Speech Recognition Experiments with Silicon Auditory Models

John Lazzaro and John Wawrzynek
CS Division
UC Berkeley
Berkeley, CA 94720-1776
lazzaro@cs.berkeley.edu, johnw@cs.berkeley.edu

Abstract

We have developed a real-time system to transform an audio signal into several specialized representations of sound. The system uses analog circuit models of biological audition to compute these representations. We report on a speech recognizer that uses this system for feature extraction, and we evaluate the performance of this speech recognition system on a speaker-independent 13-word recognition task.

1. INTRODUCTION

Neurophysiologists and psychoacousticans have made fundamental advances in understanding biological audition. Computational models of auditory processing, which allow the quantitative assessment of proposed theories of auditory processing, play an important role in the advancement of auditory science.

In addition to serving a scientific function, computational models of audition may find practical application in engineering systems. Human performance in many auditory tasks still exceeds the performance of artificial systems, and the specific characteristics of biological auditory processing may play an important role in this difference. Current engineering applications of auditory models under study include speech recognition (Jackowoski *et al.*, 1995; Ghitza, 1998; Seneff, 1988), sound separation (Cooke *et al.*, 1994), and masking models for MPEG-audio encoding (Colomes *et al.*, 1995).

Computation time is a major limitation in the engineering application of auditory models. For example, the complete sound separation system described in (Brown and Cooke, 1994) operates at approximately 4000 times real time, running under UNIX on a Sun SPARCstation 1. For most engineering applications, auditory models must process input in real time; for many of these applications, an auditory model implementation also needs to be low-cost and low-power. Examples of these applications include robust pitch-tracking systems for musical instrument applications, and robust feature extraction for battery operated speech recognizers.

One implementation approach for auditory models in these products is to design low-power special-purpose digital signal processing systems, as described in (Chandrakasan and Brodersen, 1995). However, in many of these potential products, the input takes an analog form: a voltage signal from a microphone or a guitar pickup. For these applications, an alternative architecture is a special-purpose analog to digital converter, that computes auditory model representations directly on the analog signal before digitization.

Analog circuits that compute auditory representations have been implemented and characterized by several research groups – these working research prototypes include several generation of cochlear models (Lyon and Mead, 1988; Liu *et al.*, 1992; Watts *et al.* 1992; van Schaik *et al.*, 1995), periodicity models (Lazzaro and Mead, 1989a; Lyon, 1991), spectral-shape models (Lazzaro, 1991; van Schaik *et al.*, 1995), and binaural models (Lazzaro and Mead, 1989b; Bhadkamkar, 1994).

A prime benefit of these circuit structures is very low power consumption: the circuit techniques used in most of these prototypes were originally developed for wristwatch and pacemaker applications. For example, a recent publication on cochlear design techniques reports a 51-channel cochlear filterbank that consumes only 11 microwatts at 5 volts (Watts *et al.,* 1992). Voltage and process scaling, and advances in circuit design, could reduce power consumption even further.

If auditory models offer a performance advantage over standard signal processing techniques in an application, and a compact implementation that only consumes a few milliwatts of power is needed, a hybrid system that couples a special-purpose analog to digital converter with a low-power digital processor may be a competitive alternative to a full-digital implementation. However, even if auditory models only offer comparable performance to standard techniques for an application, an analog auditory model implementation may be the best choice for front-end processing, if the system requires microwatt operation (for example, size limitations dictate a lithium watch battery power source). For such micropower systems to become a reality, micropower implementations of pattern-recognition functions must also be available: a recent report on a nanopower neural-network recognition structure (Coggins *et al.,* 1995), used in an implantable cardiac morphology classification system, is an example of progress in this area.

Standard analog performance measurements (S/N ratio, dynamic range, ect.) aren't sufficient for determining the suitability of analog implementations of non-linear, multi-stage auditory models for a particular application. This paper documents a more direct approach to evaluating analog auditory models: we have integrated a multi-representation analog auditory model with a speech recognition system, and measured the performance of the system on a speaker-independent, telephone-quality 13-word recognition task.

