PROPERTIES OF AUDITORY MODEL REPRESENTATIONS

Fernando S. Perdigão; Luís V. Sá
Dept. Eng. Electrotécnica & Instituto de Telecomunicações
Polo II Univ. Coimbra, University of Coimbra, 3030 Coimbra - Portugal
E-mail: fp@it.uc.pt, luis.sa@it.uc.pt

ABSTRACT

We address the problem of robustness of auditory models as front ends for speech recognition. Auditory models have been referred as superior front ends when speech is corrupted by noise or linear filtering, but there is not yet a deep understanding of its functioning. We analyze some commonly used auditory models and show that they present some interesting properties which are useful for robust speech recognition. In our view, the short-time adaptation provided by hair cell models is a key factor for this robustness. A disadvantage of auditory models is that the distributions of the obtained features are not well represented by gaussian pdfs. We discuss the problem of parameter transformation in order to use a standard recognizer based on CDHMMs with gaussian pdfs and present some digit recognition experiments.

1. INTRODUCTION

Several studies have been presented comparing the performance of auditory models with traditional signal processing techniques such as cepstral processing (e.g. [5],[6]). Auditory models have been shown to outperform other signal processing schemes for speech recognition tasks specially if the signal is degraded by noise or unknown linear filtering. These studies point out that the magnitude of the reduction in error rate is small compared with the increase in computational load they demand. However, the authors emphasize the need to obtain a better understanding of their functioning in order to further enhance the performance. In this paper we provide a comparative review of three of the most commonly used inner hair cell (IHC) models used in most auditory based front-ends for speech recognition and show that auditory models present some advantages and also limitations when used with standard HMM recognizers. We argue that in order to obtain a better speech recognition performance with auditory models, the raw auditory spectrum parameters need to undergo some transformation. We made some preliminary experiments with digit recognition which show that a DCT (discrete cosine transform) of auditory spectrum parameters outperforms mel-cepstrum based and auditory spectrum parameters.

2. MODELS

Almost all auditory models include a set of linear band-pass filters whose bandwidth increases nonlinearly with center frequency (usually in accordance with critical bands or equivalent rectangular bandwidth-ERB) or are obtained by the way of a more or less detailed model of the middle and inner ear (cochlear models). Besides the fact that filter responses vary from model to model, this stage is conceptually equivalent to the mel-scale band-pass triangular filters used in the MFCC framework. However, the processing is done in the time domain because of the time-domain nature of the following stages in auditory models. The next stage is a model of the functioning of the inner hair cells (IHCs) which includes half wave rectification (HWR) and short-term adaptation. The last stage corresponds to an envelope extractor that produces a mean-rate discharge spectrum.

2.1. Filterbanks

We compared several filterbank schemes namely mel-frequency triangular filters, the Seneff's critical band filterbank (stage I, [2]) and two other filterbanks (in a cascade form) based on cochlear models: one based on Neely's model, [9] and the other implemented by the authors [10], similar to stage I of Seneff's model. We have verified that, at least for the purpose of obtaining an auditory spectrum every 10ms, the filterbank choice is not critical. However, the asymmetry of filter responses that characterize cochlear filters, associated with the adaptation characteristics of the IHCs, is very important especially when detailed timing information is taken into account. In this paper we report only results obtained with our filterbank with 20 channels and CFs ranging from 200 to 3400 Hz.

2.2 Hair-Cell models

In most auditory models the filterbank outputs are applied to an IHC model in order to obtain a representation of the pattern of discharges of the auditory nerve fibers (firing rate). We have done a detailed study of 3 IHC models commonly used: Meddis' model [3], Seneff's (stage II) model [2] and Martens' model, [4], which are represented schematically in figure 1. Although not shown in the figure, the output of the IHC models are low-pass filtered in order to obtain an envelope of firing-rate patterns which are next averaged to form a mean-rate representation of the input spectrum at a needed rate.

