen in experiments.

Experiments from psychology and physiology show, lower frequency components can suppress the higher frequency components much more easier, so in the parameters' selection in formulas(2)(3)(4)(5) , these constraints can be used, but for real speech recognition task, it depends on whether the recognition rate is better or not. In addition, because our auditory has the function of self-normalization for each frame of the incoming stimulation, so a normalization operator is used after the lateral inhibition. The following part is the application of this module.

4. IMPLEMENTATION OF LATERAL INHIBITION MODEL

Because lateral inhibition operator is sensitive to noise, so a preprocessing method is used to improve prominent features and to suppress those non-prominent features of speech signal, in this paper the cepstral liftering [4]is used.

(1) **Lateral Inhibition with MFCC.** For the convenience of comparison, traditional MFCC is used first(Fig.5), new feature processed by lateral inhibition model is as MFCCI in Fig.6.

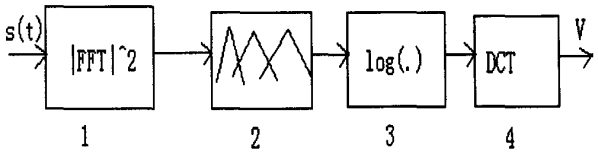

Fig.5 Traditional MFCC(Mel Frequency Cepectral Coefficients) for speech signal features extraction

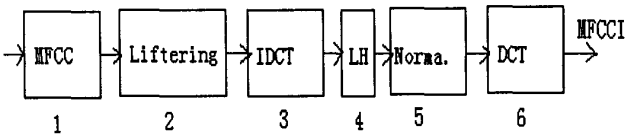

Fig.6 New MFCCI feature(Mel Frequency Cepctral Coefficients of lateral Inhibition)

In Fig.6 , module 1 is MFCC, which is the same as in Fig.5, this part can be changed into Auditory Frequency Cepstral Coefficients(AFCC)(it is discussed in the next part). In module 2, a liftering function is used:

$$L(k) = 1 + \frac{Q}{2}\sin(\frac{\pi k}{Q}), Q = 12, k = 1,2,...Q \tag{6}$$

For DCT transformation, it is a kind of linear orthogonal transformation, it can be used to compress the multi-dimensions, and de-correlate each dimension of the vector, suppose it is as the following form(for discrete form):

$$Y = CX \tag{7}$$

where X is the original spectrum vector (the log compression of energy output of the triangle filters), C is the cosine base function matrix, Y is the result vector of the DCT transformation. Then for IDCT:

$$X = C^{-1}Y \tag{8}$$

For cosine base functions matrix, it satisfies:

$$C^{-1} = C^{T} \qquad (9)$$

Module 4 is Lateral Inhibition(LH)operator. Module 5 is a normalization operator, it is used to scale the processed results. At last, a feature vector is gotten as MFCCI(Mel Frequency Cepestral Coefficients of lateral Inhibition).

The following is the comparison experiments. The database is 200 Chinese words(names of persons), sampled from telephone voices in SNR 25dB condition, sampling frequency is 8kHz, five times utterances for each words(total 200*5) for training, one time utterance for testing, white noise is added to test for the robustness, the SNR is defined as following formula:

$$SNR = 10 * \log 10(\frac{A^2}{no^2 + n^2}) \qquad (10)$$

where A^2 is the original energy of speech signal with noise signal, no^2 is the original noise energy, n^2 is the added noise energy. The SNR is chosen as 20dB, 15dB,10dB, 8dB, 5dB, 0dB.(Because there is a big decreasing in recognition rate between 10dB and 5dB, a 8dB is added between them). The experimental results are described in fig7 (for top 1 candidate)and fig 8(for top 5 candidates). The percentage is recognition rate.

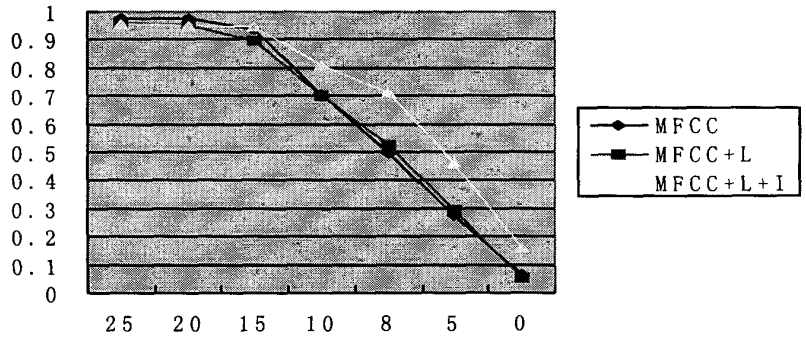

Fig.7 X-axis is SNR in dB scale, MFCC+L means MFCC with cepstral liftering processing, MFCC+L+I means MFCC with cepstral liftering and lateral inhibition processing(for Top1)

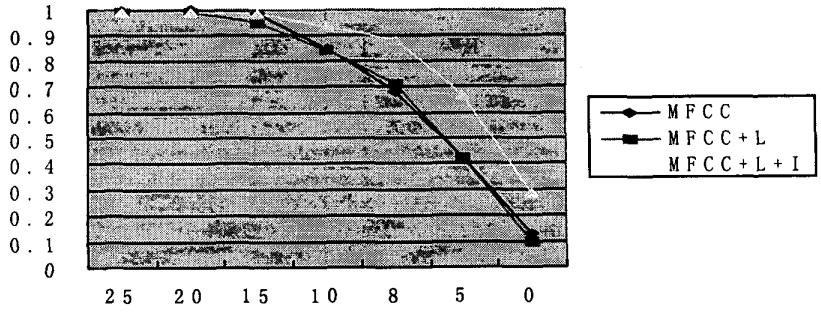

Fig.8 X-axis and Y-axis are the same as in Fig. 7 (for Top 5)

In fig 7 and fig 8, the MFCC+L means MFCC feature processed by lifter, MFCC+L+I means MFCC feature processed by lifter and inhibition. From Fig.7 and Fig.8, it is very clear that the robustness of MFCC with lateral inhibition mechanism is better than those of MFCC and MFCC only with cepstral liftering.

