A Noise Suppression Technique using an Auditory Model

Fernando S. Perdigão, Luís Vieira de Sá,
Instituto de Telecommunicações / Dep. Eng. Electrotécnica
Pólo II da Universidade de Coimbra, 3030 Coimbra, Portugal
E-Mail: fp@co.it.pt, luis.sa@co.it.pt

Abstract

In this paper we describe an efficient speech analysis model based on the properties of the peripheral auditory system. This model uses a bank of gamma-tone filters followed by a model of adaptation that occurs in auditory nerve fibers. The model produces a speech representation in terms of the mean firing rate. A noise suppression mechanism is included in order to obtain higher speech recognition robustness under noise conditions. We performed several digit recognition experiments under clean and noisy conditions in order to evaluate the effectiveness of this noise suppression method. Results show that the method substantially improves the recognition performance.

I. INTRODUCTION

Most of speech analysis front-ends for speech recognition are based on standard processing techniques such as filter banks or linear prediction (LPC). Other front-ends incorporate some properties of auditory system, such as the well known MFCC model or the PLP model [4].

Auditory models have also been used as front-ends [3,5]. They incorporate a detailed description of the auditory periphery, modeling not only the cochlea but also the inner-hair cell (IHC) transduction of basilar membrane motion into auditory nerve firing patterns. The output representation of auditory models is most often based on average firing rate of fibers corresponding to a given channel (mean-rate spectrum).

One of the most important properties of the auditory functioning is adaptation. Adaptation consists on a decrease of fiber response for a sustained excitation. In turn, if the signal energy varies suddenly, the firing rate enhances that variation in a short time course of about 10 ms. This mechanism is important for the detection of plosive sounds or the onset/offset of vowel sounds in speech.

Because of the non-linear behavior of adaptation, auditory models must operate in the time domain and are sometimes very computationally expensive. In order to alleviate this problem an efficient model was proposed [1] that uses short-term energies or RMS values instead of full rate time signals. The model produces a speech representation which is closely related to the Lin-Log RASTA (or J-RASTA) representation [4] and presents some kind of robustness against convolutional noise [1]. However, for moderate SNRs the speech representation is also affected, mainly due to the adaptation of the responses by the noise which reduces the phonetic contrast enhancement due to adaptation. When this model is used in a HMM recognizer the recognition accuracy degrades substantially due to the mismatch between the training and testing conditions.

A technique to overcome such problem that incorporates a center-clipping inside the auditory model was proposed by Vereecken and Martens [5]. This technique proved to be effective in reducing the recognizer error rate and was shown to be superior to applying power spectral subtraction on the outputs of the model. The clipping level is adaptively adjusted so that the signal level at the output of the half-wave rectification stage is almost constant for non-speech segments of the signal. An analog method was used in the present model, applied to the energy of the filter-bank outputs. Experiments show that such noise suppression mechanism is effective and improves the recognition performance.

II. THE AUDITORY MODEL

A block diagram of the auditory model is shown in figure 1. The first stage consists on a conventional filter-bank analysis (as in MFCC, PLP or RASTA analysis). The signal is blocked into frames of N=256 samples (32 ms) with a frame rate of 10 ms and a Hamming window is applied to each frame. The FFT of each frame is then applied to a gamma-tone filter-bank [2] with 35 channels. This filter-bank presents filter bandwidths and center-frequencies according to the human cochlear operation. The outputs of this first stage are the square root of energies (RMS values) at each channel, which are computed according to:

$$Y_i[m] = \frac{1}{N}\sqrt{\sum_{k=0}^{N}\left|X_m(k)H_i(k)\right|^2} \ , \qquad (1)$$

where $X_m(k)$ is the FFT of frame m and $H_i(k)$ the frequency response of the channel i filter.

