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

Ki-Young Park and Soo-Young Lee

Abstract

The masking effect of human hearing is modeled by lateral and unilateral inhibition model, and tested by the isolated word recognition task. Simultaneous masking suppresses the unwanted signal close to the dominant signal of interest, and the weak signals following dominant ones are suppressed by the temporal forward masking. These properties are to be utilized for the recognition systems since masking cuts off the uninterested signals which might make the recognition performances worse. With the parameters derived from the psychological observations, proposed model shows good analogies to them as well as the superior recognition performance.

Keywords: masking, lateral inhibition, automatic speech recognition.

1 Introduction

Although tremendous attempts have been made to make a machine which recognizes the human speeches, it still is very difficult for real world environments where the background noises exist. The modeling of human auditory system is one of the successful approaches which have been proposed to solve the problem. For example, nonuniform frequency scale such as Mel-scale which is derived from psychoacoustic observation is one of the most widely used feature in many automatic speech recognition (ASR) systems. Critical-band spectral resolution and frequency dependency of the relation between intensity and loudness are used in perceptual linear predictive (PLP) analysis [1].

One of the useful aspects which have been rarely utilized is the “masking” effect of the human hearing system. Masking has been researched for ages and used to quantify 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 component of high level suppress the adjacent spectral component of low level helps the recognition performance since the low level signals are usually noise which should be suppressed.

In this work, the masking from the psychoacoustics was implemented into the auditory model, Mel frequency cepstral coefficients (MFCC) model which is one of the most widely used feautres, and tested on the task of isolated word recognition. Even with the very simple method which does not require much amount of computation, the proposed model incorporated into the current model gave superior performance especially in noisy environments.

2 Modeling of the masking

Masking has been defined as the process or amount by which the threshold of audibility for one sound is raised by the presence of another (masking) sound [4].

As well as a signal is masked by the sound occurring at the same time, called simultaneous masking, a signal can be masked by the sound preceding it, called forward masking, or even by the sound following it, called backward masking.

Simultaneous masking helps to discriminate signals from the other to enhance the spectral resolution of the human hearing. Also by suppressing the adjacent signal in spectral domain, it enhances the signal of interest so that the unwanted signals are cut off from being fed into the recognizer.

Forward masking, which sometimes regarded as the short term adaptation of the auditory system, helps the discrimination capability between signals by cuting off the signals which changes too slowly or too fast.

2.1 Simultaneous masking with lateral inhibition model

The essence of the masking is to enhance a dominant signal component and to suppress the noise components adjacent to the signal components. To implement this concept into the current auditory

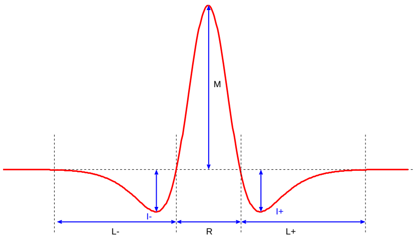


Figure 1: General shape of the filter for spectral masking.

model, the lateral inhibition by the simple convolutional filter of the shape shown in Fig. 1 were introduced. 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 transformed into frequency information first using the techniques such as Fourier transforms, then filtering can occur between the adjacent frequency components. Since the discrete Fourier transform of the blocked speech signal is explicitly computed in the processes involved with the extraction of MFCC model, this method can be easily merged into the MFCC model.

A more crucial problem is to determine the shape of the filter, i.e. the numerous parameters including the reinforcement and the inhibition gain, M, the reinforcement ranges, $R+$ and $R-$, and the inhibition ranges, $L+$ and $L-$. The filter is not necessarily symmetric, and psychoacoustic experiments state the filter is necessarily asymmetric as it can be seen in the masking pattern [5]. The range of masking is known to be the function of the signal frequency, and it usually represented by the relative masker frequency:

$$\frac{f_m - f_s}{f_s} \qquad (1)$$

where f_m is the frequency of the masker signal and f_s of the masked signal. The range of the relative masker frequency of interest extends from -0.3 to 0.3 which correspond to ±30Hz for 100 Hz signal and ±1500 Hz for 5kHz signal.

