INCORPORATION OF TEMPORAL MASKING EFFECTS INTO BARK SPECTRAL DISTORTION MEASURE

B. Novorita

Motorola
1301 East Algonquin Road
Schaumburg, IL 60196, USA

ABSTRACT

The objective of this paper is to extend a promising objective speech distortion measurement method, the Bark Spectral Distance (BSD) measure, with the auditory concepts of forward and backward temporal masking to improve its measurement accuracy. The results of this investigation show that automatic BSD-based speech quality ratings may be made to correlate better with existing MOS ratings by removing perceptually irrelevant areas of speech from the distance measure. The correlation between the objective BSD measure to the subjective MOS measure increases from 0. 91 to 0. 98. The best results were found with a window duration of 128 samples, use of exponential-slope filter characteristics for both forward and backward masking effects, forward masking delays up to 100 msec, and a backward masking time advance of 40 msec.

1. INTRODUCTION

The objective of this paper[1] is to extend the baseline work published by Wang et al. [1] which defined BSD speech quality measure. The BSD measure models the hydro-mechanical response of the human cochlea. The cochlea then transforms the hydromechanical behavior into neural activity which drives higher-level intellectual processes. This core function of acoustical to neural transduction provide a common building block for human hearing perception which the BSD exploits. Furthermore; other known supplementary acoustical properties of hearing perception exist, such as temporal masking, but to date have not been incorporated into objective speech quality evaluation models.

Improvement in the evaluation accuracy of BSD objective measure serves as the specific goal of this paper. It is hypothesized that improvement in distortion measure accuracy can be obtained by incorporating the temporal effects of auditory masking into the BSD model. The required temporal masking models are based upon established auditory masking principles [2], [3], [4], [5], [6]. The principles state that localized high-energy regions of speech "mask" or suppress lower-intensity time-contiguous regions of speech from the perception process. The psychoacoustic masking of low energy areas by neighboring (time sequence of a single critical band) high energy areas has been proven to impact human perception and discrimination [2]. These temporal masking effects have a direct influence on human perception and subsequently impact the correlation to subjective mean opinion score (MOS) ratings. Therefore temporal masking may be factored into any objective distortion measurement method as a means to provide a more accurate model of the hearing and perception process.

1.1 Bark Spectral Distortion Measure

The measure computes the spectral distance between a processed version of a source and a processed version of an output. The Euclidean distance between the input and output provides a measure of fidelity of how well the output matches the input speech. To determine the significance of the measure, it is correlated with a known subjective assessment of speech, MOS. The degree of correlation between the MOS data set and the BSD measure data set indicates a sense of tracking accuracy between the two data sets. Full details of BSD are found in Wang et al. [1].

1.2 Bark Spectral Distortion Deficiencies

The BSD measure is based on a perceptual model that incorporates the effects of simultaneous masking or frequency smearing within a given short-time frame. However the model does not take into account the inter-frame masking effects that impact human perceptual discrimination. Hence an opportunity exists to incorporate these inter-frame temporal masking concepts into the existing BSD model to improve its performance. A question of applicability surrounds the use of existing temporal masking models. Do masking threshold models derived from single-tone and narrow-band noise source stimuli accurately model the temporal auditory masking responses to complex-tone signals like speech? This paper assumes the masking models are accurate.

1.3 BSD Measure Improvement

The temporal masking properties provides the basis for the inter-frame masking concept. Figure 1 illustrates how these masking principles are incorporated into the improved BSD algorithm. The strategy is to remove those discrete Bark spectral data points from the distortion computation that are masked by higher intensity locations in that given Bark spectral track. The masked points are placed into a corresponding masking intensity matrix for all bark spectral data points and short-time frames. The error contributions associated with the masked spectral data points are removed from the distortion computation, thus the masked BSD measure is computed only from perceptual errors of the utterance; imperceptible error contributions are ignored.

[1] This paper is based on a Master's Thesis completed at the University of Illinois - Chicago.