The second stage simulates IHC transduction, basically a half- wave rectifier (HWR) operation. Simulations with several auditory models have shown that the *mean* of the HWR signal in IHCs and the fiber threshold, A, can be reasonably modeled with the following measure:

$$V_i[m] = \max\left(Y_i[m], A\right), \qquad (2)$$

which is proportional to $Y_i[m]$ except for values less than A. The constant A can also be viewed as a masking level applied to the signal representation because low level energies does not affect much $V_i[m]$. The value of A is about one thousandth of the maximum amplitude of the signals. We have also considered in this stage a simpler measure,

$$V_i[m] = A + Y_i[m]. \qquad (3)$$

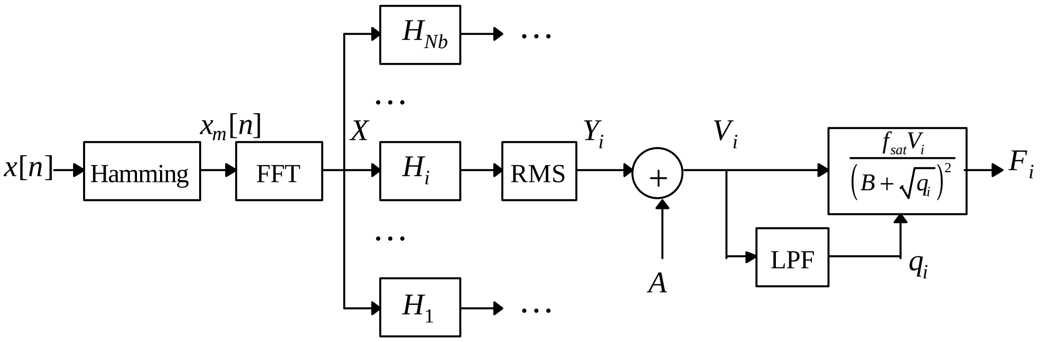

Fig. 1 Block diagram of the auditory model. The model constants are A: rate threshold (10). f_{sat}: saturation firing rate (150 fires/sec.); B: constant that determines the spontaneous firing rate, f_{spo}, (10 fires/sec.); H_i: filter i of filter-bank with N_b=35 channels. LPF: low-pass adaptation filter used in Martens-Immerseel's model with analog time constants of $\boldsymbol{t}_1$=18 and $\boldsymbol{t}_2$= 40 ms. F_i: output firing rate.

The last stage is the adaptation model and incorporates the Martens-Immerseel's adaptation model [3] operating, not at the signal sampling frequency but at the frame rate. The model output is the mean firing rate $F_i[m]$. The constants used in the model are indicated in the caption of figure 1.

This model simulates with a reasonable accuracy the main characteristics of the auditory periphery and is particularly adequate for the recognition task we have considered.

II. THE NOISE SUPPRESSION TECHNIQUE

A. *Speech enhancement*

Several solutions for the problem of speech recognition in noise have been proposed [7]. They can be classified in two categories: speech enhancement and model compensation. Speech enhancement techniques try to remove the noise from the speech signal while model compensation changes the recognition model parameters to accommodate noisy speech. In this paper we are interested in the first category in order to obtain an auditory representation that is not much affected by the noise.

One of the most used speech enhancement methods for speech recognition is spectral subtraction. The basic idea is to subtract the estimated noise power from the noisy signal power spectrum under the assumption that noise and speech are uncorrelated and additive. While this method could be applied in our model to the filter-bank output, we can use the fact that the HWR stage in the model implicitly implements a noise masking or a masking level which corresponds to the auditory threshold. This is closely related to the noise floor proposed by Klatt [8,9]. Instead of subtracting the noise power in the filter-bank output we can make a subtraction of the RMS envelope in a way that the masking level is almost unchanged with different SNRs.

This method is closely related to the center clipping used in [5] and leads to a similar formulation. However, in this case we could not assume a Gaussian pdf for the $Y_i[m]$ values because we are dealing with RMS energies and not isolated samples, $y_i[n]$.

