# Analog Auditory Perception Model for Robust Speech Recognition

Yunbin Deng, Shantanu Chakrabartty and Gert Cauwenberghs
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD 21218
E-mail: {*yunbin,shantanu,gert*}*@jhu.edu*

***Abstract*— An auditory perception model for noise-robust speech feature extraction is presented. The model assumes continuous-time filtering and rectification, amenable to real-time, low-power analog VLSI implementation. A 3mm×3mm CMOS chip in 0.5$\mu m$ CMOS technology implements the general form of the model with digitally programmable filter parameters. Experiments on the TI-DIGIT database demonstrate consistent robustness of the new features to noise of various statistics, yielding significant improvements in digit recognition accuracy over models identically trained using Mel-scale frequency cepstral coefficient (MFCC) features.**

## I. Introduction

Despite the success of speech recognition systems for clean speech in controlled environments, their performance degrades severely when they are subjected to noise in natural environments [1]. One remedy to this problem has been to reduce the mismatch by training or retraining the recognition system under noisy conditions representative of the application environment. Elaborate techniques to reduce this mismatch have been proposed in the literature [2], [3].

The motivation for auditory based spectral analysis [4], [5], [6] is to gain recognition performance in natural (noisy) listening environments from the understanding of how the human ear processes speech. Auditory based models tend to elegant and efficient, physically based implementation in analog VLSI parallel hardware [7], [8], [9].

Rather than attempting to model exact physiological detail, our goal in silicon implementation of auditory perception for robust speech recognition is to focus on effective signal processing in the human ear. A simple model abstraction of the auditory system is proposed, that leads to efficient implementation in analog VLSI while offering sufficient flexibility in tuning recognition performance by adjusting system parameters in the architecture. The architecture comprises analog continuous-time filters with digitally programmable coefficients, integrated on a single VLSI chip.

Section II introduces the auditory perception model, and its architecture. Analog circuit implementation of the front end of the model is described and characterized in Section III. Section IV presents results from speech recognition experiments using the model, and Section V provides concluding remarks, discussions and future directions.

## II. Auditory perception model

The key to noise robustness in auditory perception is a scheme to filter out noise while retaining components of the signal characteristic of speech. An auditory perceptual model was proposed in [4], that uses adaptive dynamic compression of spectral features as an alternative to static nonlinear (logarithmic) compression as used in MFCC. Significant reduction in recognition error was observed in comparison with a recognizer trained using MFCC features.

The proposed auditory perception model is based on [4] and differs mainly in the functional form of the compression stage. The compression stage of [4] is composed of five consecutive nonlinear adaptation loops, where each adaptation loop consists of a divider and lowpass filter.

The model comprises four stages: a pre-emphasis stage; a constant-$Q$ bandpass filterbank; a rectification stage; and a bandpass compression stage which abstracts the nonlinear adaptive compression stage of [4]. The signal flow is shown in Figure 1.

The pre-emphasis stage is based on the observation that the human ear is insensitive to signals of frequency lower than 50Hz. This stage is implemented by an analog second-order Butterworth highpass filter.

The 24-channel bandpass filterbank stage is composed of second-order constant-$Q$ biquad filters as an approximation to the frequency response of the human ear basilar membrane. The transfer function of each of the filters is given by:

$$H_i(s) = \frac{G\frac{\omega_i}{Q}s}{s^2 + \frac{\omega_i}{Q}s + \omega_i^2} \qquad i = 1, \ldots 24 \qquad (1)$$

with center frequency $\omega_i$, and with constant quality factor $Q$. Constant $Q$ implies that the filter bandwidth scales with the center frequency, $Q = \omega_i/BW_i$, to trade resolution in time and frequency in a wavelet-like manner. To simulate the Mel/Bark scale frequency sensitivity of the human ear, the center frequencies of the filters are chosen according to the critical band distribution [10]. The frequency response of the filterbank is shown in Figure 2.