The structure of the paper is as follows. We begin with a brief description of our multi-representation auditory model hardware implementation. We then describe in detail the specific auditory representations we use in our speech recognition experiments, and the techniques we use for generating a feature vector suitable for speech recognition systems. Next, we assess word recognition performance of the system, and compare the results with state-of-the-art feature extraction systems. The paper concludes with discussion and suggestions for further research.

2. SYSTEM DESCRIPTION

We have designed a special-purpose analog-to-digital converter chip, that performs several stages of auditory pre-processing in the analog domain before digitization (Lazzaro *et al.,* 1994; Lazzaro and Wawrzynek, 1995b). Configurable parameters control the behavior of each stage of signal processing. Figure 1 shows a block diagram of a system that uses three copies of this converter chip: by configuring each chip differently, the system produces three different auditory representations in response to an analog input.

This system acts as a real-time audio input device to a Sun workstation: a pre-amplified microphone input can be connected directly to the converters for a low-latency, real-time display of spontaneous speech. Alternatively, the system can receive analog input from the 8 Khz sampling rate, 8-bit mu-law audio output of the workstation, for controlled experiments: all experiments reported in this paper were done using this method of sound presentation. The dynamic range of the converter chip is 40 to 60 dB, depending on the signal processing configuration in use: input sensitivity is 1 mV (peak), and maximum recommended signal amplitude is 1 V (peak).

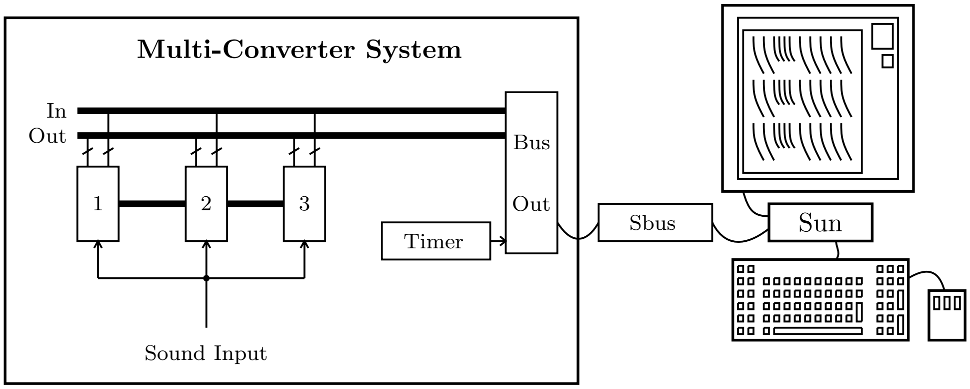


Figure 1. Block diagram of the multi-converter system.

Figure 2 shows the analog signal path of the auditory pre-processor in the converter chip. Processing begins with a silicon cochlea circuit (Lyon and Mead, 1988). A silicon cochlea is an analog circuit implementation of the differential equations that describe the traveling wave motion of physiological cochleas. The cochlea design used in this chip maps a linear, one-dimensional partial-differential equation into circuits, as a cascade of continuous-time filter sections with exponentially decreasing time constants. The second-order filter sections have a low-pass response, with a slight resonant peak before cutoff. The cascade acts as a discrete-space, continuous-time finite-element approximation of the partial differential equation.

Like wavelet filterbanks, the silicon cochlea outputs balance temporal and spectral acuity. The cochlear output response, a lowpass filter with a sharp cutoff and a slight resonant peak, derives its spectral selectivity from the collective interaction of the slightly-resonant circuits in the series cascade, not from parallel highly-resonant circuits as in a standard filterbank. By avoiding highly-resonant filters, the cochlear processing preserves the temporal details in each output channel.

This cochlear design is the first stage of processing in our chip. The cochlea consists of 139 filter stages; we use the outputs of the last 119 stages. The first 20 outputs are discarded, because their early position in the cascade results in a poor approximation to the desired differential equation solution. Four parameters control the tuning of the silicon cochlea, supporting variable frequency ranges and resonance behaviors.

