EMPIRICAL COMPARISON OF ANALOG AND DIGITAL AUDITORY PREPROCESSING FOR AUTOMATIC SPEECH RECOGNITION

Todd M. Massengill, Denise M. Wilson

University of Washington
Department of Electrical Engineering
Seattle, Washington 98195-2500 USA

e-mail: toddmass@u.washington.edu; wilson@ee.washington.edu

Paul E. Hasler, David W. Graham

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, Georgia 30332-0250 USA

e-mail: paul.hasler@ee.gatech.edu; gte168s@prism.gatech.edu

ABSTRACT

Results from digital and analog filter bank preprocessors are compared in order to establish the validity of analog processing for automatic speech recognition (ASR) systems. Three systems are evaluated using speaker and context independent phoneme recognition tasks. The three ASR systems are identical except for the preprocessing techniques used to derive three signal representations: extraction of 1)the digital mel-frequency spectrum, 2)the mel-frequency spectrum from commercial discrete bandpass filters and 3)the exponential spectrum from an analog VLSI bandpass filter bank. The discrete analog system exhibits a 38% increase in recognition accuracy over the digital preprocessing technique. The digital and analog VLSI-based techniques perform comparably (within 3% of each other).

INTRODUCTION

Real-time, low power speech recognition has long been a difficult problem for researchers who are driven by the demand for low cost portable systems. Biology is frequently inspiration for a model of an ideal recognition system due to its parallel processing and high efficiency (defined as energy required per basic operation). Current neuromorphic analog systems have been shown to be four orders of magnitude more efficient than digital techniques but still five orders of magnitude less efficient than the brain.[1]

It is clear that in order for artificial auditory systems to approach the efficiency of the brain, traditional derivations of speech features will have to undergo a paradigm shift. One example of work directed at this sort of fundamental shift in speech processing is a silicon auditory model using periodicity-based spectral shape, pitch and temporal correlation.[2] Much research in this area approaches the recognition problem by adding features such as zero crossing intervals[3] and derivatives[4] resulting in heterogeneous feature vectors.

One successful application using computational models of auditory processing is a speaker-independent 13-word recognition task with 30 features. These 30 features are subsampled from an original 119 features that are derived by the real-time pre-processor.[2] Their results showed a 4.1% error rate. Another example of using a robust auditory signal representation for recognition is an analog VLSI chip tested on the TI-DIGITS speech database. An analog cochlear bandpass filterbank with constant bandwidth and exponentially increasing center frequencies is implemented on chip along with calculation of signal energies and zero crossing time intervals.[3] The features extracted from the chip result in a recognition accuracy of 98.3%.

The difficulty that most neuromorphic signal representations face is that current recognition engines have been designed to accommodate low dimension feature sets. The size of the training data set required for good generalization and the computational complexity of the training algorithms increase as a function of the number of features. The nature of the relationship between computational complexity and number of features is greater than O(N). Biological systems are able to meet the memory and computational demands of large numbers of features generated at high frequencies. However, current computing resources typically limit the number of features that can be processed in real-time applications to under 128 features every 25ms. Therefore, compression of the feature space to comply with computational limitations can negate some of the benefits of neuromorphic processing.

Meanwhile, digital signal processing (DSP) techniques have been established as the benchmark for speech recog-

nition tasks. An opportunity exists to shift the digital systems to low power highly parallel analog systems. This paper presents an empirical argument for the replacement of the preprocessing block in an otherwise digital ASR system with analog counterparts. Our results conclude that analog preprocessing techniques perform comparably or improve upon digital techniques.

SYSTEM DESCRIPTION

The task chosen for the ASR in this paper is context and speaker-independent isolated phoneme recognition. Usually considered the smallest unit of speech of interest for speech recognition, phonemes are desirable for portable applications because of their small dictionary size (39 to 60) that requires fewer models for effective recognition. Isolated phoneme recognition is a difficult task even for the auditory system. Other research has achieved a maximum of 65% accuracy for context-independent phoneme recognition.[5] Phoneme classification is useful for a first stage in recognition to narrow down the likelihood of possible phonemes. However, to date, speaker and context independent recognition has not been evaluated for ASR systems using analog preprocessing.