2. TEMPORAL MASKING EFFECTS

Specifically of interest is temporal masking along with the associated models which are derived from single tone and narrowband noise stimuli.[2] [4]. The behavior of these forward and backward masking models mimics attributes of the human auditory system when it comes to perception.

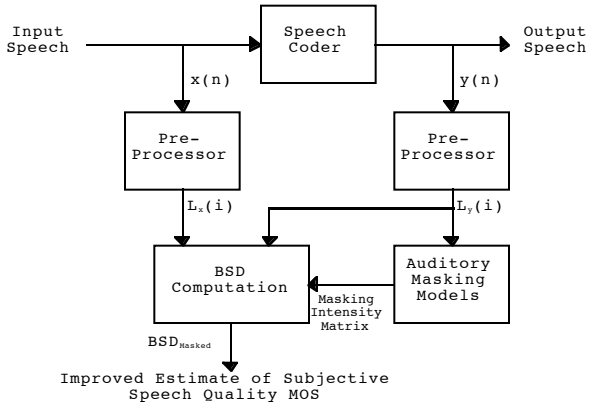

Figure 1. Block diagram of the improved BSD measurement system. Acoustic models are used to dynamically enhance the BSD measure

2.1 Forward Masking Model

Forward masking is the temporal hearing phenomenon that occurs when high energy stimuli (masking signals) precedes, in time, and suppresses later arriving and lower energy stimuli (masked signals) from perception [7], [2], [8]. Forward masking can be modeled as the decay of residual energy in the each of the Bark critical band filters which, for this bandwidth 0-4 kHz, there are sixteen (16). Aural events which have stimuli levels below the decay residual values are effectively masked from perception as illustrated in Figure 2. The curve converges at the quiet threshold for Δt greater than 200 msec.

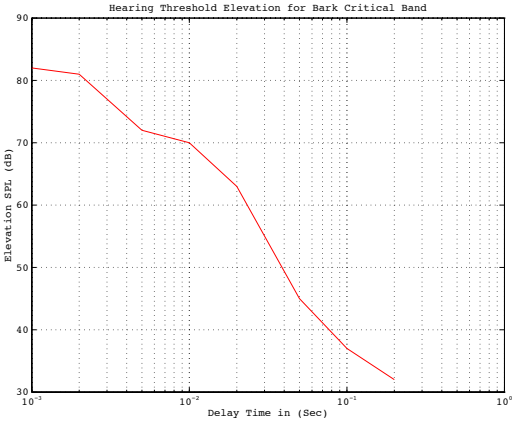

Figure 2. Post-event elevation of hearing threshold due to aural stimulation event; forward masking.

2.2 Backward Masking Model

Backward masking is the temporal hearing phenomenon exception case where later occurring high energy stimuli suppresses previous lower energy regions from perception [4]. This concept is counterintuitive to normative understanding. How can an event go back into time a modify the results? This condition is possible because human hearing processes do not occur in instantaneous time. A latency delay is encountered in the perception process which enables high energy stimuli to be experienced prior to lower energy stimuli.

The effects of backward masking are restricted to the 100 msec time interval before the masker onset [2], [9]. The significant part of backward masking occurs within the first 50 msec. After 50 msec the effects dramatically converge to quiet threshold.

2.3 Masking Filter Characteristics

The temporal masking filter models serve as the focal point for the masking calculations. To ensure that the best temporal masking model is found, four masking filter types are analyzed and evaluated. The best masking filters are assumed to be those that best match the filter functions generated by single tone and narrow band noise empirical studies. However to validate this assumption, other filter types will be investigated. The other filter types provide both more and less masking, respectively. The four filter types analyzed are the exponential, linear, second power, and half power.

