A MODEL OF DYNAMIC AUDITORY PERCEPTION AND ITS APPLICATION TO ROBUST SPEECH RECOGNITION

Brian Strope and Abeer Alwan

Electrical Engineering Department, UCLA
Los Angeles CA, 90095
bps@ucla.edu

ABSTRACT

This paper derives a non-linear model of dynamic auditory perception. The model consists of a linear filter bank with carefully-parameterized logarithmic additive adaptation after each filter output. An extensive series of perceptual forward masking experiments discussed here, together with previously reported forward masking data, determine the model's dynamic parameters. The model's prediction error of forward masking data has a standard deviation of less than 3.3 dB across wide ranging frequencies, input levels, and probe delay times. We present an initial evaluation of the dynamic model as a front end for an isolated word recognition system, and show an improvement in robustness to background noise when compared to MFCC and LPCC front ends.

1. INTRODUCTION

Short-term adaptation and recovery have been measured in individual auditory nerve firings in response to simple tones and dynamic speech [1-3]. Similarly, human speech perception is sensitive to onsets and, more generally, dynamic spectral cues [4]. Accurate characterization of dynamic audition, together with functional models, are necessary to quantify the perception of non-stationary speech. We also believe that these models will lead to significant application improvements.

Several researchers have proposed dynamic auditory models [e.g. 5-8]. Unfortunately, these models are often too computationally expensive for many recognition applications, and typically show little robustness improvement when compared to more common Mel-frequency cepstral coefficient (MFCC) front ends [9-10]. Other more direct techniques derive dynamic responses by manipulating the time evolution of the feature vectors, and have shown recognition improvements [11-12].

This paper parameterizes a relatively simple dynamic psychoacoustic model from a series of forward masking experiments. Our model is tightly coupled to measured dynamic human audition, while requiring the same order of computational complexity as MFCC front ends. After deriving the dynamic model, we include an initial evaluation using the model in a noise-robust isolated word recognition task. Other applications of the model may extend to speech coding and hearing aid design.

The model includes a linear filter bank for frequency selectivity and independent, time-varying, additive logarithmic adaptation stages after each filter output. Tonal forward masking experiments with varying masker levels and probe delays provide measurements of upward adaptation (post-stimulus recovery), while forward masking experiments with varying masker durations provide measurements of downward adaptation (attack).

We summarize data from our original forward masking experiments, derive dynamic parameters for our model, and then evaluate the noise-robustness of the model's representations.

2. FORWARD MASKING

Forward masking reveals that even though our auditory system may have a 100+ dB dynamic range, over short durations the usable dynamic range is much smaller and largely dependent on previous stimuli. A probe following a masker of similar spectral characteristics is less audible than a probe following silence. As the duration between masker and probe decreases, the probe threshold is increasingly a function of the level of the preceding masker, and decreasingly a function of the absolute probe threshold in silence. We interpret this as the auditory system adapting to the masker. After adaptation to the masker, it takes time to re-adjust so that the relatively quiet probe is audible. The amount of forward masking is also a function of the duration of the masker, reflecting the time necessary for the auditory system to adapt completely to the masker. Forward masking experiments therefore provide an opportunity to measure the rate and magnitude of short-term auditory adaptation and recovery.

We measure the amount of forward masking of a short tone following a long masking tone of the same frequency

This work was supported by NIDCD grant #DC02033.

0-7803-3192-3/96 $5.00©1996 IEEE

(and phase), across varying masker levels, probe delays, and center frequencies. These measurements extend previous data [13]. Experimental details are described in [14].

Figure 1 shows our forward masking data (averaged across five subjects) together with the model's fit, at 1 kHz. Additional data, and resulting model parameters, were also obtained for center frequencies from 250 to 4000 Hz (not shown).

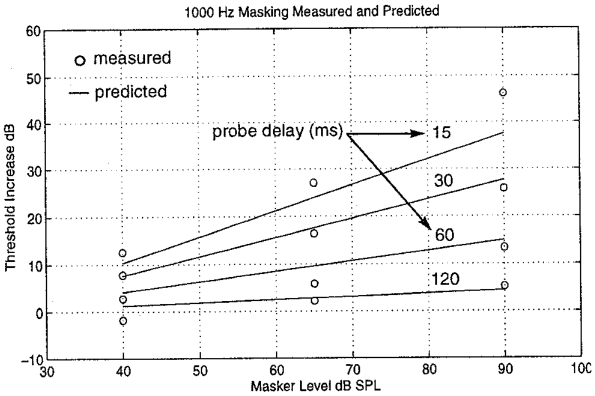

Figure 1: Forward masking data and model fit at 1 kHz.

