AN ANALOG VLSI ARCHITECTURE FOR AUDITORY BASED FEATURE EXTRACTION

Nagendra Kumar[1] *Wolfgang Himmelbauer* *Gert Cauwenberghs* *Andreas G. Andreou*

Center for Language and Speech Processing
Johns Hopkins University, Baltimore MD 21218, USA.
[1]kumar@cspjhu.ece.jhu.edu

ABSTRACT

We have developed a low power analog VLSI chip for real time signal processing motivated by the principles of human auditory system. A analog cochlear filter-bank (which is implemented on the chip) decomposes the input audio signal into several frequency bands that have almost equal bandwidth on a log scale. This step is thus similar to computing the wavelet transform. The chip then computes signal energies and zero crossing time intervals of frequency components in a cochlear filter bank. The chip is intended to work as a front-end of a speech recognition system. We include experimental results on a VLSI implementation of the auditory front-end. We present speech recognition result on the TI-DIGITS database obtained from computer simulations which model the functionality of the feature extraction VLSI hardware. We use Hidden Markov Models (HMM) in combination with Linear Discriminant Analysis (LDA) for the recognizer design.

1. INTRODUCTION

The performance of today's state of the art speech recognition systems, that mainly use LPC or cepstral speech features, markedly degrades under adverse operating conditions such as cafeteria noise low-bandwidth telephone channels. On the other hand, the human capability of understanding speech remains almost unaffected under such circumstances. It has been argued that part of this robustness can be attributed to signal representation in early stages of the auditory periphery [1, 2, 3]. Therefore, by implementing feature extraction algorithms based on physiological studies of the auditory periphery in humans and primate vertebrates, we expect to achieve significant improvement in noise robustness. Also, the auditory periphery consumes only a tiny fraction of the energy dissipated when the same algorithms are implemented in standard DSP hardware. Therefore, it is conceivable that dedicated hardware emulating the structure of the auditory periphery is more suitable for portable devices where low power dissipation is important.

This paper presents an analog VLSI [4] architecture for auditory based feature extraction, which is designed to serve as the front-end to a speech recognition system. The extracted features are the signal energies and zero crossing time intervals obtained on the frequency decomposed output channels in a cochlear filter bank [5, 6]. The combined information in these two signals does not cause any loss of information [7, 8], and it is physiologically plausible that this information is carried by the firing patterns in the auditory nerve. The extracted features are made available at the output of the chip through an asynchronous communication protocol similar to that in [9]. The developed system is intended as a versatile tool for use by neuroscientists for auditory modeling, as well as a demonstration vehicle towards a low-power, real-time, robust speech recognizer for portable applications.

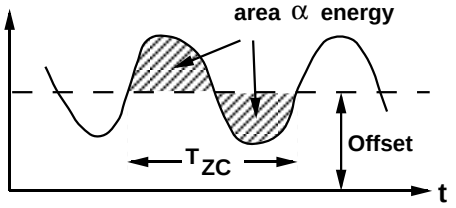

Figure 1. Information coding by zero-crossing intervals and period-energy.

The auditory processing and its VLSI implementation is described in Section 2. Some experimental results on the fabricated chip are presented in section 3. Section 4. presents recognition results on the TI-DIGITS database obtained from a software simulation using linear discriminant analysis to improve the auditory representation [10], and HMMs.

2. AUDITORY MODELING

The representation of transient signals by zero-crossings of their wavelet transforms has been investigated by S. Mallat. He pointed out that this representation is well adapted for solving pattern recognition problems [8]. From a physiological standpoint, the discrete action potentials generated by the inner hair cell in response to auditory stimuli can be considered as zero crossing events as well [11]. Moreover, zero-crossing intervals contain much information about the dominant frequency in the signal [12].

The implemented system emulating the auditory periphery includes a model of frequency decomposition in the basilar membrane of the cochlea, and a model of feature extraction (through zero crossings) in the inner hair cells of the cochlea. Subsequent processing in the auditory periphery is still unexplored terrain for physiological research, and in our implementation is left to be performed off-chip for optimal flexibility.

In our implementation, the basilar membrane consists of a bank of bandpass-shaped filters, with center frequencies spaced uniformly on a logarithmic scale from 100Hz to 8000Hz [13, 6]. The features extracted by the inner-hair cell circuitry are zero-crossing time intervals [14] and signal energies, for all of the basilar membrane filter outputs.