Next in the signal processing chain (Figure 2) are circuits that model the signal processing that occurs during the sensory transduction of mechanical motion in the cochlea. These operations include time differentiation, half-wave rectification, amplitude compression, and the conversion of the analog waveform representation into probabilistic trains of fixed-width, fixed-height spikes (Lazzaro and Mead, 1989c). Each of the 119 cascade outputs is coded by 6 probabilistic spiking circuits. Note that no time averaging has been done in this signal processing chain; the cycle-by-cycle waveform shape is fully coded in each set of 6 spiking outputs.

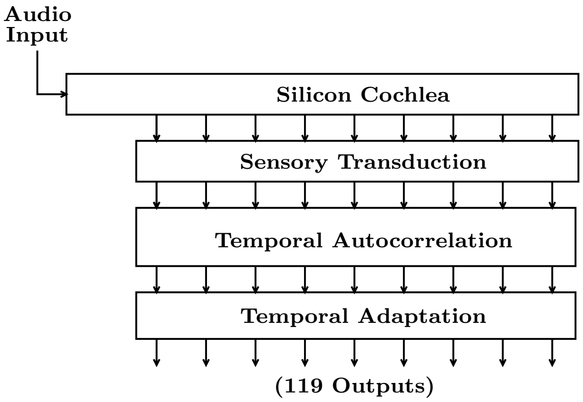


Figure 2. Analog signal path of the silicon auditory model.