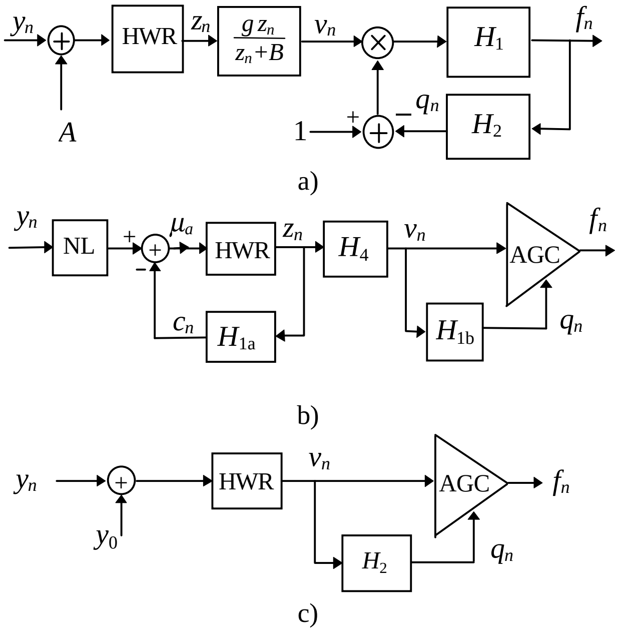

Figure 1. Hair cell models: a) Meddis' model; b) Seneff's model; c) Martens' model. The blocks marked as HWR represent half-wave rectification and AGC automatic gain control as a function of q_n. NL represents the Seneff's model nonlinearity: $1+\arctan(B{\cdot}y_n)$ for $y_n>0$ and $\exp(ABy_n)$ for $y_n<0$. The blocks H_k represent kth order filters.

Figure 1 emphasizes the differences and common aspects among these 3 models. All the models present 3 key characteristics of auditory nerve transmission: for null or very low excitation, the model output corresponds to f_{spo}, the spontaneous firing rate; for very high and sustained excitation the mean-rate output corresponds to the saturation firing-rate, f_{sat}, and for a suddenly applied signal, the firing-rate is initially very high and decays, first very rapidly (rapid adaptation - in the order of a few ms) and then more slowly (short-term adaptation - tens of ms), reaching then a steady state value.

2.2.1. Mean-rate output

In all models represented in fig. 1, the filters H_1 and H_2 ($H_1.H_2$ in the Meddis' model), are low-pass with very small bandwidth. Then, for sustained pure tone excitation of frequency, say, above 500Hz, the output of these filters is almost constant. For this case, the mean output of the models, μ_f, can be computed as a function of σ_y - the RMS input amplitude. The exact expressions, when it is possible to evaluate them, are too complicated; however the values for f_{sat} and f_{spo} can be easily calculated as a function of the model's constants. In order to compare the models it is desirable that these values, as well as the saturation threshold, be the same.

We observed that a careful change of the model constants with this goal has a small or null impact on the adaptation time constants. According to fig.1, once the maximum and minimum value of the mean of v_n, μ_v, have been obtained, then the mean output can be easily computed as follows:

Meddis' model: $$\mu_f = \frac{\mu_v H_1(0)}{1+\mu_v H_1(0) H_2(0)}$$

Seneff's model: $$\mu_f = \frac{\mu_v}{1+\mu_v K_{AGC}}$$

Martens' model: $$\mu_f = \frac{f_{sat}\cdot\mu_v}{\left(B+\sqrt{\mu_v}\right)^2}.$$

The mean μ_v is a compressive function of σ_y (almost linear in Martens' model) but, for all models, the curve of μ_f as a function of $\log(\sigma_y)$ (or with σ_y expressed in dB) is approximately sigmoidal above the threshold of excitation and constant (= f_{spo}) below this threshold. For strong excitation μ_f approaches f_{sat}. The same conclusion applies to noise excitation, but the curve due to the noise is slightly below (less than 2 fires/sec) the curve for a sinusoidal input. The transition from spontaneous to saturated activity for the Meddis' and Seneff's model is approximately 35-45 dB wide. The Martens' model presents a wider dynamic range (55-65dB) due to the different AGC function. This width depends little on the model's constants.

The simulations done with these models show that, at least for the purpose of computing a mean-rate auditory spectrum (every 10ms), they are almost equivalent: the major difference is in the adaptation time constants (fig. 2).

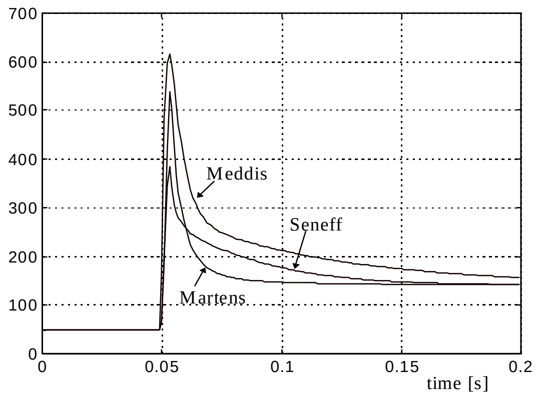

Figure 2. Envelope of firing rate responses (1ms average) to a 1kHz 60dB amplitude tone burst with 2.5ms rising time. Spontaneous and saturation firing rates are 50 and 150 spikes/sec, respectively and excitation threshold is 20dB for the three models. The model constants, according to authors' notations ([2],[3],[4]), are: for Meddis's model: A=10, B=1140, g=1000, h=76980; for Seneff's model: A=2.2; B=0.01, G_{HW}=7.6; for Martens' model: y_0=10 with an AGC function of $f_{sat}{\cdot}v_n/(2y_0+q_n)$. The other constants were unchanged. In this case the rate-intensity functions are practically identical in all models.