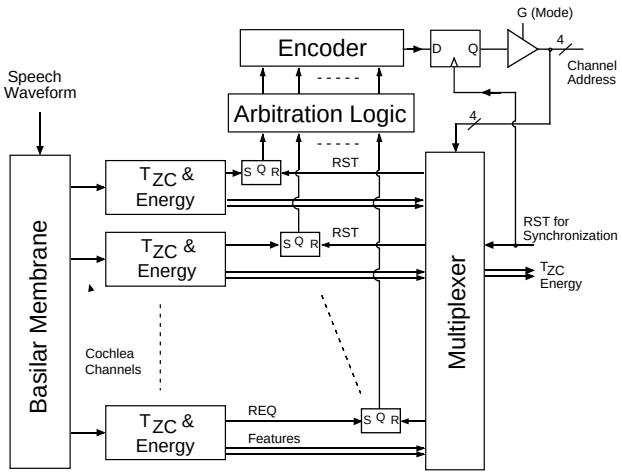

Figure 2. Block diagram of the VLSI architecture for the electronic cochlea.

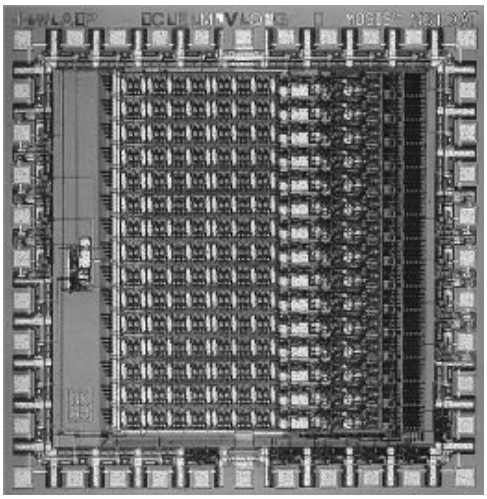

Figure 3. Analog VLSI chip implementing 15 channels of level-crossing and intensity auditory feature extraction, including channel encoding.

We need the energy feature to account for signal intensity which in the auditory system is encoded by the number of fibers firing.

Figure 1 illustrates the features extracted from the basilar membrane output (cochlea channels). We define T_{ZC} as the time interval between two consecutive upward zero crossing of the AC component of the signal and the 'energy' as the integral over the rectified AC component of the signal within the period T_{ZC}.

2.1. VLSI Architecture

Figure 2 shows a block diagram of the analog VLSI chip. The basilar membrane is implemented as a filter-bank structure, each segment of which consists of multiple linear first order sections followed by two linear bandpass filter sections [6]. The frequency decomposed time signals from the basilar membrane are processed locally. We employ a binary charge pump [15], to establish an adaptive elimination of signal offsets and inherent $1/f$ noise, through high pass filtering with channel specific time constants. A comparator detects upward zero crossing and provides control signals for circuitry computing T_{ZC}. The energy feature is obtained from integrating the full-wave rectification of the signal on a capacitor.

2.2. Asynchronous data acquisition

The outputs from every channel are time-division-multiplexed to the chip output, using an *asynchronous* protocol which is most efficient when dealing with communication problems involving a bandwidth-limited bus, and bus requests at arbitrary time and rate. At every zero crossing instant, both the time interval and energy feature are sampled and held, and a service is requested by setting the SR-latch. The arbitration logic handles multiple requests at a time, and shall favor the highest-frequency channel. The address of the winning channel is encoded and passed to a delay flip-flop (D-FF). The D-FF stores the address of the channel currently being accessed. This address controls the multiplexer that routes the channel features to the output pads of the chip. Once the outputs are sampled and held, an external reset (RST) pulse is expected. The RST pulse is multiplexed back to the SR latch of the channel just being completed. At the falling edge of the RST pulse, the new address available from the encoder is loaded into the D-FF.

The outputs available from the chip are represented by a list of events. Each event contains the address information, indicating where the event occurred, T_ZC, and the energy for that period.

Apart from the acquisition scheme just lined out, this architecture also allows for external multiplexer control. The buffer following the D-FF can be disabled and an address can be applied externally to access one particular channel at a time.