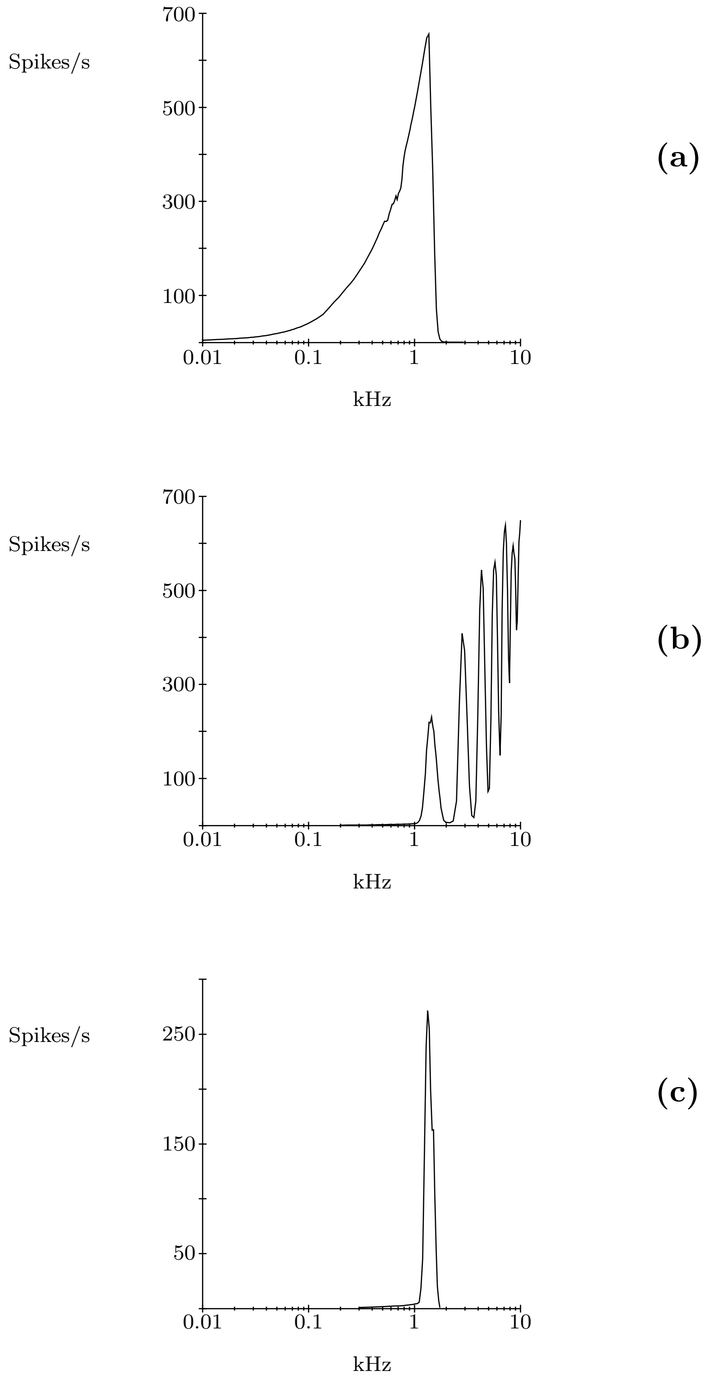


Figure 3. Periodicity-based spectral shape computation: (a) silicon cochlea tuning response (b) temporal autocorrelator tuning response (c) the combination of cochlear and autocorrelator processing.

Different secondary representations in the brain use the cochlear signal as input, and produce outputs that represent more specialized aspects of the sound. In our chip processing chain, two signal processing blocks follow the sensory transduction block, that may be used to model a variety of known and proposed secondary representations.

The first processing block (Figure 2) implements temporal autocorrelation, in a manner described in detail in (Lazzaro, 1991). The six spiking outputs associated with each cochlear output are sent into a single temporal autocorrelator, which produces a single output. Six parameters fix the autocorrelation time constant and autocorrelation window size at both ends of the representation; autocorrelation parameters for intermediate taps are exponentially interpolated.

The temporal autocorrelation block can be configured to generate a representation that codes the spectral shape of a signal. To generate this spectral shape representation, the autocorrelator associated with each cochlear channel is tuned so that the best frequency of the cochlear channel matches the first peak of the autocorrelation function (Sachs and Young, 1980). Figure 3 illustrates the algorithm: Figure 3(a) shows the frequency response of a cochlear output, Figure 3(b) shows the output of a temporal autocorrelator tuned to the best frequency of the cochlear output, and Figure 3(c) shows the effect of combining the temporal autocorrelator and cochlear filter. By cascading the cochlea and temporal autocorrelator blocks, a narrow, symmetrical filter is created; this filter is non-linear, and achieves a narrow bandwidth without using a highly resonant linear filter.

The final processing block in the signal processing chain (Figure 2) implements temporal adaptation, which acts to enhance the transient information in the signal. Figure 4 illustrates temporal adaption: in response to a tone burst (top trace), the circuit produces a series of pulses (bottom trace). The number of pulses per second is highest at the onset of the sound, and decays to a lower rate during the unchanging portion of the tone. Five parameters fix the time constant and peak activity rate of temporal adaption at both ends of the representation: parameters for intermediate taps are exponentially interpolated from these fixed values. These parameters support a wide range of adaptive responses, including temporal adaptation behaviors typical of auditory nerve fibers, as well as behaviors typical of on-cell neurons in the cochlear nucleus. The circuits used in the temporal adaptation block are described in detail in (Lazzaro, 1992).

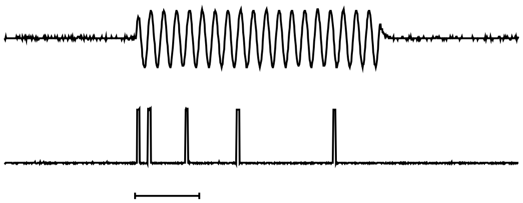


Figure 4. Temporal adaptation: top trace is audio input (gated tone burst), bottom trace shows adaptive response. Bar length is 5ms.

As shown in Figure 4, the final outputs of the auditory model take the form of pulse trains. These pulses are fixed-width, fixed-height, and occur asynchronously; they are not synchronized by a global clock. The information sent by a spike is fully encoded by its moment of onset. In collaboration with other researchers, we have developed efficient methods to transmit the information from an array of asynchronous spiking circuits off chip (Lazzaro *et al.,* 1993), and to combine the information from several chips to form a single data stream in an efficient way (Lazzaro and Wawrzynek, 1995a). We use these methods in our multi-representation system.

