AUDITORY MODELS AS FRONT-ENDS FOR SPEECH RECOGNITION

Fernando S. Perdigão, Luís V. Sá
Instituto de Telecomunicações - Pólo de Coimbra
Dept. Eng. Electrotécnica - Pólo II da Univ. Coimbra
3030 COIMBRA - PORTUGAL

ABSTRACT

In this paper we compare the performance of a speech recognizer using several front-ends based on the properties of the peripheral auditory system. An efficient model is also proposed which gives good results. Some interesting properties of firing rate based representation are analyzed. In order to evaluate the effectiveness of this model, as well as of other auditory models, we performed several digit recognition experiments under clean and noisy conditions. Results show that the auditory model's representation is considerably affected by additive noise, however a simple noise reduction mechanism substantially improves the recognition performance.

1. INTRODUCTION

Speech recognition systems include an initial processing stage that converts speech signals into a feature representation useful for further processing. Most of these front-ends are based on standard processing techniques such as filter banks or linear prediction (LPC). Auditory models can also be used to as front-ends for speech recognizers. They incorporate known properties of auditory system in the resulting representation. Most auditory models use a filter bank to represent the cochlear functioning. A second stage models the inner-hair cell (IHC) transduction of basilar membrane motion into auditory nerve firing patterns. Most often the signal representation is based on average firing rate of fibers corresponding to a given channel (mean-rate spectrum).

Due to the time domain nature of all the processing stages in auditory models they are computationally expensive. In order to alleviate this problem an efficient model is proposed that uses short-term energies instead of full rate time signals. This model simulates with a reasonable accuracy the main characteristics of the auditory periphery and is particularly adequate for the recognition task considered.

Auditory models are considered robust front-ends for speech recognition under adverse conditions. A study was carried out in order to understand the advantages they provide. Some interesting relationships between auditory model representation based on mean-rate with other processing techniques are outlined, in particular Lin-Log RASTA (J-RASTA) [3]. As with J-RASTA, model parameter adaptation to noise conditions must be done in order to obtain robustness against adverse conditions.

The recognition experiments presented here extend other published results on auditory-based speech recognition [5].

2. AUDITORY MODELS

The first stage of the models considered here uses a filter bank as a model of the cochlea. Three filter banks were considered: the gamma-tone filter bank [13], The Seneff filter bank (Stage I) [12] and a filter bank based on a physical coclear model. This last filter bank is briefly described below. The second stage uses a model of inner-hair cell/auditory nerve synapse associated to each filter bank channel. Three models of the IHC/synapses were considered as well as a new and efficient model. Some properties of this representation are outlined below and compared with conventional spectral representations.

2.1 Filter Banks

A filter bank based on cochlear properties can be obtained using a physical model of the cochlea. The cochlear model implemented is based on Neely's model [9] which includes the feedback effect of outer hair cells. This *active* model is remarkable in reproducing neural tuning curves of available cat data. However in order to have a model of the human cochlea, the basilar membrane (or partition) impedance must be changed accordingly. The definition of the partition impedance is a key feature in cochlear modeling. As pointed out by de Boer [2], several active models are formally equivalent, the partition impedance being defined as a sum of a passive impedance with another impedance related to the organ of Corti. Another generalization, useful to identify impedance parameters, consists on defining this partition impedance, in the frequency domain, as a product of a 2nd order impedance by a transfer function. In most cochlear models the partition impedance can be put in this form as a 4th order function. This form emphasizes the role of particular physical parameters in the resulting tuning curves, so the frequencies and quality factors can be chosen in order to have an impedance with grossly the same properties as in Neely's model but with CFs according to the human cochlear map. For the external and middle ear the Rosowski's model [11] was used.