The block diagram of the ASR system is shown in Figure 1. Speech is the input to one of three possible spectrum blocks as described in more detail in the next section. The output of the spectrum block is then sampled and cepstral coefficients are calculated using log operations and discrete cosine transforms. Thirteen cepstral coefficients are used to train and test the hidden markov model (HMM) used as the pattern recognition model in this effort.

CIRCUIT DESCRIPTION

Digital Filtering

The frequency spectrum is obtained with three filtering methods. The first is a digital filter bank. The input is passed through a pre-emphasis high pass filter with a coefficient of 0.95; the resulting signal is windowed and its discrete fourier transform calculated. This result is mapped into 45 values by a series of mel-frequency spaced triangular-shaped weighting functions in the frequency domain[1]. Cepstral coefficients are then calculated from these 45 values in the same way as for the remaining 2 preprocessing techniques.

Discrete Analog Filtering

The second preprocessing technique is a filterbank of 45 discrete 2nd order (two-pole) bandpass filters (Texas Instru-

1. The combination of the triangular weighting functions with a fourier transform is not, strictly speaking, an implementation of digital filtering but the result is analogous. Due to the serial nature of digital filters, digital filter banks are not used when processing time is limited; the time required is prohibitive. In this paper, the term 'digital filtering' refers to the mapping of a fourier transform into the magnitude of energy contained in each of 45 frequency bands.

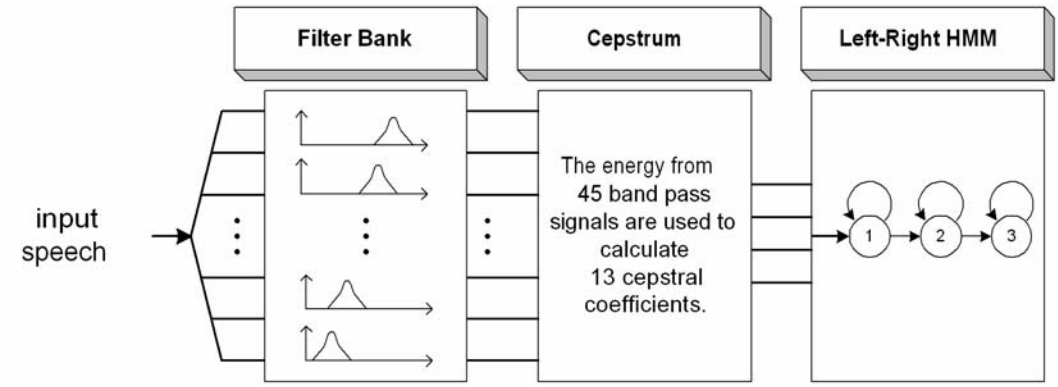


Figure 1: Block Diagram of the ASR System with Spectrum Preprocessing and Hidden Markov Model

The incoming auditory signal is first preprocessed to normalize it according to the total energy in the signal over a sliding window whose time constant is much less than the sampling rate. The normalized signal is then transferred to a highly parallel, distributed filter bank which extracts frequency information from the single input signal in a manner similar to the basilar membrane and hair cells in the ear. The preprocessing performed by this filter bank is done with one of three different methods and the results from the left-right HMM that performs the phoneme recognition from these preprocessed signals are compared.

ments UAF42AP Universal Active Filter[6]) tuned to a mel-frequency scale. Each UAF42AP consists of four op-amps and integrated capacitors and resistors. Three of the four op-amps can be wired in one of two similar configurations to implement biquad or state variable bandpass filters. The biquad configuration is used in the low frequency range (below 500 Hz) because the center frequency can be adjusted easily while keeping the bandwidth constant. The state variable filter (shown in Figure 2a) is used for the higher frequencies as its center frequency can be tuned while keeping the Q factor constant thereby enabling exponentially spaced filters. These two bandpass filter configurations are robust to component variation[7].