Full wave rectification extracts amplitude magnitude envelope information in each frequency band, modeling the response of hair cells transduction in the human auditory system. The rectified output is lowpass filtered with a cutoff frequency at 1 kHz.

Two versions of the compression stage are considered in the model, shown in Figure 1: one linear, and one with an additional log static non-linearity prior to the bandpass filtering. The log compression serves a purpose similar to that in MFCC, providing outputs that are less sensitive to spectral shaping due to the acoustics of the environment (reverberation, resonances etc.)

The absence of log compression in the linear version further simplifies the implementation in Figure 1 (b), by obviating the need for the lowpass stage since it is subsumed by the subsequent low-frequency bandpass stage. Similarly, the highpass pre-emphasis stage can be eliminated given it is subsumed by the subsequent high-frequency bandpass stage.

The critical component for noise robustness is the bandpass filterbank in the compression stage, implemented with low-$Q$ second-order biquad filters. The filterbank offers the functionality of the nonlinear adaptive compression model in [4], but simplifies the implementation.

The center frequencies of the compression stage filterbank are linearly spaced between 10 Hz and 24 Hz. This range of frequencies matches the filter characteristics observed from the adaptive compression model of [4], and corresponds to typical spectra of the amplitude envelope of human speech.

The resulting model is similar to RASTA processing of speech [6], in that it performs critical band analysis followed by low-frequency bandpass filtering. RASTA includes log static compression of the amplitude spectra prior to bandpass filtering. Hence, the noise robustness of the log-compressed version of the present model can be expected to be comparable to that of RASTA. However, the linear version is easier to implement in VLSI, and gives comparable results as shown in the experiments section below.

## III. Analog VLSI implementation

To illustrate the feasibility of implementing the model in massively parallel VLSI hardware, a general-purpose analog front-end chip is presented.

The $0.5\mu m$ CMOS chip contains 32 programmable analog continuous-time filter channels, which can be configured in parallel or cascade filterbank topologies. Each channel includes two bandpass or lowpass second-order filters, a full-wave rectifier, and a first-order lowpass filter. Hence it offers the full functionality of the simplified linear model of Figure 1 (b). The center frequency, bandwidth and gain of all filters are individually digitally programmable. As in [11], Level-crossing event detection is also provided to implement other auditory front-end models, *e.g.*, [5].

Advantages of the analog implementation include high integration density and low power consumption in comparison with digital implementation, and the continuous-time signal representation which avoids the need to sample and de-alias the speech signal. Continuous-time analog filters are conveniently implemented using transconductance amplifiers and capacitors, and achieve higher energetic efficiency than switched-capacitor filters which require excess bandwidth and thus higher power in the amplifiers.

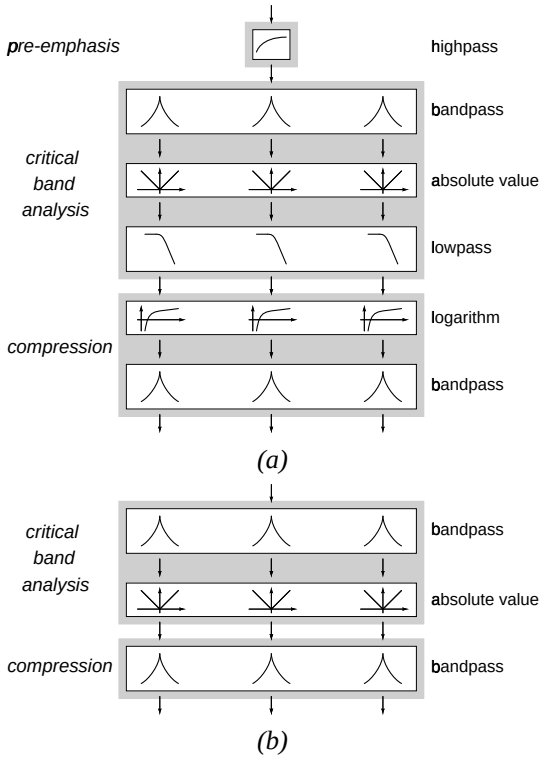


Fig. 1. Signal flow of logarithmic *(a)* and simplified linear *(b)* auditory perception model

A fully differential signal representation is adopted in the implementation, carrying the signal along with its complement throughout all stages of the architecture. While the fully differential signal format implies a doubling of circuit complexity in the implementation, it offers several advantages in maintaining high precision over single-ended implementation: doubling of the signal-to-noise ratio (SNR); high immunity to power supply noise; and elimination of all even-order distortion products (harmonics).