The noise suppression method is then based on equation (2) or (3), resulting in the following form:

$$V_i[m] = \max\left(Y_i[m] - \Delta, A\right), \qquad (4)$$

or

$$V_i[m] = A + \max\left(Y_i[m] - \Delta, 0\right) \qquad (5)$$

where Δ corresponds to a function of the RMS noise estimate. This noise estimation is done during pauses in speech, supposing that noise characteristics change slowly in time. This is a reasonable assumption in most situations. The goal in this method is to have an almost constant mean of $V_i[m]$ for speech pauses, i.e., for the noise alone. To do this the statistics of $Y_i[m]$ must be known, in particular its probability density function (pdf).

B. *Filter-bank output energy statistics*

In order to estimate the statistics of the RMS envelope of the filter-bank outputs in noise, we assumed a Gaussian density for the noise. To turn the things clear we drop the indexes m and i in the previous notation and define a frame vector $\mathbf{x}$ which corresponds to the outcome of N random variables, $\mathbf{x}_n$, normal, independent and identically distributed (iid) with zero mean and variance $\boldsymbol{s}^2$. This is commonly expressed as: $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{s}^2\mathbf{I})$, where $\mathbf{I}$ is the NxN identity matrix. The filter output frame, $\mathbf{y}$, corresponds to a linear transformation of $\mathbf{x}$, $\mathbf{y}$=$\mathbf{Ax}$. The frame energy is the quadratic form, $P = \frac{1}{N}\mathbf{y}^T\mathbf{y}$, while the RMS energy is $Y = \sqrt{P}$. It can be shown that the characteristic function associated to the pdf of P (or the moment generating function) is

$$\Phi_P(s) = E\left\{e^{sP}\right\} = \left(\prod_{k=0}^{N-1}\left(1 - 2\boldsymbol{s}^2 s\,\boldsymbol{l}_k\right)\right)^{-\frac{1}{2}} \qquad (6)$$

where $\boldsymbol{l}_k$ are the eigenvalues of $\mathbf{A}^T\mathbf{A}$. For a rectangular window these eigenvalues are simply $\boldsymbol{l}_k$ =$|H(k)|^2$ where $H(k)$ is the frequency response of the considered filter. Unfortunately, even in this case the corresponding pdf does not lead to a simple expression. However, we have found that a good match of the pdf of Y is obtained with the gamma density function,

$$f_Y(Y) = \frac{c^b}{\Gamma(b)} Y^{b-1} e^{-cY} u(Y)\,\cdot \qquad (7)$$

This pdf has mean *b/c* and variance *b/c*^{2}. Equating these values with the mean and variance of *Y* a very good fit is obtained as shown in figure 2.

The mean and variance of *P* can be also found. If *w*[*n*] is the window applied to the frames, these values are:

$$\mu_P = \frac{\sigma^2}{N}\mathrm{tr}(\mathbf{A}) = \frac{\sigma^2}{N^2}\sum_{k=1}^{N}\left|H(k)\right|^2 \sum_{n=1}^{N} w^2[n] \qquad (8)$$

and

$$\sigma_P^2 = \frac{2\sigma^4}{N^2}\mathrm{tr}(\mathbf{A}^2) = \frac{2\sigma^4}{N^4}\sum_{k=0}^{N-1}\left|H(k)\right|^2\left(\left|H(k)\right|^2 \,\widehat{*}\, \left|W_2(k)\right|^2\right) \qquad (9)$$

where the symbol $\widehat{*}$ represents circular convolution and $W_2(k)$ is the DFT of $w^2[n]$. The mean and variance of *Y* can be approximated by the following expressions:

$$\mu_Y \cong \sqrt{\mu_P}\left(1-\frac{\sigma_P^2}{8\mu_P^2}\right) \qquad \sigma_Y^2 \cong \frac{\sigma_P^2}{4\mu_P} \qquad (10)$$

These approximations turn out to be very accurate and permit us to define the gamma pdf constants *b* and *c* in (5) for every channel, normalized to the standard deviation, σ.

