An Engineering Model of the Masking for the Noise-robust Speech Recognition

Ki-Young Park[a,b] and Soo-Young Lee[a,b,c]

[a] *Brain Science Research Center,* [b] *Department of Electrical Engineering, and* [c] *Department of Bio-Systems, Korea Advanced Institute of Science and Technology, Daejeon, Korea*

Abstract

The masking effect of human hearing is modeled by lateral and unilateral inhibition, and tested for isolated word recognition tasks. Frequency masking suppresses unwanted signals close to the dominant signal of interest in frequency domain, and the weak signals following dominant ones in the time are suppressed by temporal forward masking. The masking effect filters out unimportant signals, which may improve the performance of speech recognition systems. With the parameters derived from the psychological observations, proposed model shows good analogy to psychoacoustic masking effects as well as superior recognition performance.

Key words: Masking; Lateral Inhibition; Automatic Speech Recognition

1 Introduction

Although tremendous attempts have been made to make machines which recognizes the human speeches, it is still very difficult for real world environments with the background noises. The modeling of human auditory system is one of the successful approaches which have been proposed to solve the problem. Critical-band filterbank and mel-scale frequency sampling are the most widely used properties from the psychoacoustics.

One of the useful aspects which had been seldom utilized is the "masking" effect of human hearing system. Masking has been investigated for ages and used to quantify the frequency selectivity of the auditory system. However, there have been few approaches which utilize the masking for the recognition tasks. The nature of the masking, that the spectral components of high intensity level suppress the adjacent spectral components, helps the recognition performance. The low intensity level signals are usually unimportant for speech recognition and may be noises. to be suppressed. In this paper, the time-frequency masking is modeled with lateral inhibition and incorporated

into the current auditory model, mel-frequency cepstral coefficients (MFCC) model which is the most widely used speech features, and tested on the task of isolated word recognition. The proposed algorithm does not require extensive computation, and results in much better recognition performance, especially in noisy environments.

Masking has been defined as a process in which the audible threshold for one sound is raised by the presence of another (masking) sound. Frequency masking means that signals are masked by the masking sound occurring at the same time. With temporal masking, signals can also be masked by the sound preceding it, called forward masking, or even by the sound following it, called backward masking. Frequency masking helps to discriminate signals from the other by enhancing the spectral resolution. Also by suppressing the adjacent signal in spectral domain, it reinforces the signal of critical interest so that the unimportant signals are filtered out. Forward masking, a short term adaptation process of the auditory system, help the discrimination capability between signals by emphasizing time-dependent variations.

2 Frequency masking with lateral inhibition model

The essence of the masking is to reinforce the dominant signal components and to suppress the adjacent components. To implement this concept into the current auditory system, a lateral inhibition is introduced with a simple Mexican-hat convolutional filter as shown in Fig. 1(a). The sharp peak at the center reinforces the very close stimuli and the negative values at neighborhoods inhibit the stimuli in the range.

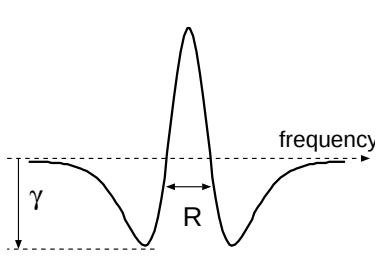
To apply the inhibition filtering in spectral domain, the blocked speech signals are first transformed into frequency information using discrete Fourier transforms, and the inhibition is applied between adjacent frequency components. It can be easily incorporated into the popular MFCC speech features as shown in Fig. 1(c).

The Mexican-hat filter is modeled by the difference of Gaussian,

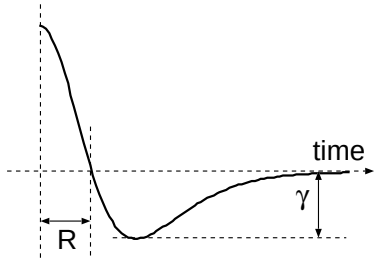
$$w(n) = A\exp(-n^2/\sigma_1^2) - B\exp(-n^2/\sigma_2^2),\ n = -L, \ldots, 0, \ldots, L \tag{1}$$

where $2L + 1$ is the length of the masking filter. The most crucial problem is to determine the shape of the filter, i.e., the parameters of A, B, σ_1 and σ_2. By applying two constraints, i.e., zero mean and the maximum value of one, one can shape the filter for the remaining two physical parameters of the excitatory range R and inhibition gain γ. The masked spectral magnitude is computed as