Analog VLSI Filtering

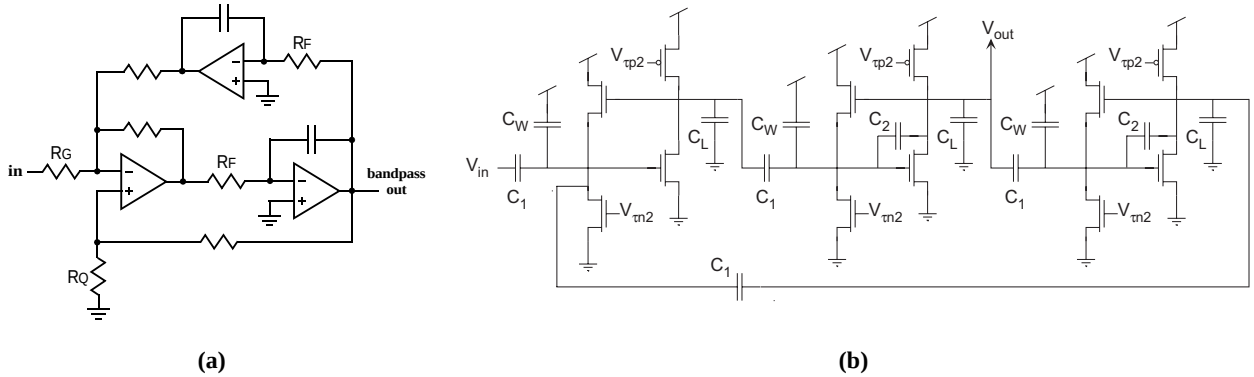
The third preprocessing technique consists of 32 filters exponentially spaced across the spectrum. The filters are fourth order (4-pole) bandpass filters consisting of capacitors and transistors. Figure 2b shows the filter structure. The frequencies are tuned by varying six gate voltages: two for a gain stage and the other four to tune the four poles. In this implementation, the gate of each filter is connected to its adjacent filter by a resistance. The voltage drop across the resistor results in an exponential shift in the frequencies of the poles. Thus, by adjusting the 12 gate voltages, (six at each end of the filter array) the top and bottom frequencies are set and the other 30 frequencies are spaced at regular intervals in log space, within the specified frequency range.

RESULTS AND DISCUSSION

The data set for these experiments is taken from the DARPA TIMIT continuous speech corpus.[8] Six voiced phonemes are used: ‘aa’ (as in w**a**sh), ‘ae’ (**a**sk), ‘aw’ (h**ow**), ‘er’ (t**er**m), ‘ow’ (g**o**) and ‘oy’ (**oi**l). A random selection of 60 recordings of each phoneme is assembled containing males and females and a representation of the eight major dialects within the United States delineated in the Timit speech corpus.

In order to remove any bias due to the small size of the data set, the training data is divided into four sets of 270 out of 360 total phonemes. For each training set, a testing set is used consisting of the 90 remaining phonemes. The four resulting data sets {training, testing} are: {1:270,271:360}, {91:360,1:90}, {181:90,91:180} and {271:180,181:270}. Four hidden markov models are trained and tested for each phoneme using each of the four data sets. The resulting confusion matrix is the average of the results from the four models thereby minimizing bias due to selection of the training and testing data.

The results of digital filtering are shown in Table 3 as a confusion matrix. Each row represents the recognition of the phoneme (designated in the row heading) by the model (designated in the column heading). It is not surprising that few phonemes are falsely recognized as ‘er’ while it is more likely that ‘ow’ and ‘oy’ will be mistaken for one another. The upper left corner shows the average accuracy for the digital filtering benchmark. Accuracy of 42.22% is not a state of the art recognition rate, a discrepancy that is



(a) **(b)**

Figure 2: Bandpass Filter Configurations

The configuration shown in (a) is the state variable discrete analog configuration of the UAF42AP. The two R_F resistors and the R_G and R_Q are external components. All other resistors and capacitors are integrated having values of 50 kOhm ±**0.5** % and 1000 ±**0.5** % pF respectively. The analog VLSI bandpass filter is shown in (b).

due to minimizing the number of triangular filters in order to match the number of discrete analog filters. The digital set serves as a benchmark with which to compare the impact of replacing digital calculations with analog equivalent circuits.