- Exponential - The exponential curve best matches with empirical data and is modeled as the exponential decay of the logarithm in equation 1.

$$threshold_b^k[\lambda] = (v^k[b] - au\_\min) \bullet \exp(\frac{k-\lambda}{eq}) \qquad (1)$$

where k = short-time frame index, λ = time offset index, b = critical band number in Barks, v is the value in phons for a given bark and time point, au_min is the convergence point for threshold response decay, eq is a factor to normalize the time constant.

- Linear, 2nd power, and 0.5 power filters are defined by equation 2.

$$threshold_b^k[\lambda] = v^k[b] - \left(m_b^k \bullet [k-\lambda]^n\right) \qquad (2)$$

The value of n sets the function type where n = 2, for squared; 1 for linear; 0.5 for square, and m = slope of the function. However m is an unknown that must be calculated. Equation 3 is computed for each spectral point.

$$m_b^k = \frac{v^k[b] - au\_\min}{[k-\lambda]^n} \qquad (3)$$

The equation provides the slope to fit a curve between the current (b,k) value and the minimum threshold point which is assigned to 40dB phon in this case. Convergence time is set to 200 msec. The result is a set of dynamic, frequency dependent masking filters which adapt their estimated hearing threshold responses to the given sound level. Figure 3 graphs the relative threshold decay over a 100 msec period. Backward masking curves are computed from the same equations with a symmetrical response about the k = t_0 point.

The time indices are placed in terms of advance time instead of decay time as in forward masking.

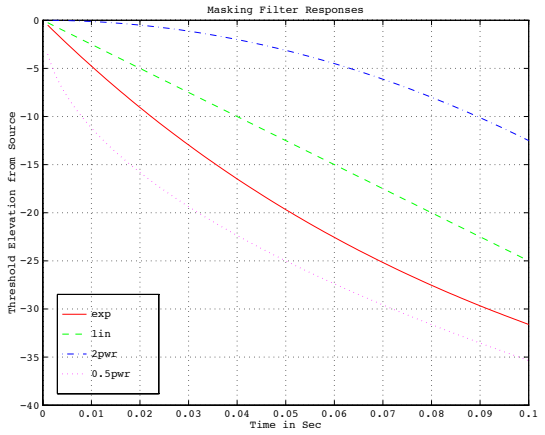

Figure 3. Masking filter curves for the four filter types

The hearing threshold estimates are then compared with the actual sound level values. Those index points where the hearing threshold estimate exceeds the actual sound level values are marked as masked in a corresponding matrix, **R**. The **R** matrix is applied to the Euclidean error distance calculation. This function remove error contributions of the masked spectral points from the overall short-time frame distortion measure, thus retain perceptible errors and rejecting imperceptible errors as modeled in equation 4.

$$BSD_{um}[k] = \sum_{b=1}^{16} \left[L_x^2[b,k] - L_y^2[b,k] \right] r[b,k] \qquad (4)$$

where the overall masked BSD measure is computed by equation 5.

$$BSD_{masked} = \sum_{k=1}^{N} \frac{BSD_{um}[k]}{L_y^2[k]} \qquad (5)$$

2.4 Temporal Masking Example

Figure 4 illustrates the operation of temporal masking. includes a trace of one of the 16 critical band filter intensity level outputs. Of interest are the peaks in the trace. Two hearing threshold curves are calculated and overlaid on the trace to illustrate where temporal masking occurs. Note the data points which fall underneath the hearing threshold and are masked and removed from the error computation.

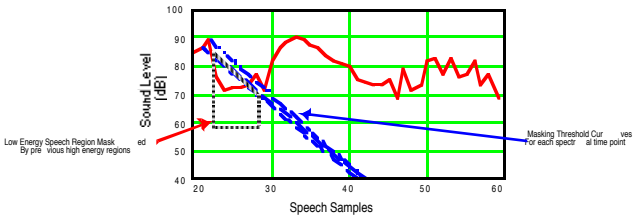

Figure 4. Forward masking illustrating

The backward masking example is very similar to the forward masking example. The only difference is the slopes and intercepts of the masking threshold curves. The backward masking curves have a positive slope (or negative time segment) and intersect the speech intensity curve at its time origin.