In the foregoing experiments, the coefficients of filters are chosen arbitrarily and only the general form of the filter were considered with the guide line above. In the engineering sense, the exact form of the filter should not have much influence on the recognition performance.

One more thing which should be considered is that the masking pattern is the function of the frequency and the masker stimulus in psychoacoustics, i.e., the shape of the filter is different if the frequency and/or the amplitude of the stimuli is different. Moreover, the masking pattern is asymmetric to the masker signal frequency. At the current stage, these properties were not implemented into the model explicitly.

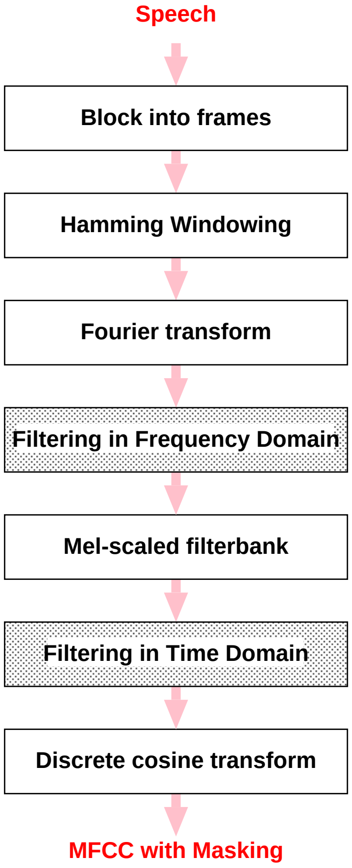


Figure 2: Proposed model of MFCC with masking.

2.2 Lateral inhibition with MFCC model

In the MFCC model, blocked speech signals are first converted into spectral domain using fast Fourier transform (FFT). Fig. 2 shows the proposed model of MFCC with lateral inhibition which has just one more step of frequency domain filtering (upper shaded box). The filter coefficients come from the difference of two Gaussian functions.

$$w(n) = \alpha \left[\exp(\frac{n^2}{\sigma_1^2}) - \gamma \exp(\frac{n^2}{\sigma_2^2}) \right], \qquad (2)$$
$$n = -L, \cdots, -1, 0, 1, L$$

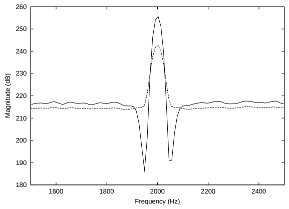


Figure 3: The effect of the lateral inhibition on the pure tone mixed with the white noise.

where $2L+1$ is the length of the masking filter. The gain of the filter is adjusted so that the sum of filter coefficients equals to one, and γ determines the inhibition gain, I. The most significant factor is the length of the filter which includes the reinforce range and the inhibition range. They are determined by the variance of the Gaussian, σ_1 and σ_2.

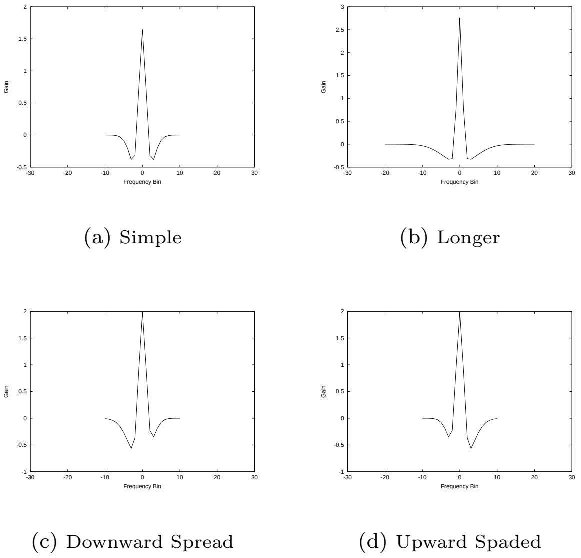
Then, the masked spectral magnitude is computed as