Figure 5 shows the programmer's model of this data stream. Data from the system takes the form of a list of "events": each event corresponds to a single spike of an output unit from a chip in the multi-representation system. Each event includes information specifying the chip sending the spike, the cochlear channel associated with the spike, and the moment of onset of the spike. The onset timestamp has a resolution of $20\mu s$; event lists are strictly ordered with respect to onset times.

We designed a software environment, **Aer**, to support real-time, low-latency visualization of data from the multi-converter system (Lazzaro *et al.,* 1994). The environment also supports a scripting language for the automatic collection of system response to large sound databases.

3. REPRESENTATIONS FOR SPEECH RECOGNITION

We configured our multi-representation system to generate specialized representations for speech analysis: a spectral shape representation for voiced speech, a periodicity representation for voice/unvoiced decisions, and an onset representation for coding transients. Figure 6 shows a screen from Aer, showing these three representations as a function of time: the input sound for this screen is a short 800 Hz tone burst, followed by a sinusoid sweep from 300 Hz to 3 Khz. For each representation, the output channel number is plotted vertically; each dot represents a pulse.

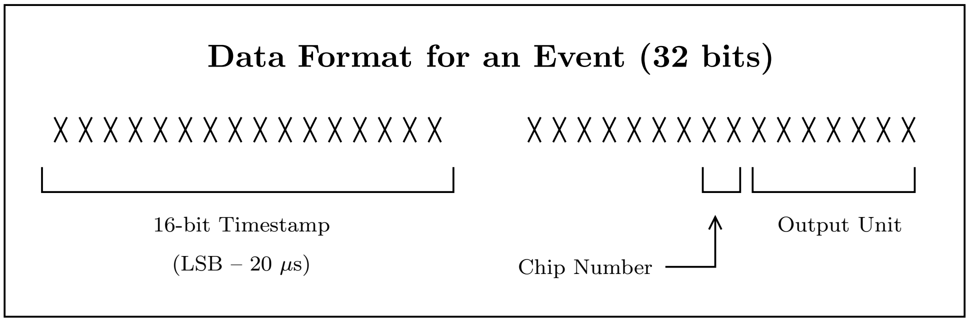


Figure 5. Programmers interface for events.