An interesting secondary effect of determining temporal masking locale is the creation of a masking intensity matrix. These masked regions of speech are presumed to be inaudible. Notice the high intensity regions of the sone-agram produce a trailing masking pattern that covers-up or masks lower energy "valley" speech regions. Thus, now we have a numerical framework for computing temporal masking locations and a visualization mechanism for their presentation.

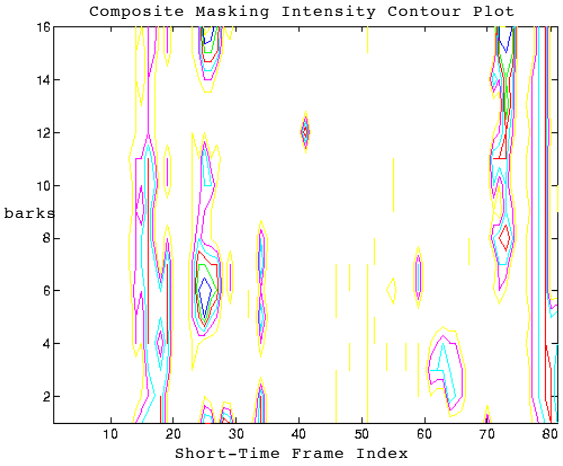

Figure 5. Masking pattern overlay for speech pattern.

3 MASKING RESULTS

3.1 Forward masking

The exponential filter provides the highest degree of correlation between the objective BSD numerical scores and the subjective assessment, in Figure 6. It exceeds both the baseline score and the half-power filter performance over the time interval. The half-power filter masking filter provides some improvement over the baseline performance in the 0-50 millisecond range. Beyond the 50 millisecond point, the performance approaches the baseline mark of 0.910 which also aligns well with a 0.92 result from [1]. The other two filter types; the linear and the 2nd-power, both begin with performance points similar to the baseline. But with increased masking delay, the performance continually declines well below the baseline. It appears that these later models over mask and remove perceptible errors. The overmasking judgment is made in terms of measurements from single tone or narrowband noise masking experiments [2], [10]. This result verifies that existing forward masking models have applicability to dynamic and complex speech signal analysis problems.

As correlation performance of the filters improves, the curves flatten out over a wider time range. This result may be due to the varying intensity levels of potential maskers which dictates varying time intervals for masking effects. Therefore

the best masking curves must cover the range of forward masking times.

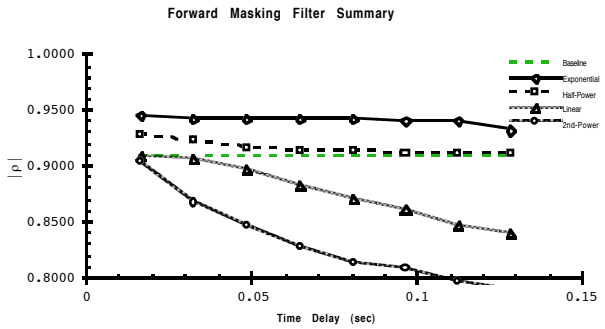

Figure 6. Forward masking filter correlation results.

Examination of the backward masking correlation coefficient performance reveals a closer alignment of the filters shown in Figure 7. This outcomes suggest less distinctive behaviors between backward masking filter types. Another interesting difference of the backwards filter is the peak in the correlation functions in the time range of 20-50 milliseconds. These results are in accordance with the observations reported in [2] which states backward masking effects are minimal at times larger than 50 milliseconds. Therefore the backward masking curves are in consistent agreement with established backwards masking data. The order of the filter performance is the same as in the forward masking case. However the variations between filter types is almost negligible and all filter types under test perform better than the baseline.