A second-order OTA-C (operational transconductance amplifier and capacitor) biquad filter topology is shown in Figure 3. It serves both as bandpass and lowpass filter, by selecting either of two outputs $Vbp$ or $Vlp$. The transfer function of the bandpass filter is given by

$$H_{bp}(s) = \frac{s\frac{G_1}{C_1}}{s^2 + s\frac{G_2}{C_1} + \frac{G_3 G_4}{C_1 C2}}, \qquad (2)$$

where $G_1$, $G_2$, $G_3$ and $G_4$ are individually programmable OTA transconductance values. The variable transconductance is obtained through current scaling techniques, as described in [12]. Capacitors $C_1$, $C_2$ are integrated on chip, with digitally selectable capacitance values. The filter parameters in (1) are determined by the programmed transconductance

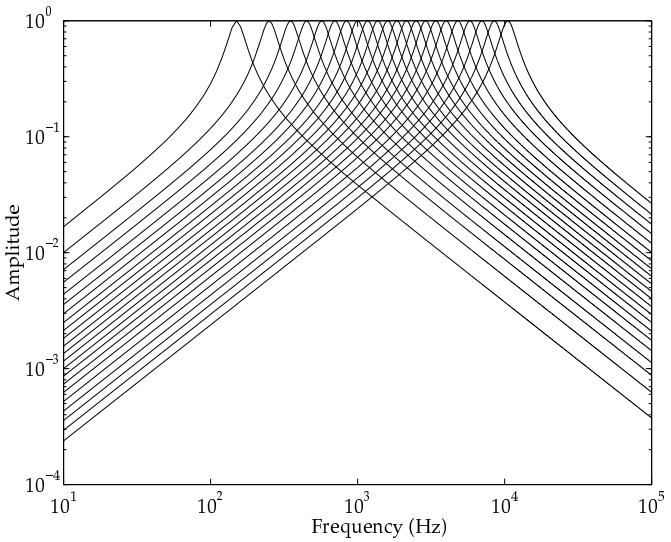


Fig. 2. Frequency response of the constant-$Q$ bandpass filterbank

and capacitance values through the relations

$$\begin{aligned} \omega &= \sqrt{\frac{G_3 G_4}{C_1 C_2}} \\ Q &= \sqrt{\frac{C_1 G_3 G_4}{C_2 G_2^2}} \\ G &= \frac{G_1}{G_2} \end{aligned} \tag{3}$$

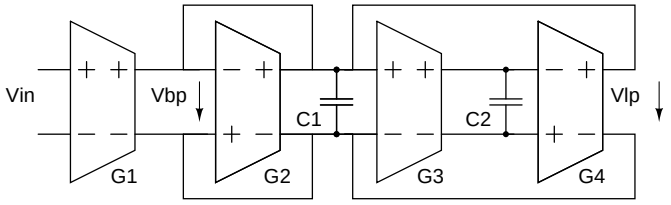


Fig. 3. Digitally programmable, constant-$Q$ OTA-C bandpass filter design

The differential signal representation provides for elegant implementation of the full-wave rectifier, using a single comparator and a cross-bar switch that either passes or inverts the polarity of the differential signal.

Although not implemented on the present chip, logarithmic compression in the model variant of Figure 1 (a) can be accomplished by linear OTA voltage-to-current conversion, followed by logarithmic current-to-voltage conversion on a diode, or a diode connected bipolar junction transistor or subthreshold MOS transistor.