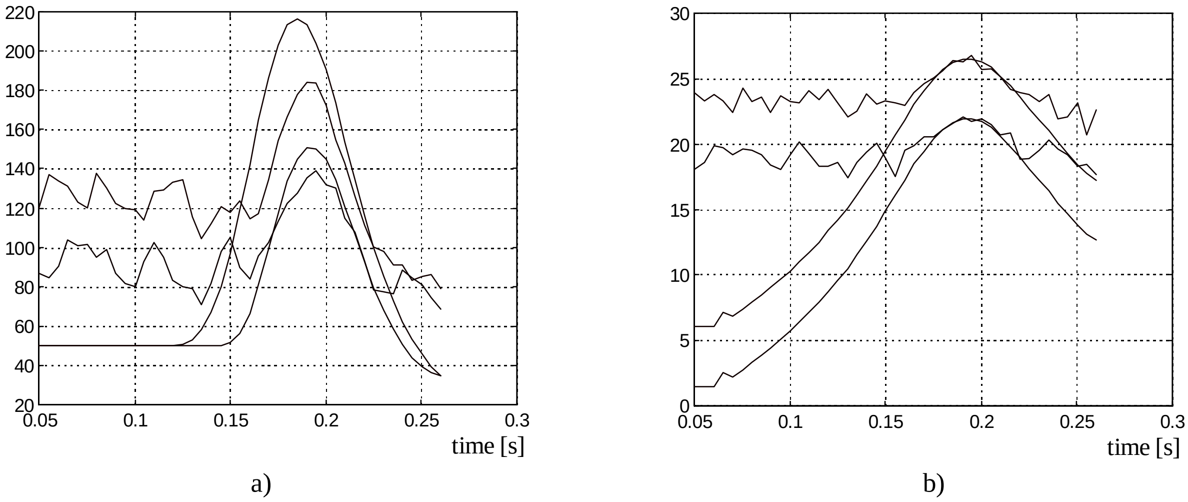

Figure 3. Output at one filterbank channel with CF=1050Hz with a signal with maximum amplitude V plus noise. a) Mean-rate output. b) Log energy. The smooth curves are for SNR=∞, V=60dB and V=80dB; other curves: SNR=20dB, V=60dB (bottom) and SNR=0dB, V=80dB (top).

We have experienced difficulties in using Meddis' model for any other sampling frequency than 20kHz. On the other hand, the functional (or phenomenological) model of Martens does not have this problem and is computationally a bit more efficient than the other two models. So, for the experiments reported here with telephone speech signals sampled at 8kHz we used Martens' IHC model without upsampling the signals.

3. COMPARISON WITH LOG ENERGIES

It has been reported that auditory models present superior encoding capabilities when compared to standard front-ends for speech recognition proposes [5],[6]. This robustness is especially evident when speech is corrupted by linear filtering as well as by additive noise. However, it not clear why auditory models work as well as they do. It seems to us that short-time adaptation is the most important aspect of auditory models in this respect. We report a simple experiment with a chirp signal with decreasing frequency (2kHz to 0.5kHz in 200ms) plus noise. Figure 3 shows the mean-rate output (Martens' model) and log energies (logE) of channel 10 (CF=1050Hz) taken every 5 ms for 2 different signal amplitudes and 3 SNRs. It is seen that noise degrades the responses in both cases; however, the mean-rate representation is more robust: if the noise picked up by the filterbank filter is insufficient to saturate the IHC, then, when the signal frequency approaches the filter CF, the IHC is not yet completely adapted and produces a stronger response. For the case of logE, not only the response curve is broader but also the noise masks almost completely the response. The adaptation has the property of enhancing changing parts of the signal while it tends to normalize stationary parts. In order to overcome the problem of, simultaneously, canceling the effects of noise and channel filtering, Hermansky and Morgan [7] propose a "lin-log RASTA", i.e. a transformation for energies that is linear for low signal energy (in order to cancel the noise), but logarithmic for higher energies (in other to remove channel distortions). It seems that auditory models present approximately this property. This could explain the robustness of auditory models when tested under those conditions. However, analyzing the histograms of auditory model outputs it is evident that the parameter pdfs are not gaussian which may hamper the speech recognition performance. We made histograms of mean-rate outputs for the vowel /E/ and fricative /s/ in the Portuguese digit /sete/ and, in fact, the distributions are far from having a gaussian form, specially for low energy signals (due to HWR they tend to have a truncated gaussian form plus an impulse at f_{spo}). Also, the noise increases the means and reduces the variances in the same way as with other parameters. This means that parameter transformation may be useful when a conventional speech recognizer, based on continuous gaussian mixtures, is to be used. An obvious method to do this transformation, due the apparent similarity of mean-rate spectrum with lin-log RASTA, consists in the use of a DCT (discrete cosine transform), in the same way as used for transforming logE to cepstrum. We also tried the method described by Nadeu et al [8] which consists in filtering a characteristic vector in the channel index domain (or in the cochlear axis domain). This operation consists in subtracting, for each frame, the average spectrum (over the channels) and then high-pass filter the result with a FIR filter. In turn, this high-pass filtering resembles the LIN (lateral inhibition network) proposed by Shamma [1], which has the effect of enhancing the peak to valley ratio of the auditory spectrum. This also tends to make the parameter distributions a bit more symmetrical . In the next section we describe some recognition experiments using these techniques.

