Nonlinear processing in auditory system*

Lu Xugang *Chen Daowen*
National Laboratory of Pattern Recognition, the Institute of Automation,
Chinese Academy of Sciences, Beijing P.O.Box 2728, China
E-mail:lxg@nlpr.ia.ac.cn

Abstract

It is no doubt that our auditory system is a nonlinear processing system in speech signal perception. In this paper many aspects of nonlinear processing are discussed. First Nonlinear frequency resolution, that is nonlinear or non-uniform sampling theory in auditory system, second, nonlinear combination of different frequency components, such as two tone suppression and lateral inhibition. At last , we mainly focus on nonlinear intensity discrimination, which is important in forward masking effect, we give our model and experimental result.

1. INTRODUCTION

Traditional speech signal processing and recognition is based on the analysis of power spectrum, in some sense, it is a simplified form of our auditory processing mechanism. In speech recognition task, the features extracted based on this method can get good result, but in very crude conditions, such as noisy condition, or multi-speech sources condition, the recognition system's performance will decrease sharply. But our auditory system can have good robustness even in crude condition. So much more knowledge of auditory processing mechanisms should be added to traditional processing frames. Apparently, nonlinear processing mechanisms are very important and difficult part of these mechanisms. But, these nonlinear processing mechanisms are necessary for auditory system, also when some of the mechanisms are added to traditional processing frame, the performance can be improved, such as Mel frequency scale, log compression, etc. of course, it is not enough for a complex processing mechanisms in auditory system. In this paper, some nonlinear processing mechanisms are discussed which are ignored by traditional processing methods. In second part of this paper, nonlinear frequency resolution of auditory processing mechanism is discussed, in the third part, nonlinear frequency combination mechanism is discussed, in the fourth part, nonlinear adaptive mechanism and forward masking effect is discussed, also in this part, we give out our model and recognition experiment. At last a general discussion is given.

2. NONLINEAR FREQUENCY RESOLUTION

As we know from auditory psychological experiments, our ears give different frequency resolution at different frequency regions. In low frequency regions, the resolution is high, but at high frequency regions, the resolution decrease. In another word, auditory system use non-uniform sampling mechanism in frequency domain, so in engineering application, nonlinear mapping method is used often. Critical band theory is get from psychological experiments, it is got from masking effects, generally speaking, critical bandwidth is proportion to center frequency(above 500 hz). So in traditional basilar membrane filters design ,critical bandwidth is often used. So the spectrum getting from auditory model is regarded as critical band rate scale. This is a mapping from physical frequency scale to the psychoacoustic frequency scale, $\upsilon = Map(f)$,Where υ is the perception frequency domain, f is frequency domain. In band pass filters design, bandwidth and center frequency are in Bark scale, Mel scale, ERB scale. Suppose the original linear spectrum of speech signal is $s(t) \Leftrightarrow S(\omega)$, a nonlinear frequency mapping method is used as $\upsilon = f(\omega)$, that is $\omega = f^{-1}(\upsilon)$, then spectrum in υ domain as $\hat{S}(\upsilon) = S(f^{-1}(\upsilon))$,then in υ domain, uniform sampling is used. Apparently if $\upsilon = f(\omega) = a\omega$ that is linear mapping, then $\hat{S}(\upsilon) = S(\frac{1}{a}\upsilon)$, then it is just linear frequency warping.

* This work is supported by national sciences fund, No. 69635020

In traditional channel normalization, it is often used for speaker independent recognition. Suppose $\upsilon = \ln(\omega)$,that is nonlinear frequency warping, then if it is linear sampling in υ domain , it will be corresponding to nonlinear sample in ω domain. In auditory subband coding analysis, Bark domain or Mel domain are often used, then they are uniform sampled in Bark domain or in Mel domain, then in frequency ω domain will be approximate to exponential sampling. In fact many scale analysis method can give this kind of function. In auditory model, band-pass filters are often designed to simulate the function of basilar membrane, character frequency and critical frequency band are often important in design. With this constriction, nonlinear frequency resolution can be got. So traditional auditory model can have the function of channel normalization as frequency warping can do.(they are also regarded as constant Q filters[4][6]) Nowdays, the traditional recognition method is about HMM+MFCC, in features MFCC, it is a kind of nonlinear frequency domain which is enlightened by auditory perception. In fact this kind of frequency warping is related to auditory invariance.(Scale invariant transform), in traditional MFCC, with Mel cepestrum transformation, there is a scale invariant transformation which is derived from auditory perception experiments. Also in Bark domain, it acts as the same function. Many traditional normalization methods are derived from this kind of frequency mapping mechanism, thus in an transformation domain, the parameters distribution are unchanged for any dilation or compression. Channel normalization is a kind of scale transformation which can bring scale invariant to parameters.