The photomicrograph of the fabricated front-end processor is shown in Figure 4. The chip measures 3mm $\times$ 3mm in 0.5 $\mu$m CMOS technology. Table I summarizes the specification and measured performance figures of a single OTA. The frequency response of second-order bandpass filters programmed at different center frequencies is shown in Figure 5.

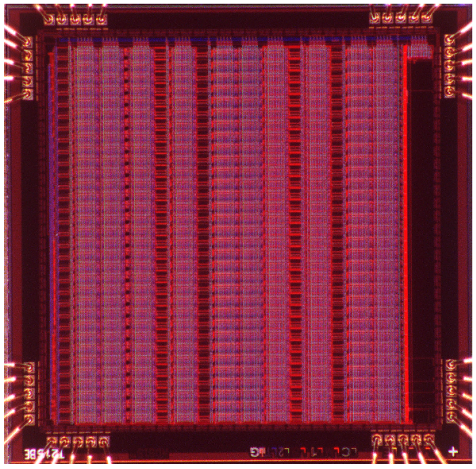

Fig. 4. Photomicrograph of programmable and reconfigurable OTA-C array.

Further details on circuits, measurements and characterization are presented in [12].

TABLE I
MEASURED OTA CHARACTERISTICS

| Parameter Specification | Measured |
|---|---|
| Max $Gm$ | 1.6 $\mu$A/V |
| Min $Gm$ | 0.78 nA/V |
| Programming ratio | 1/2048 |
| Input offset voltage | 20 mV |
| Max dynamic input range | $\pm$1.2 V |
| Common mode input voltage range | 0.5-3 V |
| Common mode output voltage range | 1.0-4.0 V |
| Common mode rejection ratio | 40 dB |
| Silicon area | 0.014 mm$^2$ |
| Power supply | 5 V |
| Power consumption | 12 $\mu$W |

## IV. Recognition Experiments with Auditory Perception Model

To compare the recognition performance of the proposed auditory perception model with that obtained using MFCC features, we chose as benchmark the standard TI-DIGIT isolated digit recognition dataset, with a vocabulary size of 11 (zero to nine plus 'O'). During recognition, the acoustic signal from the TI-DIGIT was subjected to additive noise from the NOISEX database, described further below.

Parameters of the simulated auditory model are as follows. The quality factor of the second stage bandpass filters is $Q = 4$, and that of the compression stage $Q = 0.7$. Without loss of generality it has been assumed that the input signal is rescaled such that $x[n] \leq 1, \forall n$. After all stages of auditory front-end processing, the resulting data were down sampled to 100 Hz rate, and DCT (discrete cosine transfer) was applied. Only the first 12 DCT features are retained to reduce the feature dimension.

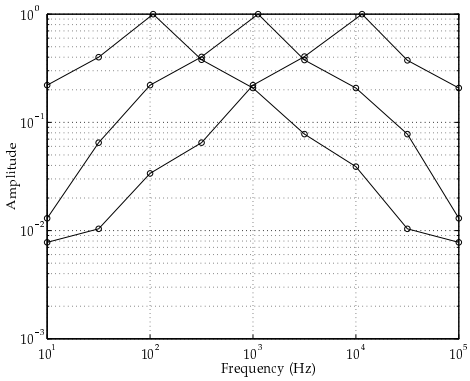


Fig. 5. Measured frequency response of second-order bandpass filter

For comparison, MFCC features were generated by applying a Hamming window of size 25 ms and overlap 15 ms to the same pre-emphasized 24-channel Mel-scale filterbank. The cepstral features were obtained from DCT of log-energy over the 24 frequency channels.

Figure 6 shows a sample comparison between auditory features and corresponding MFCC features for digit *five* obtained before DCT operation, at different SNR levels. The degradation of spectral features for MFCC in the presence of white noise is evident, whereas auditory features prevail at elevated noise levels.