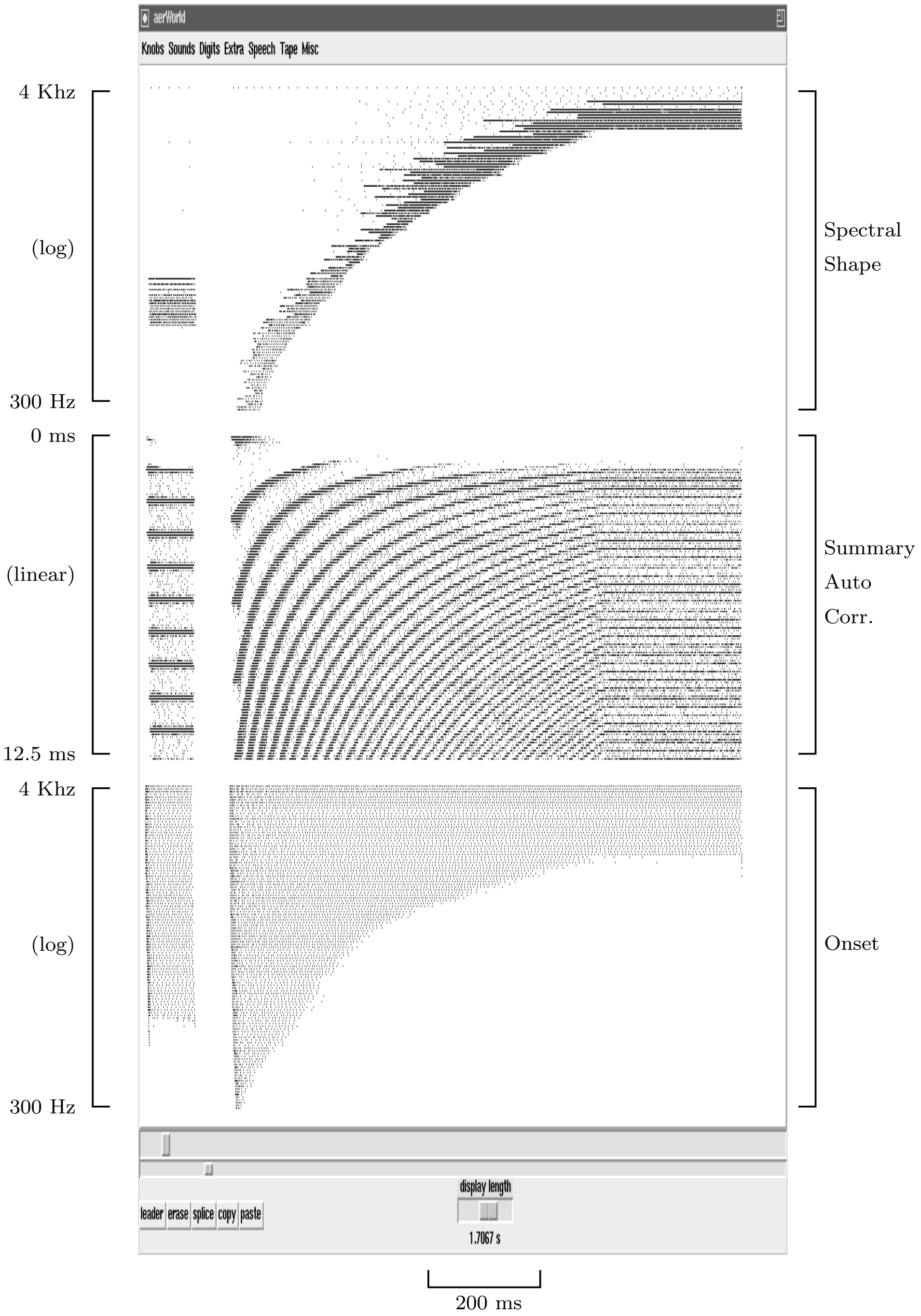
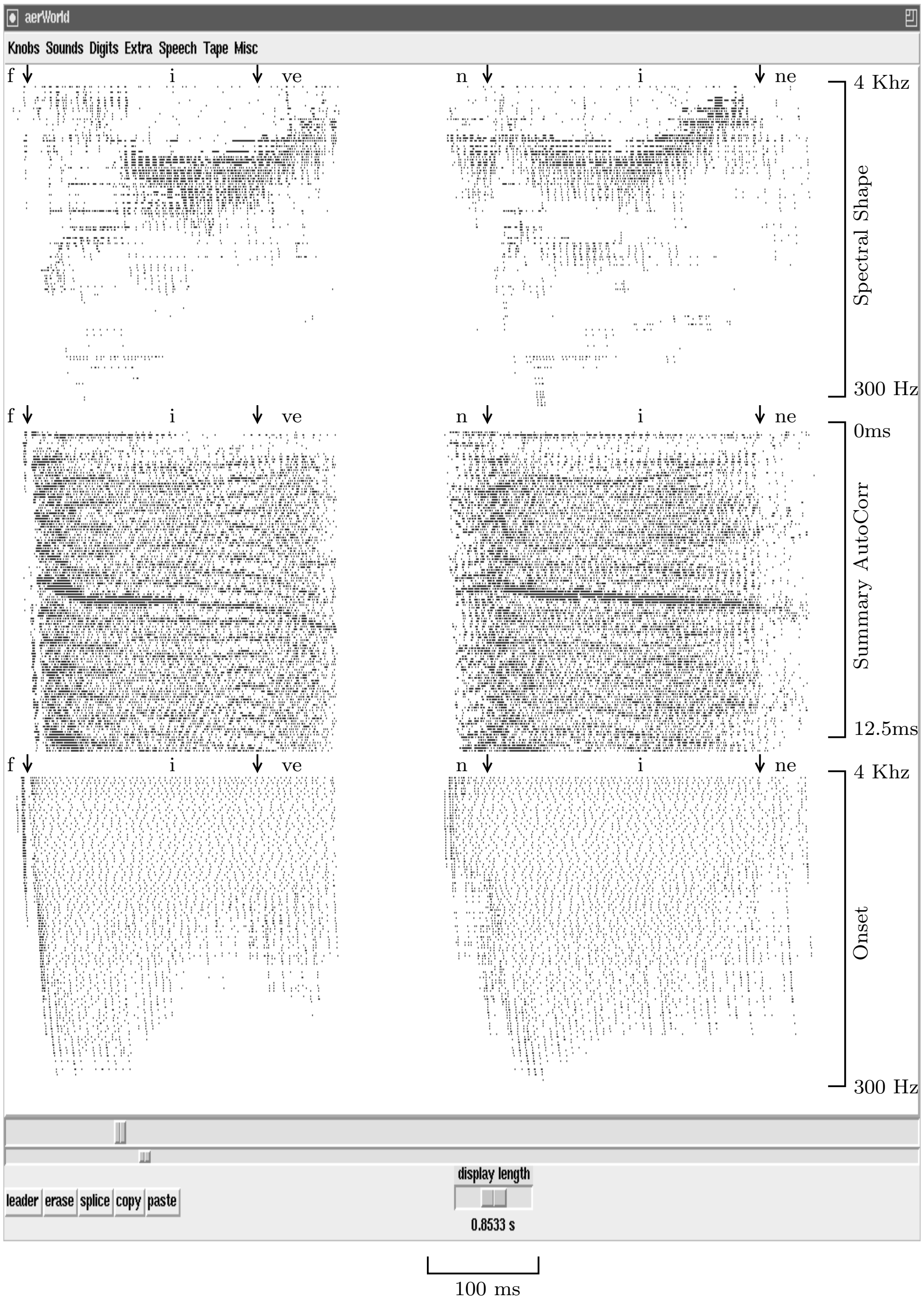