In real recognition task, in one aspect, in higher frequency regions, the small ripples in spectral domain can be smoothed by this broad band pass filters, so in theory, the representation should be more robust than traditional representation based on linear band pass filters. But in another aspect, when the bandwidth of band pass filters are large, much more noise energy will mask the other frequency components. So it is difficult to judge which is better in real speech recognition task. Maybe auditory system can adapt it bandwidth for different detection task.

3. NONLINEAR FREQUENCY COMBINATION

Many nonlinear processing mechanisms in spatial domain are used in auditory system, such as two tone suppression, lateral inhibition, nonlinear frequency combination of different frequency components, of course , they have some relations with each other, such as two tone suppression[3] and lateral inhibition. Traditional features extraction in speech signal processing and recognition, such as MFCC, PLP BPFG etc., power spectrum is used, taking MFCC as example, the processing frame is as following:

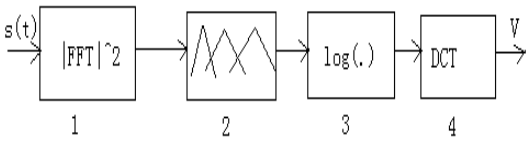

Fig.1 Traditional MFCC processing method

The power spectrum is calculated by FFT in module 1, a series of triangle filters are used to weight the power spectrum in module 1, after a log compression in module 3, a DCT is used to de-correlate the dimension in module 4, this processing method has it's drawback, first, linear input(dB) and output(dB) transformation(we will discuss this aspect in next part), second, in every triangle filter, linear combination of every frequency component is used to calculate the energy output of the triangle filter, it is an average processing method, because special characteristics of noisy and periodic information are not used in the processing, they are processed in the same status, such as following fig.2, for pure tones in left figure, the output of this Mel scale filter is $O = \sum_i W_i E(f_i)$, where W_i is weight of triangle filter, and $E(f_i)$ is

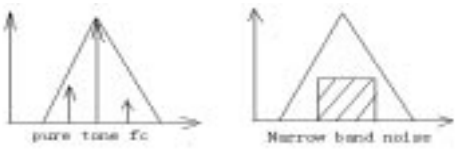

Fig.2 Linear combination of every frequency components for a triangle filter

the power spectrum of frequency f_i ,for right figure the output is $O = \sum_i W_i N(f_i)$,where $N(f_i)$ is the noise power spectrum, the weight is the same for the same frequency, so this processing method is an equalitarian method, no prior information of noise and speech is used. But in fact for noisy component, the weights will be different from those of pure tones. Also, for speech component, periodic property should be used, especially formants of speech signal. What we want to discuss in this paper is nonlinear combination of different frequency components, that is the suppression effect of multi-

frequency components.

Traditional model is based on pure tone stimulation tuning curves, but in fact, for multi-tone stimulation, two tone suppression is very obvious, that is to say, there are two kinds of stimulation, one is exciting stimulation, another is suppression stimulation, such as following tuning curve, The shadow area is suppression zone. So if the fiber is stimulated only by one single pure tone, there is exciting response only, if the fiber is stimulated by two pure tones simultaneously, the firing response will decrease because of the existence of the second stimulation.

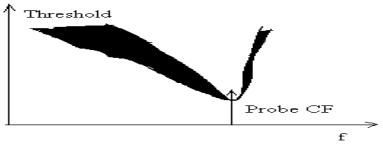

Fig.3 tuning curve of auditory neural fiber

Lateral inhibition has great relation with this two tone suppression. What the effect of this lateral inhibition? It is well known that lateral inhibition can enhance the edge of an image in computational vision, in auditory system, it can sharpen the changing part in temporal and spatial domain. So if this lateral inhibition mechanism is used to process the spectrum, it can sharpen the contrast between peaks and valleys.

In recognition task, the following lateral inhibition form(fig.4) is often chosen, it is obvious that low frequency regions can provide much more suppression than high frequency regions. Also, we define, the total area is zero, that is $\int_{Il+Ir+E} W(f)df = 0$,where Il is left suppression region, Ir is right suppression region, E is exciting region, $W(f)$ is the power spectrum at frequency f ,if the spectral structure is flat, the total output will be zero after processed by this lateral inhibition module, so this lateral inhibition can sharpen the contrast in spectral domain. In fact, not all the frequency component can give suppression, only those larger frequency components can provide suppression to other smaller frequency components, that is the suppression threshold.