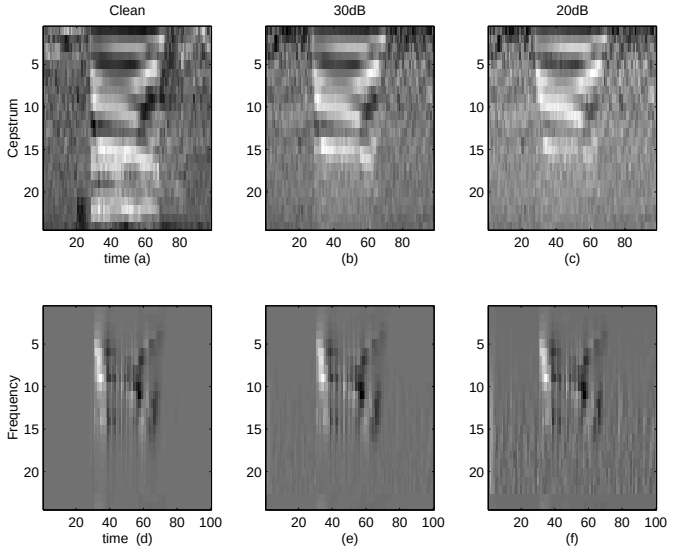


Fig. 6. *MFCC features (a)-(c) and linear auditory perception features (d)-(f) for digit* five *obtained before DCT, under white Gaussian noise conditions at different SNR (clean, 30 dB and 20 dB).*

As in [13], the training set contained two utterances of isolated digits each from 35 male speakers comprising a total of 770 utterances, and the test set contained isolated digits from 25 other male speakers for a total of 440 utterances. A recognition system was developed using the Hidden Markov Toolkit HTK, implementing a 14-state left-to-right transition model for each digit where the probability distribution on each state was modeled as a four-mixture Gaussian. Noise samples for the experiments were obtained from the $NOISEX$ database and were added to clean speech to obtain test data. We considered four types of noise common in application environments: white noise ($W$), speech babble noise ($B$), factory noise ($F$, plate-cutting and electrical welding equipment) and car interior noise ($C$, Volvo 340 at 75 mph under rainy conditions).

Table II summarizes the recognition rates obtained based on the two features under different noise statistics and under different SNR levels. Recognition rates are obtained with an identically trained HMM system for all models, and with speech subjected to additive car ($C$), babble ($B$), factory ($F$) and white Gaussian ($W$) noise, at various SNR levels.

TABLE II
RECOGNITION RATES FOR MFCC AND AUDITORY FEATURES

| | | Clean | 30dB | 20dB | 10dB |
|---|---|---|---|---|---|
| $C$ | MFCC | 98.8% | 98.6% | 98.1% | 96.8% |
| | Lin. Auditory | 98.6% | 98.6% | 98.6% | 98.6% |
| | Log Auditory | 98.8% | 98.6% | 98.8% | 98.6% |
| $B$ | MFCC | - | 97.2% | 93.8% | 60.7% |
| | Lin. Auditory | - | 98.6% | 97.9% | 74.1% |
| | Log Auditory | - | 98.4% | 93.2% | 72.5% |
| $F$ | MFCC | - | 95.9% | 67.7% | 28.6% |
| | Lin. Auditory | - | 98.6% | 93.4% | 44.3% |
| | Log Auditory | - | 97.7% | 91.8% | 61.6% |
| $W$ | MFCC | - | 81.1% | 27.5% | 12.2% |
| | Lin. Auditory | - | 98.6% | 92.3% | 27.7% |
| | Log Auditory | - | 98.4% | 95.7% | 72.7% |

The following can be inferred from the tabulated results:

1) For clean speech the performance of MFCC and auditory systems are comparable. For contaminated speech, the auditory models show superior performance for all noise types at all SNR levels, but one entry in the table (log auditory at 20 dB SNR).
2) For all noise types at low noise level (30 dB SNR), the auditory models do not degrade in performance.
3) For car noise, the auditory models maintain constant performance down to 10 dB SNR level.
4) For white noise and factory noise, auditory features demonstrate significantly better performance than MFCC. The improvement is most significant with the logarithmic auditory features.
5) For babble noise, improvements by the auditory features are least significant, but still noticeable.