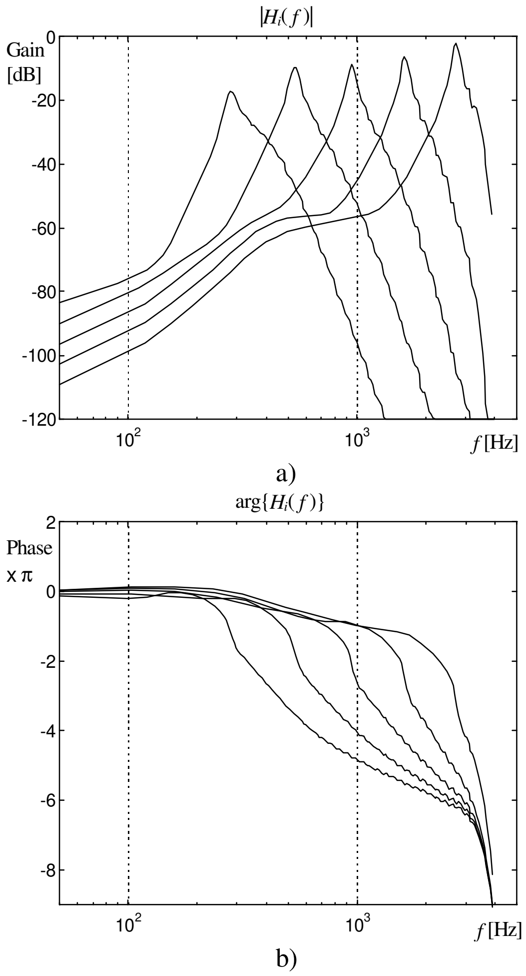

Figure 1: Frequency responses of the filter bank derived from the cochlear model. The responses shown corresponds to channels 4, 11, 18, 25 and 32. a) Magnitude. b) Phase.

From the computed responses a bank of cochlear digital filters was implemented by taking the transfer functions between successive and equally spaced responses along the cochlear path and translating them directly to discrete time by a LMS criterion (deterministic Prony method in the frequency domain [15]). This results in a filter bank represented as a cascade of digital filters. The outlined procedure works fairly well for a linear model. Another procedure must follow if non-linear characteristics are to be included in the model as, for instance as in [6]. However this was not been considered in this work. Figure 1 shows some responses of a filter bank with 35 channels with CFs ranging from 3400 Hz to 200Hz for a sampling frequency of 8kHz. The filters are 6th order except the first one which has a higher order (12) since it includes all the effects from external ear to the first cochlear place considered (CF of 3.4kHz).

The stage I of Seneff's model is a bank of 40 filters working at a sample rate of 16kHz. In order to use this model in the present work we take the first 35 channel responses and translate them to a sampling frequency of 8kHz again using a LMS criterion. The resulting responses are in good accordance with the original responses.

For the gamma-tome filter bank the Slaney implementation [14] was used .

2.2 IHC/Synapse Models

Three models of the inner hair-cells (IHCs) and auditory nerve synapses were considered: the Meddis' model [8], Seneff's model (stage II) [12] and Martens-Immerseel's model [7]; [10]. All IHC/Synapse models basically include half-wave rectification (HWR) to model the transduction of IHCs and an automatic gain control system to model the short-term adaptation. The simulations done with these models show that they are almost equivalent for the computation of a mean-rate auditory spectrum every 10 ms. The main differences among the models are due to the adaptation stage, in particular adaptation time constants and the peak value of responses at the onset of bursts.

2.3 Proposed Functional Model

Due the time domain nature of all processing stages in IHC/synapse models they are computationally expensive when compared with frequency domain techniques such as MFCC representation (mel frequency cepstral coefficients, [1]). However, when a mean-rate representation is desired every 10ms or so, a much more efficient procedure can be followed. The model stage corresponding to IHC transduction is basically modeled by a half wave rectifier operation which preserves the RMS envelope of the subband signals. If adaptation characteristics could be modeled in terms of the sort-term RMS signal at the output of this stage instead of the full rate signal, a considerable computational gain would be obtained. Preliminary experiments done with such a model confirm that the short-term adaptation can be reasonably modeled. To do so the Martens-Immerseel model was used but with a sampling frequency of 100Hz. The first stage of the model was simplified using the following expression:

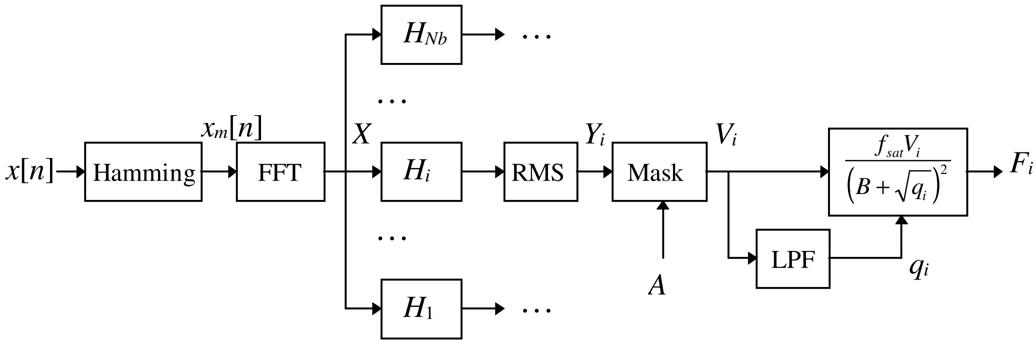

Figure 2: Block diagram of the functional model. A: rate threshold (masking level); f_{sat}: saturation firing rate; B: constant that determines the spontaneous firing rate (f_{spo}); F_i: output firing rate. H_i: filter i of the filter bank.