3. EXPERIMENTAL RESULTS

Figure 3 shows a micro-graph of the 2mm×2mm chip that has been fabricated through MOSIS using 1.2μm technology. Some experiments have been performed to verify chip functionality. Figure 4 demonstrates the time-interval feature computation. The lower trace is the BM output signal of the highest frequency segment in response to a triangular FM-modulated input signal in the audio range. The upper trace shows the resulting time-interval feature voltage. Its time discrete nature and 1/f (hyperbolic) characteristic can be observed easily. Also note that the magnitude envelope in the BM output signal in response to different signal frequencies.

The energy feature is extracted every period of the signal. If the period is held constant, then the amplitude modulation of the signal will reflect in this feature. Figure 5 illustrates this operation. The BM is supplied with an AM modulated sinusoidal of constant frequency (upper most trace). The trace below is the corresponding BM output. The lower two traces are the energy feature and the time-interval feature, respectively. We observe that signal energy changes, but due to constant frequency the time-interval feature remains constant.

The asynchronous communication scheme has also been tested and found functional.

4. DIGIT RECOGNITION EXPERIMENTS

Practical statistical recognition systems work best with compact signal representations that contain only the infor-

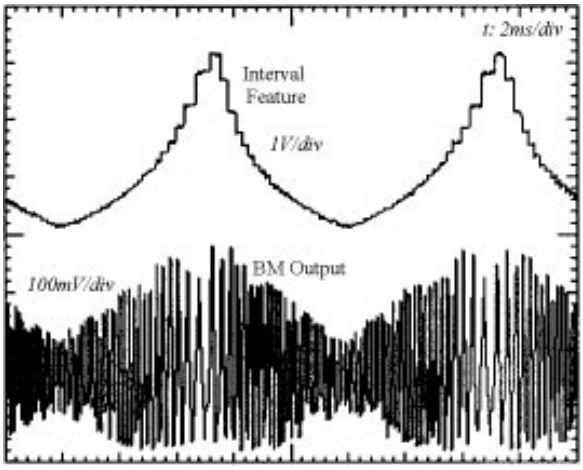

Figure 4. Interval Feature for FM-modulated Input.

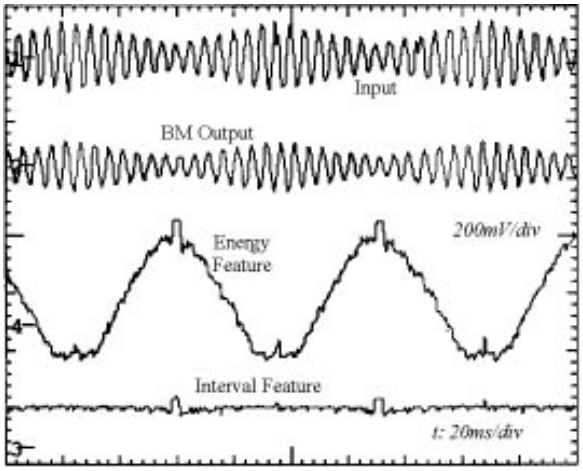

Figure 5. Energy and Interval Feature for AM-modulated Input.

mation that is relevant for the recognition task. We have used the Interval Histogram (IH) as feature vector. An IH is generated by creating several bins corresponding to different zero-crossing intervals T_{ZC}. For any zero crossing event, the appropriate bin is chosen depending on the value of the event's T_{ZC}, and filled with the non-linearly compressed event energy. For the current implementation, we use cubic root compression. The IH is computed from at most last twenty crossing events, at a rate of 100 Hz. In addition, events that have occurred more than forty milliseconds prior to IH generation are discarded. This approach is similar to that of computing EIH [1]. As an alternative, we may choose the IH generation to be triggered every upward zero crossing in the lowest frequency channel. The motivation behind this approach is that in a realistic biological system, there is no reason to believe that feature computation occurs at uniform intervals. However it seems plausible that some low frequency events may drive the computation necessary for recognition. One such possible event is the zero crossings in low frequency channel, that roughly correspond to the pitch period of the voiced speech signal. The