Since babble is essentially speech, further advances would require a means to separate multiple speech components in the signal, *e.g.*, using [14]. For other types of noise that overlap minimally with the spectral and temporal structure of speech, the improvements are significant.

A higher level of noise robustness (down to 0 dB SNR), at the expense of a higher complexity in implementation, can be achieved using features obtained by growth transformation in reproducing kernel Hilbert space [13].

## V. Conclusions

A functionally simple, auditory perception model for noise-robust speech recognition was presented. The model maps directly onto parallel VLSI hardware using analog continuous-time filters, for efficient low-power and real-time implementation. To illustrate the filter characteristics achievable from such analog implementation, a prototype CMOS chip was fabricated and analyzed experimentally. Simulations of the model using HMM recognizers trained on the TI-DIGIT database demonstrated the robustness of the auditory features to noise of different statistics, significantly outperforming MFCC features at elevated noise levels, down to 10 dB SNR.

## Acknowledgment

This work was supported by ONR N000149910612, and a grant from the Catalyst Foundation, New York. The chip was fabricated through the MOSIS foundry service.

## References

[1] Gong, Y., "Speech Recognition in Noisy Environments: A survey", *Speech Communication*, vol. 16, pp. 261-291, 1995.

[2] Vaseghi, S.V. and Milner, B.P, "Noise-Adaptive Hidden Markov Models Based on Wiener Filters", *Proc. European Conf. Speech Technology*, Berlin, 1993, Vol. II, pp. 1023-1026.

[3] Nadas, A., Nahamoo, D. and Picheny, M.A, "Speech Recognition Using Noise-Adaptive Prototypes", *IEEE Trans. Acoust. Speech Signal Process.* vol. 37 (10), pp. 1495-1502, 1989.

[4] J. Tchorz and B. Kollmeier, "A Model of Auditory Perception as Front End for Automatic Speech Recognition," *J. Acoust. Soc. Am.*, vol. 106, pp. 2040-, 1999.

[5] Ghitza, O., "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, ed. by S. Furui and M.M. Sondhi (Marcel Dekker, New York), Ch. 15, pp. 453-485.

[6] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Processing*, vol. 2 (4), Oct. 1994.

[7] W. Liu, A. Andreou, and M. Goldstein "Voiced-Speech Representation by an Analog Silicon Model of the Auditory Periphery," *IEEE Trans. Neural Networks*, vol. 3 (3), pp. 477-487, 1992.

[8] R.F. Lyon and C.A. Mead, "An Analog Electronic Cochlea," IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, pp. 1119-1134, July 1988.

[9] E. Fragniere, A. van Schaik and E.A. Vittoz, "An Analog VLSI Model of Active Cochlea," in *Neuromorphic Systems Engineering: Neural Networks in Silicon*, T.S. Lande, Ed., Kluwer Academic, 1998.

[10] E. Zwicker, G. Flottorp and S.S. Stevens, "Critical Bandwidth in Loudness Summation," *J. Acoust. Soc. Am.*, vol. 29, pp. 548-557, 1957.

[11] Kumar, N., Himmelbauer, W., Cauwenberghs, G., and Andreou, A.G., "An Analog VLSI Chip with Asynchronous Interface for Auditory Feature Extraction", *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Proc.*, vol. 45 (5), pp. 600-606, 1998.

[12] Y. Deng, S. Chakrabartty and G. Cauwenberghs, "Three Decade Digital Programmable OTA Design," *IEEE Int. Symp. Circuits and Systems (ISCAS'04)*, Vancouver Canada, May 23-26, 2004.

[13] S. Chakrabartty, Y. Deng and G. Cauwenberghs, "Robust Speech Feature Extraction by Growth Transformation in Reproducing Kernel Hilbert Space," *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP'2004)*, Montréal Canada, May 17-21, 2004.

[14] G. Cauwenberghs, "Monaural Separation of Independent Acoustical Components," *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'99)*, Orlando FL, vol. 5, pp. 62-65, 1999.