$$V_i[m] = \sqrt{Y_i^2[m] + A^2}\ , \tag{1}$$

where $V_i[m]$ represents the RMS signal at the output of HWR stage; A represents the model threshold; $Y_i[m]$ is the short-term RMS energy at the output of channel i of the filter bank and m represents the frame index. We have also considered this operation simply as:

$$V_i[m] = \max\left(Y_i[m], A\right), \text{ or} \tag{2}$$

$$V_i[m] = A + \max\left(Y_i[m], 0\right), \tag{3}$$

with approximately the same results. A more accurate relationship between the RMS signal and the short-term energy for the case of sinusoidal signals can be used.

The next stage uses a low pass filter which determines the adaptation time constants and an AGC stage as in Martens-Immerseel's model. The filter time constants were adjusted by comparing model responses to sinusoidal bursts with the original full rate responses. A good match was obtained with time constants of 18ms and 40ms.

This model can be made much more efficient by using the FFT to compute the RMS values. In this way this functional model presents almost the same computational complexity as the MFCC front-end while it preserves the main characteristics of the auditory periphery. Figure 2 shows a block diagram of the model.

2.4 Mean-Rate Representation

IHC/synapse models present 3 main characteristics of auditory nerve transmission: a) an excitation threshold below which the firing rate is constant - the spontaneous firing rate; b) a dynamic range of operation in which the firing rate grows almost linearly with the log-power stimuli but saturates for high levels, and c) a short-term adaptation that enhances firing rate for rapid variations in the signal energy. These characteristics can be related to noise reduction techniques used in speech processing. In fact, from a signal processing point of view, the model threshold imposes a "masking level" below which the firing rate is constant (spontaneous rate) and this spontaneous rate can be viewed as a "noise floor" in the resulting representation. For stationary signals the rate-intensity curve relating firing rate to excitation level is, above the threshold, an approximately sigmoidal function of the logarithmic power. This represents a compression function which, in the fiber's dynamic range, is not much different from a log compression and means that the signal representation based on firing rate can be viewed as a "pseudo log-spectrum". Moreover, the short-term adaptation can be viewed as a band-pass filtering of compressed time energy trajectories. The characterization of adaptation as a filtering mechanism can easily be done for simple signals such as tone bursts. In this case the responses can be associated to a step response of a linear filter if the effect of compression and previous adaptation state is considered in a preceding stage. Figure 3 shows the model used in this characterization and figure 4 shows the frequency responses of such linear filters for several adaptation states and for the case of Meddis' model. For speech signals we can expect similar responses since responses do not vary much with the state of adaptation. As a consequence, mean-rate representation has a clear resemblance to RASTA processing [3] specially with Lin-Log RASTA. In this technique a masking level is also applied to the perceptual subband power before log compression and a band pass filtering (RASTA filtering) of the log-energy trajectories is performed to compensate against spectral variability. The model of figure 3 also applies to the RASTA case, however the RASTA band-pass filter has a zero DC gain which effectively removes long-term spectral distortions.

These characteristics are essential for the robustness observed when auditory models are used in speech recognition systems under adverse conditions. On the other hand, the mean-rate representation can be strongly affected by noise if model thresholds are not adjusted to the noise conditions. Due to the limited dynamic range of fibers, large speech level variation may also deteriorate recognition performance. Both a level compensation and a noise suppression technique based on a center clipper applied in front of IHC models have been already used with good results [16]. Results presented here also confirms the effectiveness of these techniques.

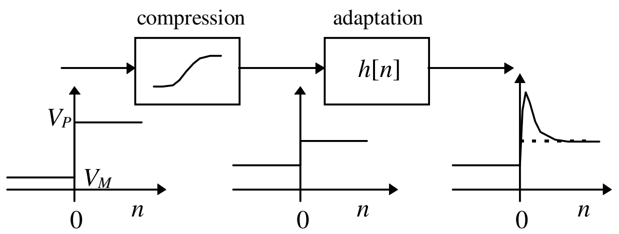

Figure 3. Model to characterize adaptation as a linear filtering in the case of sinusoidal bursts. V_M represents the sinusoid amplitude that adapts the response to a burst of amplitude V_P.