These forward masking data are used to derive the upward adaptation (recovery) parameters for the model, while other data [15], which measure the change in short-delay forward masking with masker duration, are used to derive downward adaptation (attack) parameters.

3. DERIVATION OF MODEL PARAMETERS

In our perceptual model, a dynamic adaptation stage follows each output of a linear filter bank. The filter bank separates sound into appropriate bands, and the adaptation stages provide a dynamic response largely dependent on preceding inputs. For this computationally-efficient version of the model, Mel-scale filters are implemented by weighting and adding power spectrum points from a Discrete Fourier Transform incrementing at approximately a 100 Hz rate.

We refer to the dynamic adaptation stages as automatic gain control (AGC). However, it is significant that the AGC is implemented as an adapting additive increase to the log energy of the signal, and not as an adapting multiplicative factor after the logarithm [2].

The AGC stages adjust an additive logarithmic offset slowly in time to keep the output level on an I/O target. The rate of adaptation of the AGC stage is described in terms of time constants. An AGC time constant can be defined as the time for the output to settle to within 2 dB of the I/O target in response to an abrupt 25 dB change of the input [16]. Different time constants are used for decreasing offset (attack), and increasing offset (release). Over short durations relative to the time constant, the AGC stage has little time to adapt, and is therefore nearly linear. On an I/O curve, when the input changes abruptly, the output initially tracks the input, moving in a 45 degree line.

For this first-order model, we impose a piece-wise linear I/O curve with a slope of 1 below threshold, a constant compression slope less than 1 from threshold to a high value of equal loudness across center frequency (90+ dB SPL), and a slope of 1 thereafter. The shape of the prototypical I/O curve is motivated by the motility of outer hair cells [17]. Here we generalize this I/O shape to derive a generic model of adaptation. We also impose a fixed internal threshold, corresponding to the static threshold.

Figure 2 describes the geometry necessary to measure the model's prediction of the forward masking threshold with long maskers as a function of masker level and probe delay. Before the masker offset, the output trajectory reaches the target on the I/O curve (point A). As the masker shuts off abruptly, the output trajectory instantly falls along the diagonal. Once the trajectory is below the compressive region, the distance to target is constant, and the model adapts by slowly increasing toward maximum additive offset (region B). At some point during this adaptation, the onset of the probe causes an abrupt transition from below threshold back up along a new diagonal (point C). If the probe level is high enough to place the trajectory above threshold (at the instant of the probe onset) the probe is audible, if the internal level just reaches threshold, the model predicts a forward masking threshold.

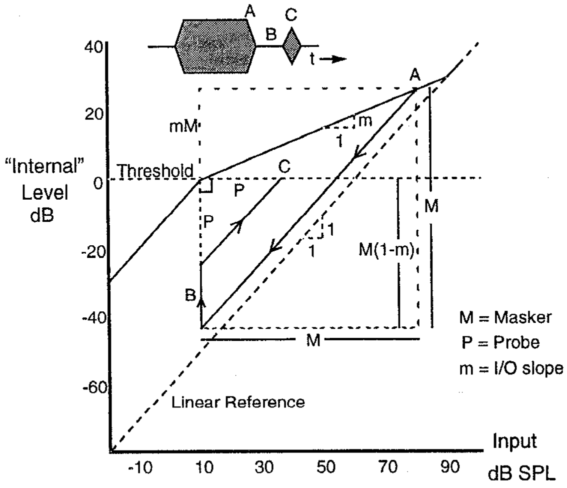

Figure 2: Geometry to derive model parameters.

In our digital implementation, the adaptation of the model is controlled by a first order difference equation which leads to a corresponding exponential decay of the (logarithmic) distance to target, while the input is constant.

From Figure 2, the probe threshold P as a function of masker level M, probe delay nu, I/O slope m, and incremental upward adaptation a is:

$$P = M\,(1 - m)\; a^{nu}$$

Instantaneously after masker offset, nu is zero, and the slope parameter determines the predicted short term usable dynamic range below masker. Therefore forward masking data with the shortest probe delays provide an initial estimation for the slope parameters. Iterative minimization of the MSE of the model's forward masking predictions determine final parameters.