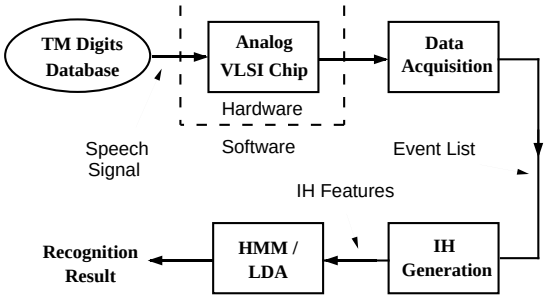

Figure 6. System architecture for the use of silicon cochlea as a pre-processor to a speech recognition

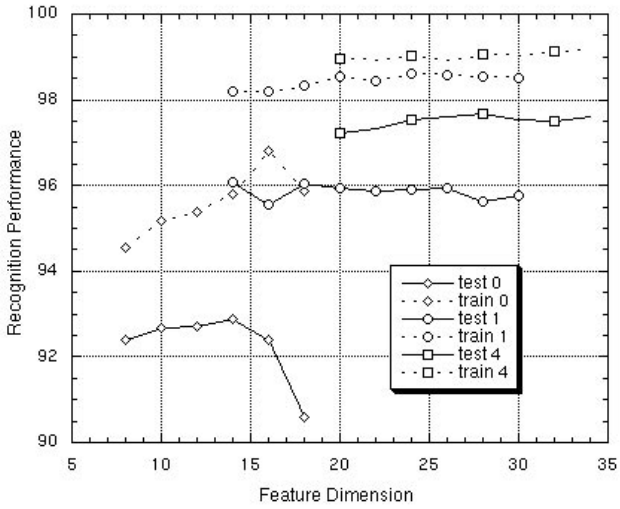

Figure 7. Recognition performance with 20 bin IH features computed every 10 milliseconds

bins are of equal width on a log-frequency scale, and are normalized by dividing the energy in each of them by the total sum of all the energy. The energy is then scaled adequately and added as a separate feature. Figure 6 depicts the intended recognition system architecture. The analog VLSI chip serves as the front end. The acquisition system collects zero crossing intervals as well as the corresponding energy measure from all channels. The events are stored in turn to form an event list. Subsequently, a software module computes interval histograms, which are fed into a statistical recognizer. For software simulations, the analog VLSI chip and the acquisition system has been replaced by an equivalent software module. Digit recognition experiments were performed according to Figure 6. A five state single mixture HMM is trained for every digit. Suppose the feature is n dimensional. At any frame, the C previous frames, and the C following frames are concatenated to form a new (2C+1)n dimensional feature. C is defined as the context size. Linear Discriminant Analysis is applied to the expanded features [16, 17, 10, 18] to reduce the feature dimensionality.

Figure 7 shows the results of experiments with the IH features computed every 10 milliseconds. Results are shown for both training and the test data. The context size C is indicated in the legend. Recognition accuracy is plotted against the reduced feature dimension. Our objective here

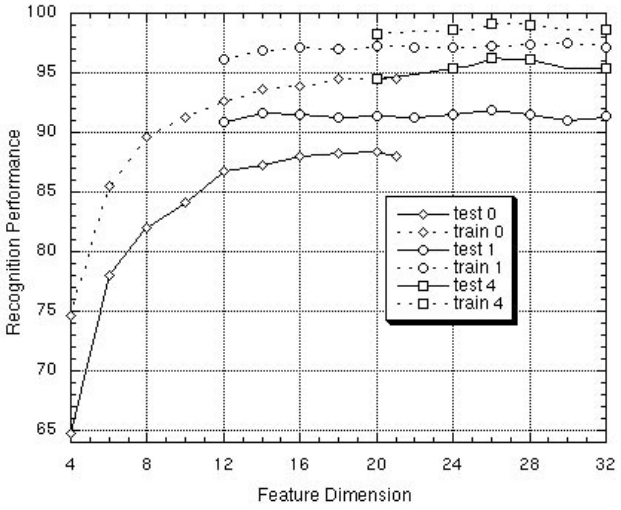

Figure 8. Recognition performance with 20 bin IH features computed at every zero crossing of the lowest frequency channel

is to develop a silicon hardware that can serve as a tools for research in auditory representations. Once the VLSI hardware is developed, the representation choice would also have to take into account the robustness of the representation in presence of device mismatch and temperature variations. Given the fact that not much effort was put into choosing the right representation to suite the Gaussian models, the performance appears to be reasonable. When the number of mixtures is increased to four, the recognition performance improves further to 98.3%. Experiments with zero-crossing triggered IH generation (Figure 8) also indicate that further research and optimization may be helpful. These results should be treated as preliminary when compared to the state of the art systems [19]. We believe that the difference in performance is due to the fact that our features are not rich enough, and models are too simple to represent the data. The performance may potentially be further improved by applying the more general models of LDA [20, 21]. However, the recognizer performance certainly suggests the applicability of analog VLSI cochlea to auditory based research.