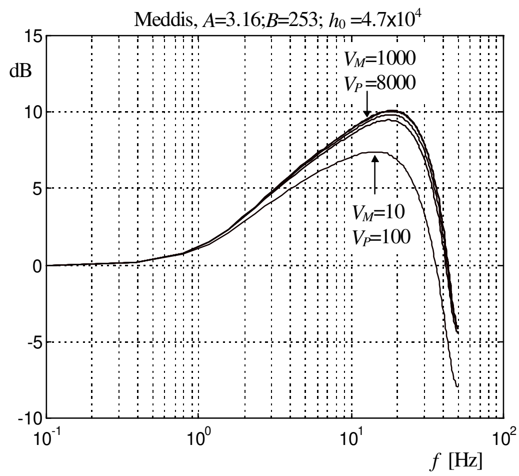

Figure 4. Filter responses of filter $h[n]$ according to figure 3 for several states of adaptation. This responses are for Meddis' model. The other models yield similar responses.

2.5 Noise Subtraction

In the case of the functional model a spectral subtraction based on subband energies was used with good results. Like in [16] the goal is to obtain an almost fixed noise floor at $V_i[m]$ for pure noise segments independently of the noise level. To attain this goal a knowledge of the pdfs of RMS signals ($Y_i[m]$) for noise conditions, is required to evaluate the expected value of $V_i[m]$ (or $q_i[m]$ because the low-pass filter has an unitary DC gain). An empirical study with gaussian input noise has shown that these signals have a distribution than can be approximated in an excellent way by a gamma distribution. The two parameters of the gamma pdf depend on the noise variance and on the filters' frequency response shape. One parameter can then be pre-computed in each channel. The spectral subtraction was applied to $Y_i[m]$ in equation (2) or (3). The value to subtract is a strong non-linear function of the noise standard deviation and is defined in a way similar to the method described in [16], but with a gamma instead a gaussian pdf. This spectral subtraction represents a small computational overhead with respect to the other computations.

4. RECOGNITION EXPERIMENTS

4.1 Conditions

In order to evaluate the proposed model as well as the performance of several auditory models in a speech recognition task, several recognition experiments were carried out using a Portuguese telephone speech database of isolated digits collected all over the country. This database has about 800 speakers and about 4200 isolated digits. Approximately one half of these digits was used for training and the other half for testing. A CDHMM recognizer [17] was used in the experiments. Diagonal covariance matrices and 7 emitting states for each word with 5 or 6 component gaussian mixture were used. Pauses were represented by a 3 state HMM.

In auditory-based front-ends feature vectors were generated by performing a DCT of the mean-rate spectrum in order to use diagonal covariance matrices in the HMMs. The filter banks have 35 channels and 12 DCT coefficients were used. In most cases the energy was also included as the first DCT coefficient.

The recognizer was also evaluated under noise conditions. In this case white noise was added with a signal-to-noise ratio of 20dB. Energy levels were measured averaging the signal energy of speech frames and discarding frames corresponding to pauses. Experiments were also conducted with speech shaped by a linear filter that approximates the response of a telephone set. An additional experiment uses a combination of additive noise and filtered speech.

4.2 Results

The first set of experiments were conducted in order to evaluate the filter bank performance alone in terms of the subband short-term energies. A triangular filter bank used in MFCC analysis and the trapezoidal filter bank used in PLP analysis [4] were considered in addition to the above referred filter banks. All filter banks use 35 channels. The recognition rates are approximately identical, about 90% with a small variation, less than

1.4%. This suggests that these filter banks are essentially equivalent for a representation based on short-term energy.

The inclusion of the IHC/synapse models to obtain a representation based on mean-rate spectrum enhances the recognition rates from 2% to 6% depending on the models and values for spontaneous and saturation firing rate. The gamma-tone filter bank associated with IHC/synapses models gave best results in most cases. The best result (96.74% recognition rate) was obtained with the functional model. However when level normalization and a center clipper is applied (to the Martens-Immerseel's model only) the corresponding front-end led to a better result.

When we add noise to the signals a dramatic decrease of recognition performance is observed in all representations, specially in auditory models. In fact, the auditory models have shown worse results than MFCC or LPC cepstral representations. This contrasts with the results presented in [5] which indicate a slightly reduction of the error rate in noise with the use of auditory models when compared to the MFCC representation. However in that work delta coefficients were always used and the results pertain to speaker dependent recognition. In the present case this performance degradation occurs because the mean-rate responses in speech pauses increase substantially as a consequence of a low threshold used in IHC/synapse models. With noise compensation the performance degradation is much less evident which confirms previous results [16].