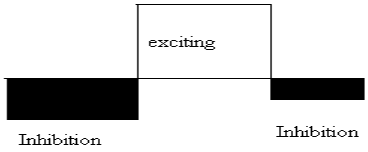

Fig.4 Lateral inhibition

4. NONLINEAR DYNAMIC INTENSITY PERCEPTION AND FORWARD MASKING EFFECT

It is proved that there is a very clear nonlinear intensity resolution in intensity discrimination [1][2][5]. Maybe the most famous psychological formula is Weber's law, that is JND theory $JND = \frac{\Delta I}{I}$,where I is the stimulation intensity, ΔI is the changes of intensity, it is supposed that JND is constant. In masking effect experiments , suppose masker intensity at level M_1 can give masking amount m_1, at masker level M_2, it can give masking amount m_2. If increase masker level, such as masker level is $M_1 + \Delta M$, then masking amount will increase to $m_1 + \Delta m_1$, for masker level $M_2 + \Delta M$, the masking amount will be $m_2 + \Delta m_2$, if $M_2 > M_1$,then $\Delta m_1 > \Delta m_2$, because with increasing masker intensity, there is nonlinear compression in intensity, the increased masker intensity will not so effective in masking. There are many nonlinear compression processing in auditory system, such as in basilar membrane, in auditory nerve fibres, etc. The possible biological path[8][9] is as following fig.5, BM means basilar membrane, IHC means inner hair cell, OHC means out hair cell, ANF means auditory neural fibers.

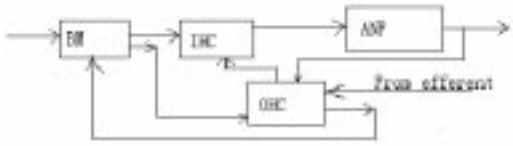

Fig.5 Nonlinear adaptation path in auditory system

Of course, it is difficult to model it in detail, a forward path is designed as following fig.6, the real output of each frequency channel is controlled by an adaptive gain.

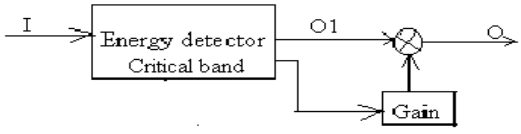

Fig.6 Forward adaptive gain controlling frame

Experiments can provide us with precise nonlinear I/O transformation, such as following curve,

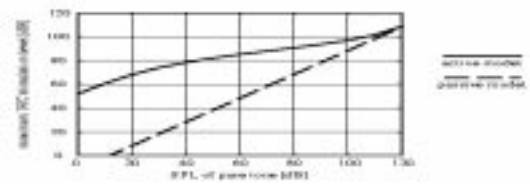

fig.7 Nonlinear I/O for IHC

Suppose, in fig.7, the relation of input and output is as following $O: I \to O(I)$

the slope of the curve as : $\frac{dO(I)}{dI} = k(I)$

Because in fig.7, dB scale is used, then gain can be got at different input intensity, that is , $O(I) = G(I) + I$ then the differential of this formula is $\frac{dO(I)}{dI} = 1 + \frac{dG(I)}{dI}$ so the derivative of gain to intensity can be got, $\frac{dG(I)}{dI} = k(I) - 1$

Now we define compression ratio in fig.7 as following, Suppose the input intensity is I ,corresponding output is $O(I)$,with a little increment ΔI , the corresponding output is $O(I + \Delta I)$, the compression ratio is $g(I) = \frac{O(I + \Delta I) - O(I)}{\Delta I} = k(I)$,that is to say, the compression ratio is just the slope of the I/O curve.

For the convenience of discussion, the simple gain curve is as following fig.8, Left curve is I/O curve, right is static gain curve. It is easy to show the relation between input and corresponding gain.

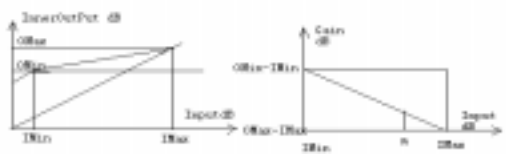

Fig. 8 Left: Simplified I/O curve, Right: Gain curve for different input intensity

The above discussion is about the static nonlinear compression analysis. In fact, the gain of cochlea can not change so rapidly as the input stimulation changes, so a time constant can be introduced in this adaptive processing. So for a input intensity, the output is as following curve, the dot line is the final output of the transformation, the solid line is the real adaptation processing.