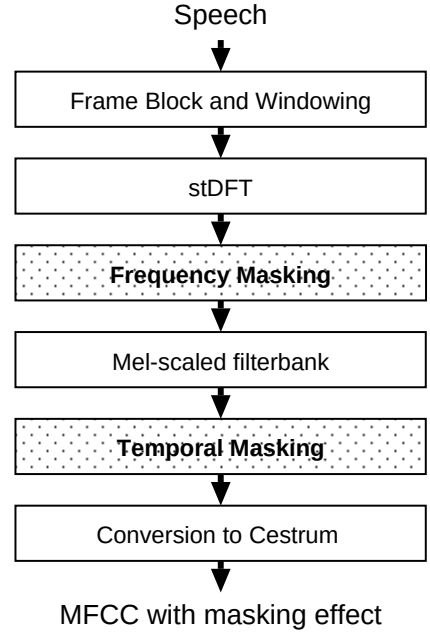
$$m(k) = \sum_{n=-L}^{L} w(k-n)S(k), k = 0, 1, \cdots, N-1 \tag{2}$$



(a) Frequency masking



(b) Temporal masking



(c) MFCC model

Fig. 1. The filters for the frequency(a) and temporal(b) masking and the MFCC model with the proposed masking mechanism(c)

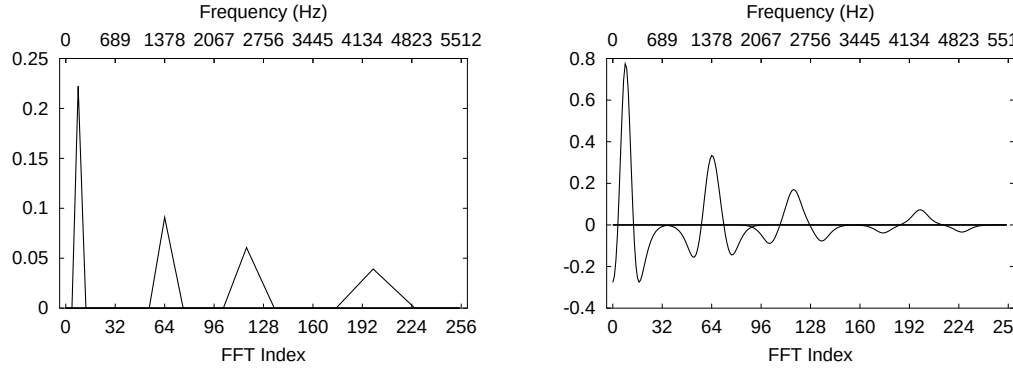


Fig. 2. Mel-scale filterbank with masking effect (right) compared to the original mel-scale triangular filterbank (left).

where $S(k)$ is the magnitude of the FFT output of the blocked speech signal and N is the number of FFT points. The mel-scale filterbank in the MFCC is originated from the critical band analysis. Since the implementation of the filterbank can be regarded as the weighted sum of the frequency components, the two process can be combined into one as shown in Fig. 2. The figure shows that the Mexican-hat filter in frequency domain results in the modification of the mel-scaled filterbank. The output value of each critical band are inhibited by those of adjacent bands, and the inhibition range increase nonlinearly as the center frequency increases. These are consistent with psychoacoustic

observations.

3 Temporal Masking

The short-term adaptation and the temporal integration [5] are the possible mechanisms of the temporal masking. Many researches have modeled the temporal masking as the temporal integration of the response of the auditory nerve [1,6]. If we assume that the temporal masking is due to the temporal integration of the response of the auditory nerves, temporal masking can be modeled as

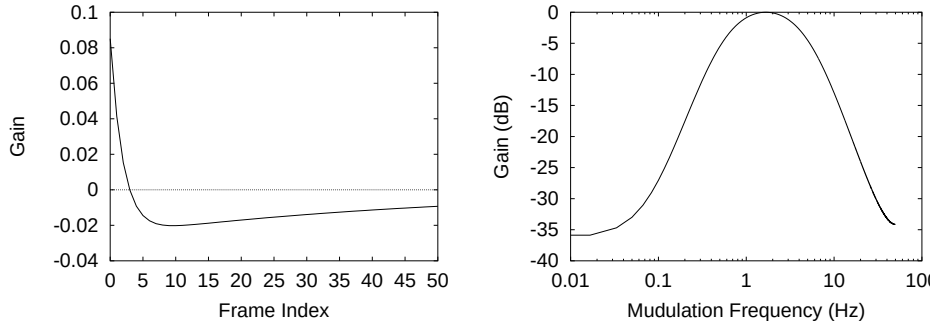
$$y(n) = x(n) + A\sum_{k=1}^{\infty}\alpha^{-k}x(n-k) - B\sum_{k=1}^{\infty}\beta^{-k}x_(n-k) \tag{3}$$

where, $x(n)$ is the output signal before temporal masking, $y(n)$ the output of the temporal masking, α and β are time constants of integration. The second term in the right side of Eq.(3) denotes the slow response of the neurons, in other words, the responses of the neurons could not catch up the changes of the stimuli, and the previous inputs are accumulated to affect the current output. The third term indicates the masking integration. The current response of the neurons are suppressed by the preceding signal of high intensity.