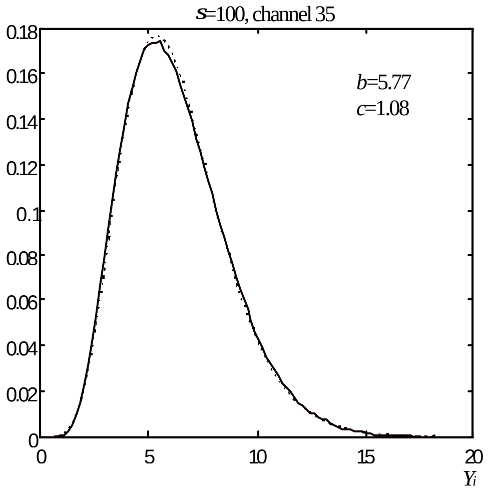

Fig. 2 Normalized histogram of Y_i for a channel *i* of the filter-bank (solid curve), and the gamma density function superimposed (dashed curve).

C. *Implementation*

Knowing the pdf of $Y_i[m]$ the mean of $V_i[m]$, $E\{V_i[m]\}$, can be expressed as a function of Δ and σ, as well as the ratio $\Delta/E\{Y_i[m]\}$ (or Δ/σ because the mean of $Y_i[m]$ is proportional to σ) [12]. In figure 3 we show the ratio $\Delta/E\{Y_i[m]\}$ as a function of $E\{Y_i[m]\}$ which is a very non-linear function of the mean of $Y_i[m]$. Then, from an observed mean value of $Y_i[m]$ in a previous segment of a speech pause, a new value for Δ is computed multiplying the mean of $Y_i[m]$ by this ratio. In this way the value of Δ changes adaptively. This method implies an accurate speech/non-speech detector which is a drawback in this kind of speech enhancing methods.

In order to estimate the mean of $Y_i[m]$ we can use the output of the adaptation low-pass filter (LPF) which has unit gain at DC, obtaining a relatively smooth estimate of the mean at no extra cost. Furthermore, we can use the method proposed in [5] which integrates the speech/pause detector within the model by monitoring the minimum value of the LPF output with a window large enough to ensure that at least one speech pause exists in the utterance.

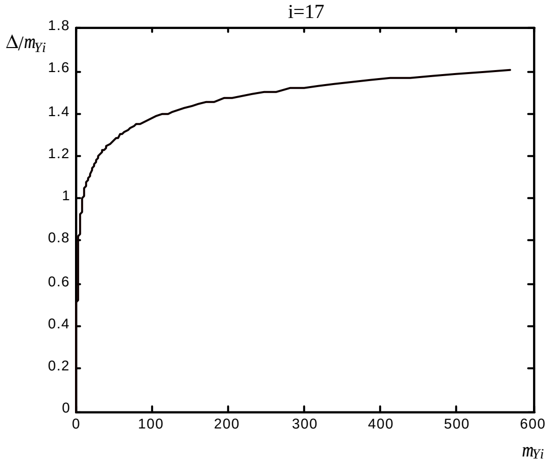

Fig. 3 Ratio $\Delta/E\{Y_i[m]\}$as a function of $E\{Y_i[m]\}$ for a constant value of $E\{V_i[m]\}=A+1$.

III. SPEECH RECOGNITION EXPERIMENTS

A. *Method*

In order to evaluate the proposed noise suppression technique as well as the performance of the auditory model in a speech recognition task, several recognition experiments were carried out. We used a subset, corresponding to isolated digits, of a Portuguese telephone speech database collected all over the country (the TELEFALA database). 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 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.

The observation vectors were generated by performing a DCT of the 35 coefficient mean rate vector leading to a 13 coefficient vector, including the first DCT coefficient as energy coefficient. This is a common practice in order to use diagonal covariance matrices in HMMs.

For noisy conditions we added artificial white noise to the speech signals with a signal-to-noise ratio of 20dB. The signal power was computed by averaging frame levels with more than 30 dB. Experiments were also conducted with speech shaped by a linear filter combined with the additive noise.