Figure 6. Data from the multi-converter system, in response to a 800-Hz pure tone, followed by a sinusoidal sweep from 300Hz to 3Khz.



display length
0.8533 s

Figure 7. Data from the multi-converter system, in response to the word “five” followed by the word “nine”.

The top representation codes for periodicity-based spectral shape. For this representation, the temporal autocorrelation block generates responses as shown in Figure 3, and the temporal adaptation block is inactive. Spectral frequency is mapped logarithmically on the vertical dimension, from 300 Hz to 4 Khz; the activity in each channel codes the presence of a periodic waveform at that frequency. The difference between a periodicity-based spectral method and a resonant spectral method can be seen in the response to the 800 Hz sinusoid onset: the periodicity representation shows activity only in a narrow band of channels, whereas a spectral representation would show broadband transient activity at tone onset.

The bottom representation codes for temporal onsets. For this representation, the temporal adaptation block is active, and the temporal autocorrelation block is inactive. The spectral filtering of the representation reflects the silicon cochlea tuning: a low-pass response with a sharp cutoff and a small resonant peak at the best frequency of the filter. Temporally, the representation produces a large number of pulses at the onset of a sound, decaying to a small pulse rate with a $10m$s time constant. The black, wideband lines at the start of the 800 Hz tone and the sinusoid sweep illustrate the temporal adaptation; the tuned response throughout the sinusoid sweep illustrates the low-pass spectral tuning.

The middle representation is a summary autocorrelogram, useful for pitch processing and voiced/unvoiced decisions in speech recognition. This representation is not raw data from a converter; software post-processing is performed on a converter's output to produce the final result. The frequency response of the converter is set as in the onset representation; the temporal adaptation response, however, is set to a $100m$s time constant. The converter output pulse rates are set so that the cycle-by-cycle waveform information for each output channel is preserved.