(2) **Lateral Inhibition in AFCC(Auditory Frequency Cepstral Coefficient).** Before calculating the lateral inhibition feature for auditory spectral representation, AFCC must be gotten, the AFCC can be gotten from the processing frame in Fig. 9:

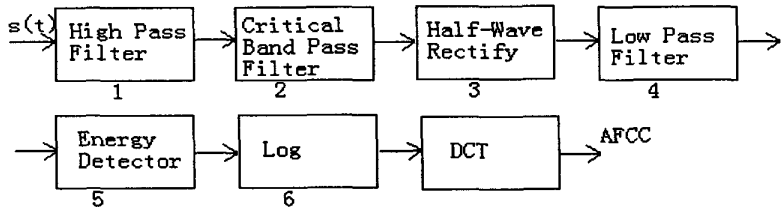

Fig. 9 Feature extraction for AFCC

From Fig.6, the MFCC is replaced by AFCC, then the processing frame is as in Fig.10, then the AFCCI can be gotten. The parameters of each modules in Fig.10 are the same as in Fig.6, the new feature is AFCCI, then robust recognition experiments are done as described in fig11 and fig 12, in the figures, AFCC+L+I means AFCC feature processed by lifter and inhibition. From Fig.11 and Fig.12, it shows, after processed by lateral inhibition model, the robustness of the recognition is improved.

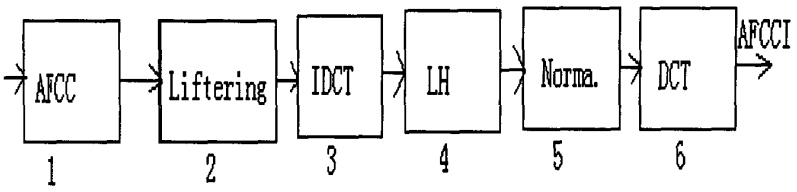

Fig.10 Feature extraction for AFCCI(Auditory Frequency Cepstral Coefficient of Lateral Inhibition)

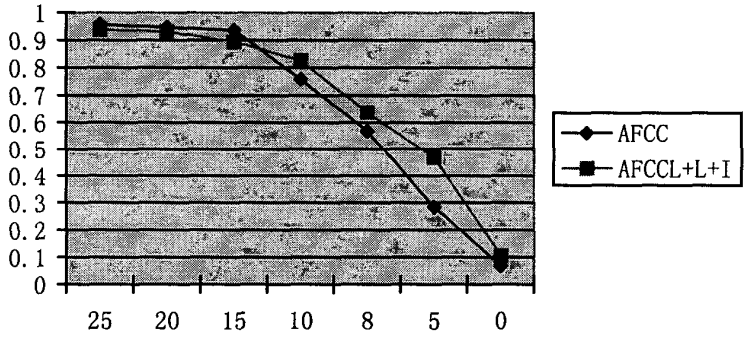

Fig.11 AFCC and AFCC+Liftering+Lateral Inhibition, the recognition rate (for top 1)X-axis is the SNR , Y-axis is the recognition rate

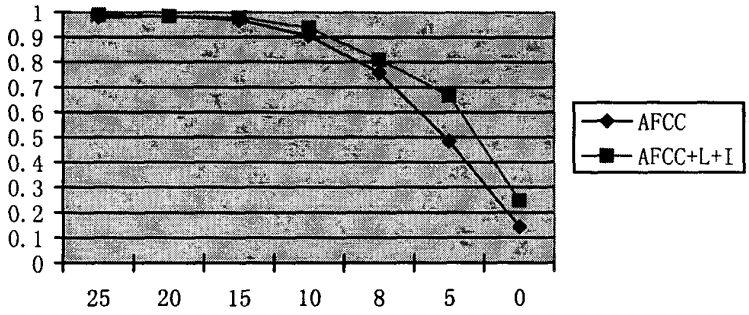

Fig.12 AFCC and AFCC+Liftering+Lateral Inhibition the recognition rate(for top 5), X-axis is the SNR, Y-axis is the recognition rate

5. DISCUSSION

In auditory information processing, many nonlinear mechanisms are used, such as masking effect, two tone suppression, lateral inhibition, etc. These nonlinear mechanisms can help auditory system to focus on prominent parts of the coming stimulation. But in traditional speech representation, few of these mechanisms is considered. Auditory system is the most efficient but complex model, new ideas can be borrowed from auditory psychological and physiological experiments. In this paper, the nonlinear mechanism of lateral inhibition is used in speech recognition, experiments show, it can improve the robustness of speech recognition system.

Reference

[1] Shihab A. Shamma, Speech processing in the auditory system I:The representation of speech sounds in the responses of the auditory nerve, J.Acoust.Soc.Am., 78(5), 1985, P1612-1621

[2] Shihab A. Shamma, Speech processing in the auditory system II: Lateral inhibition and central processing of speech evoked activity in the auditory nerve, J.Acoust.Soc.Am., 78(5), 1985 , P1622-1632

[3] Brain Strope, Abeer Alwan, A model of dynamic auditory perception and its application to robust word recognition, IEEE Trans. On speech and audio processing, Vol.5, No. 5, 1997, P451-464

[4] Bing-Hwang Juang,etc., On the use of band-pass liftering in speech recognition, IEEE Trans. On acoustics. Speech and signal processing, Vol. ASSP- 35,No. 7,1987, P947-953