Because the digit utterances have short pauses, we used a fixed Δ taken according to the minimum value of the LPF output, corrected with the step-response of the filter.

B. *Results*

In table 1 we show several results with the present model. Three situations are indicated: clean speech, speech in noise and filtered speech in noise. For comparison, the first row refers to J-Rasta results. The second and third rows are results for the present model. As we can see, while the model is good in the situation of clean speech, the recognition performance degrades very substantially in noise conditions. This is a common result with auditory models as we reported in a previous paper [1]. The two HWR configurations, corresponding to equations (2) and (3), give almost the same results. As we referred earlier this is due to the adaptation in the model by the noise. A simple subtraction of the minimum value of the $Y_i[m]$ in the whole digit utterance gives much better results (row 4). However, this corresponds to an underestimate value of the noise level. As we would expect, a smoother estimate, obtained with the LPF output, is better (row 5). The results with the noise suppression technique described in section II are indicated in the last two rows, giving the better results in the situation of additive noise or noise plus filtering.

Table 1 - Digit recognition rates

Configuration	***Clean***	***Noise***	***Noise+Filt.***
J-RASTA	96.27	89.59	88.56
$V_i = A + Y_i$	96.82	67.04	45.03
$V_i = \max(Y_i, A)$	96.74	67.11	45.09
$V_i = \max(Y_i - Y_{\min}, A)$	96.76	90.38	78.30
$V_i = \max(Y_i - q_{\min}, A)$	96.89	93.40	90.43
$V_i = A + \max(Y_i - \Delta, 0)$	96.62	92.68	**93.93**
$V_i = \max(Y_i - \Delta, A)$	97.19	**94.45**	91.73

IV. CONCLUSIONS

We have presented an auditory model in which we applied a noise suppression technique closely related with spectral subtraction. In non-linear spectral subtraction the same principle is applied: remove less noise in high SNR segments and more noise in low SNR segments. While in this method the non-linear function weighting the subtraction process is chosen in an ad hoc form, the method presented here corresponds to an analytical solution. The presented method combines a constant noise floor with noise subtraction which seems to be a good practice in the framework of HMMs.

The presented results show that this method is effective in removing the noise conducting to much better recognition results. Some kind of noise normalization in the model is indeed needed as the model is very sensitive to the noise. Moreover, short-term adaptation could provide a better segmentation of speech signals into states in the HMM framework. This would lead to more impressive results in continuous speech recognition. Experiments with continuous digit recognition are scheduled to a near future.

The model computational complexity is comparable to the MFCC or RASTA front-ends. Furthermore, the noise suppression method described here can be applied to those front-ends with little modifications because they use a filter-bank and all that is needed is to perform a noise power (or level) subtraction at the output of these filters.

V. REFERENCES

[1] F. Perdigão, L. Sá, "Auditory Models as Front-Ends for Speech Recognition", *NATO ASI on Computational Hearing*, Il Ciocco, Italy, July, 1998.

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

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

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

[5] 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.

[6] 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.

[7] J-C. Junqua, J-P. Haton, "Robustness in Automatic Speech Recognition - Fundamentals and Applications", The Kluwer Int. Series in Eng. and Computer Science, Kluwer Academic Publishers, 1996.

[8] D. Klatt, "A Digital Filter Bank for Spectral Matching", ICASSP-76, pp. 573-576, 1976.

[9] B. Mellor, A. Varga, "Noise Masking in a Transform Domain", Proc. ICASSP-93, pp. II-87 - II-90, 1993.

[10] P. Lockwood, J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Speech Recognition in Cars", Speech Communication, 11 (2-3), pp. 215-228, June 1992.

[11] H. Hirsch C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition", ICASSP-95, pp. 153-156, 1995.

[12] F. Perdigão, "Modelos do Sistema Auditivo Periférico no Reconhecimento Automático de Fala", Ph.D. thesis, Univ. Coimbra, 1997.

ACKNOWLEDGMENT

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