$$m(k) = \sum_{n=-L}^{L} w(k)S(k+n), \qquad k = 0, 1, \cdots, N-1 \tag{3}$$

where $S(k)$ is the magnitude of the FFT output of the blocked speech signal and N is the number of FFT points. The filtered FFTed values are then integrated using the critical bandpass filters like the conventional models. Simple inhibition filter of Fig. 4(a) was used for the preliminary experiments, where σ_1 and σ_2 are set for the masking range extends to ±100 Hz in Eq.(2).

Fig 3 shows the effect of filtering. The input stimuli was the pure tone signal mixed with high level of white noise. The magnitude of the FFT output is indicated as a dotted line. By filtering this in spectral domain, we get the desired result shown as a solid line.

With the various filters of Fig. 4, the isolated word recognition task was performed. 50 Korean words spoken twice by the 9 men are used to train multi-layer perceptron (MLP), and the same words spoken three times by the another 16 men are used to test the recognition performance. Asymmetric filters which reflect the asymmetry of the psychoacoustic masking pattern were also considered. The recognition results with the various filters of different shapes in Fig. 4 were shown in Fig. 5. The performance with various filters were not much different except with the longer filter which extends up to ±200 that might be too wide. This means the masking is not so sensitive to the exact shape of the filter.



(a) Simple

(b) Longer

(c) Downward Spread

(d) Upward Spaded

Figure 4: Simple filter to simulate lateral inhibition.

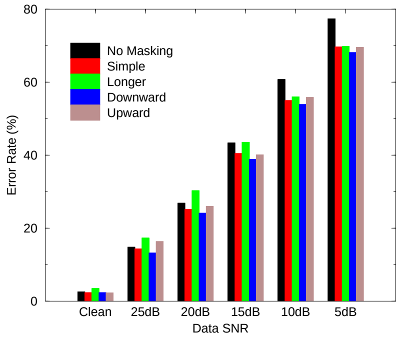


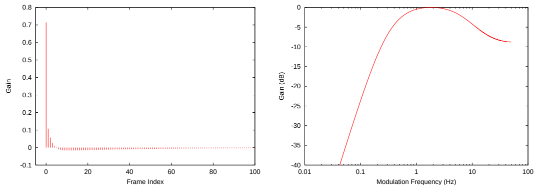
Figure 5: Word recognition error rates with the filters shown in Fig. 4.

2.3 Temporal Masking Effect

The short-term adaptation and the temporal integration [6] 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 [7, 8, 9].

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 following in general.

$$y(n) = x(n) + A\sum_{k=1}^{\infty} \alpha^{-k} x(n-k) - B\sum_{k=1}^{\infty} \beta^{-k} x_(n-k) \tag{4}$$



(a) Impulse response.

(b) Frequency response.

Figure 6: Characteristics of the filter used for the integration model of temporal masking.

where, $x(n)$ is the output signal without temporal masking, $y(n)$ the output of the temporal integration, α and β are time constants of integration. The first term in the right side of Eq.(4) covers the slow response of the neurons, in other words, the responses of the neurons could not catch up the changes of the stimuli so the previous inputs are accumulated to affect the current output. The second term indicates the masking integration. The current response of the neurons are suppressed by the preceding signal of high intensity.