4. RECOGNITION EXPERIMENTS

In order to test the validity of the proposed transformations on the auditory spectrum, we made 4 tests with digit recognition using CDHMMs with 7 states and 6 mixtures per state. The digits are from a Portuguese telephone speech database collected all over the country. In all the 4 experiments we trained and tested the recognizer with 1864 and 1759 digits, respectively. We did not use delta coefficients since we wanted to test only the spectral parameter richness. In the first test we used mel-cepstrum coefficients (MFCC) with triangular filters and 12 cepstral coefficients. The second test consisted in mean-rate spectrum with 20 parameters. Next we tried Nadeus' et al method [8], with a first order FIR filter, $1\text{-}0.5z^{-1}$, and also 20 parameters per characteristics vector. Finally we made a test with 12 DCT coefficients taken from the mean-rate spectrum. The recognition performance was, in the same order as described: 72.1%, 76.4%, 76.8% and 81.4%.

The auditory spectrum seems to be more robust than mel-cepstrum. This is even more evident for the DCT of the mean-rate spectrum which provided an increase of about 9% in recognition performance. The reason why the 3rd test did not obtain a better score, as we expected from Nadeu's et al results, may be related with the "gaussianity" of the parameters which is much more evident for the DCT parameters then for the other case.

CONCLUSION

We have shown that auditory models produce a rich representation of speech signals, useful for robust speech recognition. The most important aspect of this robustness seems to come from adaptation of IHC. IHC models can be used in an computationally-efficient way without losing the main properties of these cells. Transformation of the mean-rate spectrum can be done with advantage in order to use standard recognizers based on CDHMMs. We have shown that a DCT of mean-rate spectrum is much more robust than the MFCC representation.

REFERENCES

[1] Wang, K; Shamma, S., "*Self-Normalization and Noise-Robustness in Early Auditory Representations*", Speech & Audio Proc., Vol. 2 No. 3, July 1994.

[2] Seneff, S., "*A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing*", J. Phonetics, 16, 1988, pp 55-76.

[3] Meddis, R. "*Simulation of Mechanical to Neural Transduction in the Auditory Receptor*", J. Acoust. Soc. Am. 79 (3), March 86, pp 702-711.

[4] Martens, J.; Immerseel, L., "*An Auditory Model Based on the Analysis of the Envelope Patterns*", ICASSP'90, pp. 401-404.

[5] Jankowski, C.; Vo, H; Lippmann, P., "*A Comparison of Signal Processing Front Ends for Automatic Word Recognition*", Speech & Audio Proc., Vol.3 No.4, July 1995.

[6] Stern, R. et all, "*Signal Processing for Robust Speech Recognition*", in "*Automatic Speech & Speaker Recognition-Advanced Topics*", Kluwer, 1996.

[7] Hermansky, H.; Morgan, N., "*RASTA Processing of Speech*", Speech & Audio Proc., vol.2, No.4, Oct. 94.

[8] Nadeu, C. et al, "*On the Decorrelation of Filter-Bank Energies in Speech Recognition*", EUROSPEECH'95, pp. 1381-1384.

[9] Neely, S., "*A Model of Cochlear Mechanics with Outer Hair Cell Motility*", J. Acoust. Soc. Am., 94 (1), July 1993, pp 137-146.

[10] F. Perdigão; L. Sá, "*A Cochlear Model for Speech Processing*", RECPAD'95, Aveiro-Portugal, 1995.