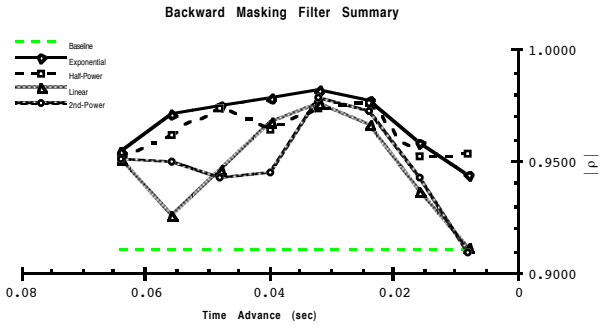

Figure 7. Backward masking filter correlation results.

The filter performance is primarily dominated by the amount of time advance with the filter type being a secondary factor. This result contrasts with the forward masking case where the primary factor was filter shape not the masking time offset.

All filters exhibited similar performance having a correlation peak around 30 milliseconds. This result is consistent with single tone or narrowband noise backward masking experiments [2], [9], [11] verifying applicability to dynamic and complex speech signal analysis.

The backward masking correlation results imply a different set of behaviors governing their effect. It appears backward masking may be more simple to distinguish and be bounded by a smaller region of effect coverage. Therefore its effects are concentrated in a smaller region thus minimizing the filter shape impact on the performance.

4. SUMMARY

The incorporation of temporal masking properties into an objective speech distortion numerical model improves its correlation performance to subjective assessment data by removing inaudible speech samples. Both forward and backward temporal masking properties individually improve the base objective distortion model performance over the baseline model.

Existing temporal (forward and backward) masking models that have been derived from single tone and narrow-band noise experiments are applicable to complex speech analysis to varying degrees. Temporal masking models that conform to exponential responses have the best performance. Backward masking properties have more impact on improving the correlation to subjective assessment data than forward masking. Backward masking models show best performance improvement in the 20-40 millisecond range; consistent with the existing models. Forward masking effects do not produce a peak in the correlation coefficient curve at any delay value rather the performance in the best case is mainly flat over the masking analysis range.

5. REFERENCES

[1] Wang, S., Sekey, A., Gersho, A.; An objective measure for predicting subjective quality of speech coders. IEEE Journal on Selected Areas in Communication, vol. 10, no. 5; 819-829, June, 1992.

[2] Schraf, B.; Buus, S.; Chapter 14 "Audition I" Stimulus, Physiology, Thresholds;" in Basic Sensory Processes II.

[3] Flectcher, H.; "Hearing and Speech in Communication;" D. Van Nostrand Company, Inc.; New York, NY, 1953

[4] B.C.J. Moore, B.R. Glasberg; "Auditory Filter Shapes Derived in Simultaneous and Forward Masking;" J. Acoust. Soc. Am; Vol.70, Num.4; Oct, 1981; pp.1003-14.

[5] G. Ruske; "Auditory Perception and its Application to Computer Analysis of Speech;" Computer Analysis and Perception, Vol. II; CRC Press, Boca Raton, FL; 1982.

[6] Schroeder, Manfred R.; "Models of Hearing;" Proceedings of the IEEE, vol 63, no. 9; September 1975.

[7] Quackenbush, S. R.; Barnwell, T. P. III; Clements, M. A.; "Objective Measures of Speech Quality;" Prentice Hall, Englewood Cliffs, New Jersey; 1988.

[8] Duifuis, H.; "Consequnces of Peripheral Frequency Selectivity for Non-simulatenous Masking;" J. Acoust. Soc. Amer., vol 54, pp. 1471-88, Dec. 1973.

[9] Wright, H. H.; "Temporal summation and backward masking;" J. Acoust. Soc. Amer., no 36; 1964.

[10] Jesteadt, W.; Bacon, S. P.; Lehman, J. R.; "Forward Masking as a function of frequency, masker level, and signal delay;" J. Acous. Soc. Amer.; 1982, 71, pp. 950-962.

[11] Sekey, A.; "Short-term auditory frequency discrimination," J. Acoust. Soc. Amer., vol. 35; pp. 682-690, May 1963.