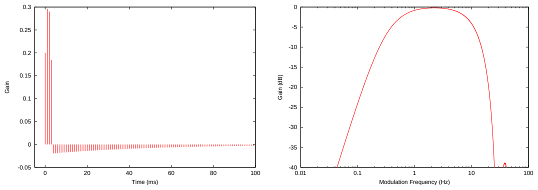
The time constant of two integration terms, α and β represent the extent to which the previous term affects the current output. It is known that α is much greater than β, typically α is a few tens of milliseconds and β a few hundreds of milliseconds. The gain factor A and B imply the amount of accumulation and masking respectively. Eq.(4) is represented in z domain as following.

$$\begin{aligned} H(z) &= \frac{1 - \mathcal{X} z^{-1} + \mathcal{Y} \alpha\beta z^{-1}}{1 - (\alpha+\beta) z^{-1} + \alpha\beta z^{-2}} \qquad (5) \\ \mathcal{X} &= (1-A)\alpha + (1+B)\beta \\ \mathcal{Y} &= (1-A+B)\alpha\beta \end{aligned}$$

Again for the preliminary experiments α is set to 0.6 which corresponds to 20ms, and β is set to 0.98 corresponding to 200ms. A and B are set to 0.3 and 0.03 respectively which are determined arbitrarily. With the parameters above, Eq.(5) yields

$$H(z) = 0.72 \frac{1 - 1.43 z^{-1} + 0.43 z^{-2}}{1 - 1.58 z^{-1} + 0.58 z^{-2}}. \qquad (6)$$

Fig. 6(a) and Fig. 6(b) show the impulse response of the time integration filter of Eq.(6).



(a) Impulse response.

(b) Frequency response.

Figure 7: Characteristics of the filter in RASTA model

Time integration model can be explained in a different way. Hermansky and Morgan used the filtering method in feature domain, called RASTA [10]. They also showed that the temporal filtering of the sequence of feature vectors could give the masking effect [11]. They used the infinite impulse response (IIR) filter with the transfer function

$$H(z) = 0.1 \frac{2 + z^{-}1 - z^{-3} - 2z^{-4}}{1 - 0.98 z^{-}1}. \qquad (7)$$

Fig. 7(a) and Fig. 7(b) show the impulse response of the filter in frequency and time domain respectively. Both filters represent the band-pass filters with similar characteristics. The biggest difference between filter of RASTA model and that of time-integration model is that while RASTA filter does the band-pass filtering in feature domain, the proposed time-integrating model does not explicitly imply the band-pass filtering, and it can be any kind of filtering.

As for the filtering domain, we can use the MFCC model as the input vector again. The output of MFCC, however, is the cepstrum which is cosine transformation of the log of the filterbank output. They have both positive and negative values and are not proportional to the intensity of the stimuli. So, the outputs of the MFCC model are not suitable to apply temporal masking. Instead the outputs of the filterbank are good for the filtering since they correspond to the responses of the cells in the human auditory system which are nonlinearly proportional to the input stimuli. Fig. 2 shows the modified MFCC model including both the temporal masking stage (upper shaded box) and the simultaneous masking (lower shaded box).

To quantify the time integration model of masking on MFCC model, the experiments by Jesteadt *et. al.* were reproduced by simulation with the 2 kHz pure tone signal [12]. Fig. 8 shows the input signal used for the simulation. Each shaded box indicates the pure tone wave of the same frequency. The detection of probe of the small power is influenced by preceding masker signal of the high power, and the amount of masking is the function of the time delay between probe and masker signals, Δt,

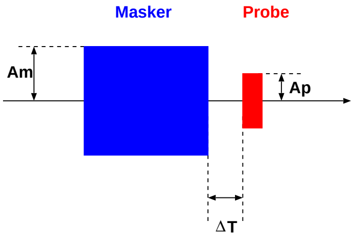
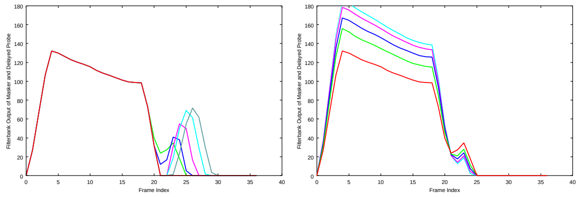


Figure 8: Input stimuli for the forward masking.



(a) Varying Delay

(b) Varying Masker Intensity

Figure 9: Typical outputs of the temporal masking model.