Table 1: Recognition rates for several front-ends

Label	*Clean*	*SNR=20dB*	*Filter*	*Noise+Filter*
LRasta	97.57	90.22	97.55	89.66
JRasta	96.27	89.59	93.71	88.56
MFCC_E_D	97.67	80.49	94.81	47.57
LPC_E_D	97.95	85.90	94.96	62.77
GMedH1	94.00	63,33	88.96	38.16
SSenM1n	88.01	68.00	87.30	54.25
HMarH2c	92.84	84.43	88.88	76.64
HMarH2nc	96.95	86.93	94.22	87.13
GMarH2nc	97.85	90.76	96.79	93.01
GMFAns	97.19	94.45	93.77	91.73

Table 1 shows some of the recognition results obtained for clean and distorted speech. The first 4 rows concern conventional approaches. Label JRasta refers to Lin-Log RASTA with J=10^{-6} [3]. The letters "_E_D" in the labels refers to addition of energy and delta parameters (26 parameter feature vectors). The other rows concern auditory models with different composition of a filter bank (H: cochlear; S: Seneff; G: gamma-tone), IHC model (Med: Meddis; Sen: Seneff; Mar: Martens-Immerseel); high or medium spontaneous rate (H/M), level normalization (n) and noise reduction with level clipping (c). The last row refers to the functional model with the noise subtraction described above. The result obtained with this model is a good one if we compare its computational complexity with the other representations for the same size of feature vectors.

Also remarkable is the recognition performance obtained with the RASTA representation (feature vectors with 9 parameters).

5. CONCLUSION

The results obtained in this work show that auditory models and the mean-rate representation in particular present interesting properties useful for robust recognition. It was shown that there is a close relationship between mean-rate spectrum and J-RASTA processing. This could explain the why auditory models could provide some noise robustness and normalization to spectral variations. In addition, short-term adaptation could provide a better segmentation of speech signals into states in the HMM framework. Noise can degrade substantially the performance of recognizers if no compensation is provided against it in this representation. A functional model of IHC/synapses functioning was proposed which has a computational load equivalent to the MFCC analysis but leads to a much better recognition performance. This model seems to be a reasonable compromise between accuracy of modeling and computational efficiency.

ACKNOWLEDGMENT

This work was done in the scope of the national project REC - Speech Recognition Applied to Telecommunications (program PRAXIS XXI).

REFERENCES

[1] S. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE trans. ASSP, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[2] E. de Boer, "On Equivalence of Locally Active Models of the Cochlea", J. Acoust. Soc. Am. 98(3), pp. 1400-1409, Sept. 1995.

[3] H. Hermansky, N. Morgan, "RASTA Processing of Speech", IEEE trans. Speech and Audio Proc., vol 2, No. 4, pp. 578-589, 1994.

[4] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", J. Acoust. Soc. Am. 87(4), pp. 1738-1752, Apr. 1990.

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

[6] J. Kates, "A Time-Domain Digital Cochlear Model", IEEE, Signal Proc., Vol.39, No.12, pp. 2573-2592, Dec. 1991.

[7] J. Martens, L. Immerseel, "An Auditory Model Based on the Analysis of the Envelope Patterns", ICASSP-90, pp. 401-404, 1990

[8] R. Meddis, "Simulation of Auditory-Neural Transduction: Further Studies", J. Acoust. Soc. Am. 83(3), pp. 1056-1063, Mar. 1988.

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

[10] F. Perdigão, L.V. de Sá, "Properties of Auditory Model Representations", Proc. EuroSpeech'97, Vol. 5, pp.2499-2502, Rhodes, Greece, Sept. 1997.

[11] J. Rosowski, "Models of External- and Middle-Ear Function", in "Auditory Computation", R. Fay, A. Popper, Eds., Springer Handbook of Auditory Research, Springer-Verlag, vol. 6, pp. 15- 61, 1996.

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

[13] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank", Apple Tech. Rep. No. 35, 1993.

[14] M. Slaney, "Auditory Toolbox", Apple Tech. Rep. No. 45, 1993-1994.

[15] C. Therrien, "Discrete Random Signals and Statistical Signal Processing", Prentice Hall Signal Processing Series, Prentice Hall, NJ, 1992.

[16], H. Vereecken, J.P. Martens, "Noise Suppression and Loudness Normalization in an Auditory Model-Based Acoustic Front-End", trans. ICSLP'96, pp. 566-569, 1996.

[17] S. Young, J. Jansen, J. Odell, Dave Ollason, P. Woodland, "The HTK Book - Hidden Markov Model Toolkit", Cambridge University, Dec. 1995.