5. CONCLUSION

We have proposed an approach to real-time auditory based speech recognition using analog VLSI as a front-end feature extractor. A chip has been designed and the algorithm was tested on the TI-DIGITS spoken digit database. Simulation results show that one can obtain encouraging recognition results while keeping model complexity low.

REFERENCES

[1] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy enviornment," *Computer Speech and Language*, vol. 1, pp. 109–130, 1986.

[2] C. R. Jankowski Jr., "A comparison of auditory models for automatic speech recognition," Master's thesis, M.I.T., May 1992.

[3] C. V. Neti, "An auditory model for speech recognition in noise: Some recognition experiments," Tech. Rep. TR54.881, IBM, 1994.

[4] C. A. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.

[5] R. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 36, pp. 1119–1134, July 1988.

[6] P. M. Furth and A. G. Andreou, "Cochlear models implemented with linearized transconductors," in *International Symposium on Circuits and Systems*, vol. 3, pp. 491–494, IEEE, 1996.

[7] B. F. Logan, Jr., "Information in the zero crossings of bandpass signals," *The Bell System technical Journal*, vol. 56, pp. 487–510, April 1977.

[8] S. Mallat, "Zero-crossings of a wavelet transform," *IEEE Transactions on Information theory*, vol. 37, no. 4, 1991.

[9] J. Lazzaro, J. Wawrzynek, M. Mohwald, M. Sivilotti, and D. Gillespie, "Silicon auditory processors as computer peripherals," *IEEE Transactions on Neural Networks*, vol. 4, pp. 523–528, May 1993.

[10] N. Kumar, C. Neti, and A. G. Andreou, "Application of discriminant analysis to speech recognition with auditory features," in *Proceedings of the Fifteenth Annual Speech Research Symposium*, (Johns Hopkins University, Baltimore, MD 21218), pp. 153–160, June 1995.

[11] E. Young and M. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.*, vol. 66, pp. 1381–1403, 1979.

[12] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, pp. 1477–1493, Nov. 1986.

[13] W. Liu, A. G. Andreou, and M. G. Goldstein, "Voiced speech representation by an analog silicon model of the auditory periphery," *IEEE Transactions on Neural Networks*, vol. 3, pp. 477–487, May 1992.

[14] N. Kumar, G. Cauwenberghs, and A. G. Andreou, "A circuit model of hair-cell transduction for asynchronous analog auditory feature extraction," in *International Symposium on Circuits and Systems*, vol. 3, pp. 301–304, IEEE, 1996.

[15] G. Cauwenberghs and A. Yariv, "Fault-tolerant dynamic multi-level storage in analog VLSI," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 41, no. 12, pp. 827–829, 1994.

[16] X. Aubert, R. Haeb-Umbach, and H. Ney, "Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models," in *Proc. of ICASSP*, vol. 2, pp. 648–651, 1993.

[17] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. of ICASSP*, vol. 1, pp. 13–16, 1992.

[18] P. F. Brown, *The Acoustic-Modelling Problem in Automatic Speech Recognition*. PhD thesis, Carnegie Mellon University, 1987.

[19] R. Haeb-Umbach, D. Geller, and H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," in *Proc. of ICASSP*, vol. 2, pp. 239–242, 1993.

[20] N. Kumar and A. G. Andreou, "On generalizations of linear discriminant analysis," Tech. Rep. JHU/ECE-96-07, Johns Hopkins University, 1996.

[21] N. Kumar and A. G. Andreou, "A generalization of linear discriminant analysis in maximum likelihood framework," in *Proceedings of the Joint Statistical Meeting*, vol. Statistical Computing, ASA, August 1996.