Similar geometric reasoning relates the model's downward adaptation to the change in probe threshold as a function of the duration of the masker. Shorter duration maskers leave less time for downward adaptation, which leads to less forward masking shortly after the masker offset. The change in probe threshold ΔP, as a function of the incremental downward adaptation b, and masker duration nd is:

$$\Delta P = M\,(1 - m)\; b^{nd}\, a^{nu}$$

We solve the probe difference equation for b, and then estimate its value from the differences reported in [15], using the m and a parameters derived from the first experiments. Table I summarizes the model parameters and effective time constants [16] across frequencies. The a and b terms are with respect to sampling the spectrum every 10 ms. Two general trends are clear: the model's slope increases with increasing center frequency; and downward (attack) time constants are 3 to 4 times shorter than upward (recovery) time constants.

TABLE I: Adaptation Parameters

Freq. Hz	Slope m	a	b	up TC	down TC
250	0.19	0.864	0.474	159 ms	31 ms
500	0.20	0.854	0.510	146 ms	34 ms
1000	0.26	0.816	0.543	109 ms	45 ms
2000	0.29	0.851	0.525	135 ms	34 ms
4000	0.34	0.858	0.507	139 ms	31 ms

Finally, the adapting model predicts dynamic relative loudness, and forward masking consistent with five points summarized previously [18]: 1) The amount of forward masking drops nearly linearly in dB with the log of the probe delay. 2) The rate of decay of the amount of masking is greater for increased amounts of masking. 3) An increase in the masker level leads to a fractional increase of the amount of masking (determined by the slope parameter in our model). 4) Shorter duration maskers lead to less forward masking. 5) Forward masking is a function of the frequency relationship between the masker and probe.

Figure 3 shows Mel-scale spectrograms before and after the adaptation stages for the digits "nine six one three," spoken by a male talker. Notice that the adaptation stages emphasize the onset of "nine" and the formant transitions in "one" and "three."

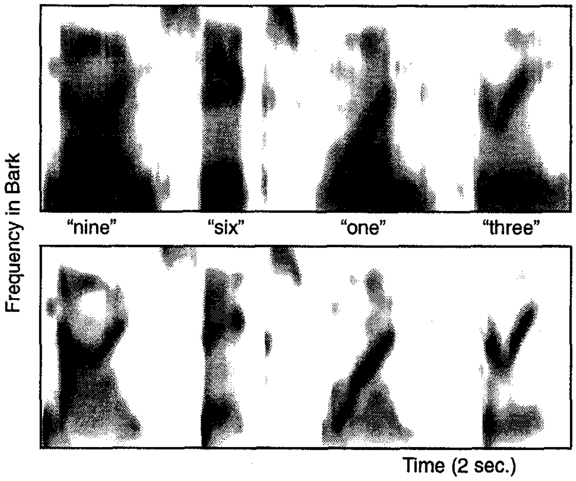

Figure 3: Spectrograms before and after adaptation.

4. ROBUST RECOGNITION EVALUATION

Our goal is to derive a functional model of dynamic audition as a tool for understanding speech perception which can be readily applied to engineering applications. Preliminary evaluation with a dynamic programming-based word recognition system, shows improved robustness to background noise when the dynamic perceptual model is compared with more standard MFCC and LPCC front ends. The dynamic model highlights onsets and spectral transitions which may remain as perceptually salient cues after the addition of background noise. Motivated by the sensitivity of our auditory system to the frequency location of spectral peaks, we also evaluate the dynamic model together with a novel processing technique which isolates local spectral peaks. Interestingly, the combination of a dynamic response with peak isolation is significantly more robust than either process by itself.

Figure 4 shows the degradation of recognition performance in background noise with four front ends: LPCC, MFCC, the dynamic model (DYN), and the dynamic model with peak isolation (DYN+PK). A 13 element cepstral vector, and its temporal derivative are obtained for each front end. The local distance measure is Euclidean, but does not include the undifferentiated cepstral level term (c_0).

The peak isolation algorithm is a novel form of raisedsin cepstral liftering [19]. After liftering the cepstral vector, we transform back to the frequency domain. Points in frequency that are below zero are set to zero. Then the liftered and zero-clipped spectral vector is combined with the original spectral vector to derive a peak- (but not valley-) isolated spectral estimation. Finally, we transform the vector back to the cepstral domain.

Increasing amounts of noise shaped to match the long-term average speech spectrum [20] complicate a speaker-dependent isolated digit recognition task. Consistent with previous results [9], we find MFCC more robust than LPCC. However, both the dynamic model (DYN), and the dynamic model with peak isolation (DYN+PK), are significantly more robust to background noise than MFCC. Following this initial evaluation, future experiments will evaluate the robustness of the model's representations with more difficult recognition tasks, using more sophisticated recognition systems, and including comparisons with additional front ends.