The time constants of two integration terms, α and β, represent the extent to which the previous neural signals affects the current signals. It is known that α is much greater than β. Typically α is a few tens of milliseconds and β a few hundreds of milliseconds. The gain factor A and B imply the amount of accumulation and masking respectively. Esq.(3) is represented in z domain as

$$H(z) = \frac{1 - [(1-A)\alpha + (1+B)\beta]z^{-1} + (1-A+B)\alpha\beta z^{-2}}{1 - (\alpha+\beta)z^{-1} + \alpha\beta z^{-2}} \tag{4}$$

Again for the preliminary experiments α is set to 0.72 which corresponds to 30ms, and β is set to 0.97 corresponding to 300ms. A and B are set to comply with the constraints of zero-mean and the maximum value of one as in the case of frequency masking. Fig. 3(a) and Fig. 3(b) show the impulse response of the time integration filter of Eq.(4) with the parameters above. From the frequency response of the model, it is shown that the proposed model with the parameters given actually do the role of bandpass filtering which cuts off the signal components which vary too fast or too slow. Then the model can be regarded as the generalized version of RASTA model proposed by Hermansky and Morgan, where they used the bandpass filtering method in feature domain to enhance the recognition performance [3]. They also showed that the temporal filtering of the sequence of feature vectors could give the masking effect [2]. The temporal masking filter is applied at the logarithm filterbank output. Fig. 1(c) shows the modified MFCC model including both



(a) Impulse response. (b) Frequency response.

Fig. 3. Characteristics of the filter used for the integration model of temporal masking.

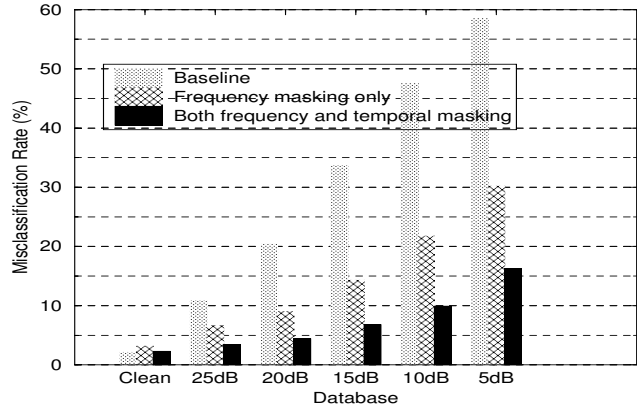


Fig. 4. False Recognition rates of the baseline system without any masking, with the frequency masking only and both the temporal and frequency masking.

the temporal masking stage (upper shaded box) and the frequency masking (lower shaded box).

4 Experimental Results

The temporal masking using the integration model in the feature domain as well as the frequency masking is applied to the isolated word recognition task. For the simulation, 50 Korean words spoken by 13 men three times each were tested by nearest neighbor classifier for simplicity. Fig.4 compares the false recognition rates of the baseline without any masking, with the frequency masking, and with both the frequency and the temporal masking. Frequency masking reduces the misclassification rate greatly and the temporal masking reduces the error rate even further. We found the optimal shape for the recognition task with the exhaustive calculation by varying two control variables, R and γ. The simulation showed that the value of excitatory range R between

60 and 100Hz and the value of inhibition gain γ between -0.4 and -0.2 gave much better performance than the baseline system for frequency masking. For temporal masking, the values of R between 40 and 100 ms and γ between -0.05 and -0.15 gave the consistent improvements on the recognition performance. We also found that parameters deviated from the optimal values gave much better performance than the baseline system. Therefore broad ranges of parameter values could be used.

The values of parameters were found by the brute-force search and after many trials of simulations, we found that parameters deviated from the optimal values also gave much better performance than the baseline system which meant broad ranges of values for each parameters could be for practical use.

5 Conclusions and discussions

In this paper, the psychoacoustical phenomena of masking is modeled by simple convolutional filtering in both spectral and time domain. The concepts of lateral inhibition in spectral domain and of unilateral inhibition in time domain model the frequency masking and temporal masking, respectively. Proposed model results in efficient features for speech recognition and provides much better performance than popular MFCC features especially in noisy environments.

Acknowledgment

This research was supported by Korean Ministry of Science and Technology as Brain Neuroinformatics Research Program.

References

[1] Torsten Dau and Dirk Püschel. A quantitative model of the "effective" signal processing in the auditory system. I. model structure. *The Journal of Acoustical Society of America*, 99(6):3615–3631, 1996.

[2] Hynek Hermansky. Shoould recognizers have ears? *Speech Communications*, 25(1):3–27, 1998.

[3] Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transaction on Speech and Audio Processing*, 2(4):587–589, 1994.

[4] Brian C. J. Moore. *An introduction to the psychology of Hearing.* Academic Press, third edition, 1997.

[5] Andrew J. Oxenham. Forward masking: Adaptation or integration? *The Journal of Acoustical Society of America*, 109(2):732–741, February 2001.

[6] Brian Strope and Abeer Alwan. A model of dynamic auditory perception and its application to rebust word recognition. *IEEE Transactions on Speech and Audio Processing*, 5(5):451–464, September 1997.