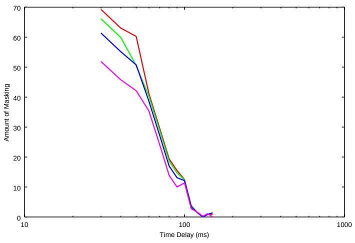


Figure 10: The amount of masking as the function of delay between masker and probe signal.

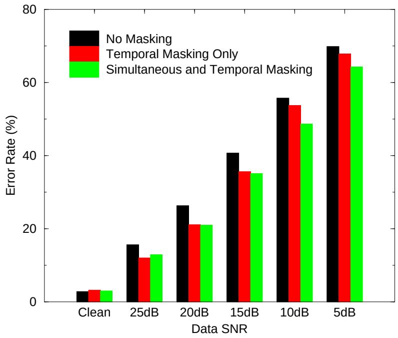


Figure 11: Recognition error rate using temporal masking and the simultaneous masking.

and the ratio between the levels of two signals.

Fig. 9 shows the typical result of the temporal masking using the Eq. 6 as the temporal filter. Fig. 9(a) shows the output as the delay between the masker and probe is changed, and Fig 9(b) varying the intensity of the masker signal. It is clearly shown that as the delay increases, the amount of masking also decreases, and as the intensity of the masker increases, the amount of masking decreases. The gradual decrease during the masker signal is due to the cut-off of the low frequency components of the filter used.

The amount of the masking is the function of the delay between masker and probe, Δt. To quantify the degree to which the delay influence on the detection of the probe signal, the amount of the masking was defined as the difference between the peak values of the outputs with and without a preceding masker signal. Usually this is linearly decreasing function of time delay [12]. Fig. 10 shows the amount of masking as the function of log of delay Δt. It shows the linear dependency between the amount of masking and the logarithm of the delay.

The temporal masking using the integration model in the feature domain is applied to the isolated word recognition task. The simulation system and database were identical to the experiments in simultaneous masking. Fig.11 shows the recognition error rates with the only temporal masking and with both the simultaneous and the temporal masking. As for the simultaneous masking, simple filter of Fig. 4 was used. The temporal masking by time integration model consistently reduces the recognition error rate especially in noisy environments.

3 Conclusions and discussions

In this work, the psychoacoustical observation of masking is introduced and modeled by simple filtering in both spectral and time domain. Masking, which is a quite subjective measure of the amount by which the significant signal of high level suppresses the adjacent signals.

The concepts of lateral inhibition in spectral domain and of unilateral inhibition in time domain model the simultaneous masking and temporal masking respectively. Proposed model was tested by the task recognizing the isolated words. The proposed model gave the superior performance to the current auditory models consistently especially in noisy environments.

The most crucial issue in this work would be the design of the filter which modeled the lateral inhibition. The filter used in this work was designed quite arbitrarily and had no theoretical background. We just assumed that only the overall shape of the filter would be important and not the exact shape. However, this assumption was not verified theoretically and only some experimental results were presented. To design the filter properly, we should rely on the psychoacoustics, where human hearing is measured quantitatively.

References

[1] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of Acoustical Society of America*, 87(4):1738–1752, April 1990.

[2] Oded Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):115–132, January 1994.

[3] Doh-Suk Kim, Soo-Young Lee, and Rhee M. Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on Speech and Audio Processing*, 7(1), January 1999.

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

[5] James P. Egan and Harold W. Hake. On the masking pattern of a simple auditory stimulus. *The Journal of Acoustical Society of America*, 22(5):622–630, September 1950.

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

[7] 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.

[8] 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.

[9] Jürgen Tchorz and Birger Kollmeier. A model of auditory perception as front end for automatic speech recognition. *The Journal of Acoustical Society of America*, 106(4):2040–2050, October 1999.

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

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

[12] Walt Jesteadt, Sid P. Bacon, and James R. Lehman. Forward masking as a function of frequency, masker level, and signal delay. *The Journal of Acoustical Society of America*, 71(4):950–962, April 1982.