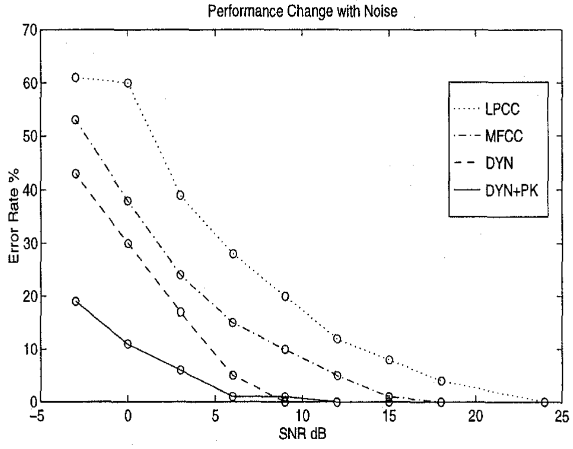

Figure 4: Recognition degradation with background noise.

5. CONCLUSION

This paper derives a model of dynamic auditory perception. Non-linear adaptation and recovery are parametrized from perceptual forward masking data. The complete model provides a tool for analyzing the perception of non-stationary speech.

As a front end to a simple speech recognition system, the model shows promise improving noise-robustness. Other potential applications of the model include hearing aid design and speech coding. Future work will include perceptually-parameterized peak isolation and continuing recognition evaluations.

REFERENCES

[1] N. Y. S. Kiang, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, MIT Press, Cambridge, MA, 1965.

[2] R. L. Smith and J. J. Zwislocki, "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Bio. Cybernetics*, vol. 17, pp. 169-182, 1975.

[3] B. Delgutte, "Representations of speech like sounds in the discharge patterns of auditory nerve fibers," *J. Acoust. Soc. Am.*, vol. 68, pp. 843-857, 1980.

[4] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.* vol. 80, pp. 1016-1025, 1986.

[5] R. S. Goldhor, "Representation of consonants in the peripheral auditory system: a modeling study of the correspondence between response properties and phonetic features," *RLE Technical Report No. 505*, MIT, Cambridge MA, 1985.

[6] J. Kates, "An Adaptive Digital Cochlear Model," *Proceedings, IEEE ICASSP* (Toronto), pp. 3621-3624, 1991.

[7] R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proceedings, IEEE ICASSP* (Paris), pp. 1282-1285, 1982.

[8] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 85, pp. 55-76, 1988.

[9] C. R. Jankowski Jr., Hoang-Doan H. Vo, and R. P. Lippman, "A Comparison of Signal Processing Front Ends for Automatic Word Recognition," *IEEE Trans. Speech and Audio Processing*, vol 3., pp. 286-293, 1995.

[10] S. Sandhu and O. Ghitza, "A comparative study of mel cepstra and EIH for phone classification under adverse conditions," *Proceedings, IEEE ICASSP* (Detroit), pp. 409-412, 1995.

[11] H. Hermansky, N. Morgan, B. Aruna, and P. Kohn, "RASTA-PLP speech analysis technique," *Proceedings, IEEE ICASSP* (San Francisco), pp. 121-124, 1992.

[12] K. Aikaiwa and T. Saito, "Noise robust speech recognition using a dynamic-cepstrum," *Proceedings, ICSLP* (Yokohama), pp. 1579-1582, 1994.

[13] W. Jesteadt, S. Bacon, and J. Lehman, "Forward Masking as a function of frequency, masker level, and signal delay," *J. Acoust. Soc. Am.*, vol. 71, pp. 950-962, 1982.

[14] B. Strope, "A model of dynamic auditory perception and its application to robust speech recognition," Master's Thesis, Dept. of Elec. Engr., UCLA, 1995.

[15] G. Kidd Jr. and L. L. Feth, 1982. "Effects of masker duration in pure-tone forward masking," *J. Acoust. Soc. Am.*, vol. 72, pp. 1364-1386, 1982.

[16] H. Dillon, and G. Walker, *Compression in Hearing Aids*, NAL Rep. No. 90, Austral. Gov. Pub. Ser., Canberra, 1982.

[17] B. Johnstone, R. Patuzzi, and G. K. Yates, G. K., "Basilar membrane measurements and the travelling wave," *Hearing Research*, vol. 22, pp. 147-153, 1986.

[18] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. 3d ed., Academic Press, London, 1989.

[19] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[20] D. Byrne and H. Dillon, "The NAL's new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing*, vol. 7, pp. 257-265, 1986.