The results from the discrete analog filtering in Table 4 show a marked improvement in recognition accuracy at 58.33% with an uncharacteristically low 28.33% accuracy for 'aa'. The analog VLSI-based filtering results in Table 5 at 39.72% are less consistent. Lower accuracy is most likely attributable to the use of 32 filters rather than the 45 used with the other two preprocessing techniques. Despite the reduced size of the filter bank, analog VLSI filtering produces comparable results to the digital spectrum derivation. The number of filters is easily scalable and we expect to see improvements in these results as we approach the architecture of the digital and discrete analog preprocessing used in this work.

Table 3: Digital Mel-Frequency Confusion Matrix

42.22	**aa**	**ae**	**aw**	**er**	**ow**	**oy**
aa	**40**	10	18.33	1.67	20	10
ae	10	**55**	11.67	1.67	15	6.66
aw	8.33	21.67	**33.33**	1.67	16.67	18.33
er	10	13.33	6.67	**50**	13.33	6.67
ow	13.33	5	15	11.67	**31.67**	23.33
oy	10	1.67	10	1.67	33.33	**43.33**

Table 4: Discrete Filter Bank Confusion Matrix

58.33	**aa**	**ae**	**aw**	**er**	**ow**	**oy**
aa	**28.33**	6.67	38.33	5	18.33	3.33
ae	3.33	**56.67**	18.33	6.67	10	5
aw	11.67	5	**61.67**	1.66	11.67	8.33
er	3.33	11.67	5	**68.33**	10	1.67
ow	1.67	5	10	1.67	**75**	6.66
oy	5	6.67	0	10	18.33	**60**

Table 5: VLSI Filter Bank Confusion Matrix

39.72	**aa**	**ae**	**aw**	**er**	**ow**	**oy**
aa	**51.67**	36.67	1.66	10	0	0
ae	41.67	**36.67**	6.67	13.33	0	1.66
aw	45	31.67	**10**	13.33	0	0
er	16.67	21.67	3.33	**30**	5	23.33
ow	0	0	0	20	**21.67**	58.33
oy	0	0	0	5	6.67	**88.33**

ACKNOWLEDGEMENTS

The authors wish to acknowledge the National Science Foundation for partial support of this research (Award #ECS-9907464).

CONCLUSIONS

We have successfully demonstrated the ability of biologically inspired analog preprocessing to supplement digital processing efforts and traditional auditory recognition algorithms in a mixed signal implementation. Both analog preprocessing techniques demonstrate recognition comparable to digital techniques with the discrete analog technique producing a marked 38% improvement over digital. Our future work will focus on modifying the analog VLSI-based filtering to more accurately match the architecture of digital and discrete analog techniques.

REFERENCES

[1] Carver Mead, "Neuromorphic Electronic Systems", *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1629-1636, October 1990.

[2] J.Lazzaro and J. Wawrzynck, "Speech Recognition Experiments with Silicon Auditory Models", http://www.pcmp.caltech.edu/anaprose/lazzaro/recog.pdf.

[3] N. Kumar, W. Himmelbauer, G. Cauwenberghs, A.G. Andreou, "An Analog VLSI Architecture for Auditory Based Feature Extraction", *ICASSP*, vol. 5, 1997, pp. 4081-4084

[4] Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition", *Proc of the IEEE*, vol. 81, no. 9, pp. 1215-1247, September 1993.

[5] R. Chengalvarayan, L. Deng, "HMM-Based Speech Recognition Using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features", *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, pp. 243-256, May 1997.

[6] UAF42AP Product Folder, Texas Instruments, http://focus.ti.com/docs/prod/productfolder.jhtml?genericPartNumber=UAF42.

[7] P. Horowitz and W. Hill, The Art of Electronics, *Cambridge University Press*, 1989, pp. 277-278.

[8] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1, October 1990.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from incomplete data via the EM algorithm," *J. Roy Stat. Soc*., vol. 39, no. 1, pp. 1-38, 1977.

[10] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc of the IEEE*, vol. 77, No. 2, pp. 257-286, February 1989.