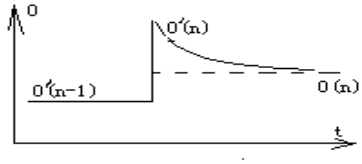

Fig.9 Output for adaptation

The adaptation processing is as following, suppose the input intensity is $I(f,n)$, f is center frequency of one channel , n is time（in MFCC，it is the energy output of triangle filters in dB scale , the static gain is $G(I(f,n))$, then the static output is

$$O(f,n) = I(f,n) + G(I(f,n)) \text{ (dB)}$$

The adaptation processing frame by frame is as following :

if $I(f,n) < I(f,n-1)$

$$G'(I(f,n)) = G(I(f,n)) + (G(I(f,n-1)) - G(I(f,n))) * a_h$$

where a_h is adaptation constant (from lower gain to higher gain)

if $I(f,n) > I(f,n-1)$

$$G'(I(f,n)) = G(I(f,n)) + (G(I(f,n-1)) - G(I(f,n))) * a_l$$

where a_l is adaptation constant (from higher gain to lower gain, it is different from the adaptation constant above).

Thus the real output should be:

$$O'(f,n) = I(f,n) + G'(I(f,n))$$

In real recognition task, an isolated speaker dependent word recognition system is used to test the effect of the nonlinear adaptation model, the original features are 12 order MFCC, and their first difference, and second difference, the total dimensions are 36. White noise is added to test database to test the robustness. The recognition resuts are as following table:

Table 1 (without nonlinear adaptation) Correct rate

SNR(dB)	Top 1	Top 5
25	0.910526	0.957895
20	0.900000	0.947368
15	0.821053	0.910526
10	0.531579	0.694737
5	0.194737	0.315789
0	0.057895	0.115789

Table 2(with nonlinear adaptation) Correct rate

SNR(dB)	Top 1	Top 5
25	0.921053	0.973684
20	0.900000	0.968421
15	0.836842	0.910526
10	0.568421	0.710526
5	0.194737	0.336842
0	0.057895	0.126316

The results show, there is only a little improvement in

robustness. The reason maybe lie in the differential of cepstrum, it can provide temporal information. But in theory, features processed by this nonlinear dynamic model can sharpen the changes in temporal domain, the representation should be much more robust.

5. DISCUSSION

Nonlinear processing mechanisms are very common for biological system, it is no doubt that these nonlinear mechanisms are very important for them to adapt to environment, as to auditory system, these nonlinear mechanisms can bring good mechanisms for auditory system to improve performance in noisy condition, but how to integrate these nonlinear mechanisms to traditional processing frames and to the engineering application is a very challenging topic.

6. REFERENCE

1. Brain Strope A.,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
2. Christopher J.Plack,Andrew J.O., Basilar-membrane nonlinearity and the growth of forward masking, J Acoust.Soc.Am.,103(3),1998.
3. C.Daniel Geisler, A.L.Nuttal, Two-tone suppression of basilar membrane vibrations in the base of the guinea pig cochlea using "low-side" suppressors. J.Acoust.Soc.Am. 102(1) 1997
4. Jams M.Kates ,A time-domain digital cochlear model, IEEE Trans. On Signal processing , Vol.39,No.12 1991.
5. Walt Jesteadt, Sid P.B.,James R.L., Forward masking as a function of frequency, masker level, and signal delay. J.Acoust.Soc.Am. 71(4) 1982
6. Hirahara T and Komakine T., A computational cochlear nonlinear processing model with adaptive Q circuits, ICASSP'89, pp.496-499.
7. M.P.Gorga and P.J. Abbas, AP measurements of short-term adaptation in normal and in acoustically traumatized ears, J.Acoust.Soc.Am., 70(5)1981
8. G.Kidd,Jr&L.L.Feth, Effects of masker duration in pure-tone forward masking,J.Acoust.Soc.Am. 72(5),1982.
9. E. Zwicker, Dependence of post-masking on masker duration and its relation to temporal effects in loudness, J.Acoust.Soc.Am.,75(1) 1984.
10. Stephanie Seneff, A joint synchrony/mean-rate model of auditory speech processing, Journal of Phonetics, 1998,Vol 16, P55-76