To complete the representation, a set of running autocorrelation functions $x(t)x(t-\tau)$ is computed for $\tau = k\,105\mu s$, $k = 1 \ldots 120$, for each of the 119 output channels. These autocorrelation functions are summed over all output channels to produce the final representation, a summary of autocorrelation information across frequency bands. τ is plotted as a linear function of time on the vertical axis. The correlation multiplication can be efficiently implemented by integer subtraction and comparison of event timestamps; the summation over channels is done by merging event lists. Figure 6 shows the qualitative characteristics of the summary autocorrelogram: a repetitive band structure in response to periodic sounds. In contrast, the summary autocorrelation function of a noise signal shows no long-term spatial structure.

Figure 7 shows the output response of the multi-converter system in response to telephone-bandwidth-limited speech; the phonetic boundaries of the two words, “five” and “nine”, are marked by arrows. The vowel formant information is shown most clearly by the strong peaks in the spectral shape representation; the wideband information in the “f” of five is easily seen in the onset representation. The summary autocorrelation representation shows a clear texture break between vowels and the voiced “n” and “v” sounds.

4. FEATURES FOR SPEECH RECOGNITION

The data shown in Figures 6 and 7 share many properties with neural responses. Each output unit codes information asynchronously, and the effective sampling period is adaptive and data dependent. Each representation in the system is specialized for a certain property of sound. These representations are not uncorrelated: there is considerable redundancy between the representations, and among output units of a single representation. These properties of neural auditory representations are summarized in Figure 8.

Also shown in Figure 8 are the contrasting properties of conventional feature representations used in speech recognition systems. These representations generate features at a single uniform frame rate, typically 10-20ms, unsynchronized to acoustic features. A single, general-purpose spectral representation is typically used: often, both signal energy information and pitch contour information is removed from this representation. Finally, low-dimensional representations are used (5 to 15 elements, typically), and the components of the representation are often uncorrelated. These front-end properties reflect the statistical and architectural properties of recognition systems.

Figure 8 depicts a "representation-recognizer gap" that complicates the use of auditory models for speech recognition. We address this issue in two ways: by transforming the representations shown in Figures 6 and 7 to have properties closer to conventional front-end representations, and by choosing speech recognition technology that is more compatible with auditory models. The method we used to extract a feature vector from our multi-representation system output is described below.

The first step in feature extraction is to convert the asynchronous, event-list representation into a sequence of uniformly sampled frames. Each frame output consists of 3 vectors (one for each representation) with 119 floating point elements (one for each output unit), and codes the spike activity that occurs during a 25ms interval. Subsequent frames overlap in time by 12.5ms. To generate each frame element, the spiking pattern during the 25ms interval is considered as a train of delta functions with unit height: this function is multiplied by a Hamming window. After multiplication, the heights of the delta functions are summed to yield the final floating-point feature element value. These operations are graphically shown in Figure 9.

Auditory Models	**Speech Recognition**
Adaptive Sampling	Uniform Sampling
Specialized Features	General-Purpose Features
Multiple Representations	Single Representation
High-Dimensional	Low-Dimensional
Correlated Features	Uncorrelated Features

Figure 8. Comparison of auditory representations and current speech recognition technology.

To reduce the size of the spectral-shape and onset representations, we subsample the original 119-element vectors using symmetrical triangular filters with a 50-percent filter response overlap. This subsampling produces a 5-element vector coding onsets, and an 8-element vector coding spectral shape. The subsampling procedure is graphically shown in Figure 10.

To reduce the size of the summary autocorrelogram, we compute the discrete cosine transform of the 119-element vector, and use the first two components of the transform as the summary autocorrelogram feature vector. We choose this reduction technique to enhance the coding of voicing in the representation, while de-emphasizing formant information.

These reduction techniques resolve the feature-size table entries in Figure 8; they do not, however, address the correlation between feature elements. As detailed in the next section, we use recognition technologies that are relatively insensitive to correlations between feature elements to address this issue.

To complete our feature vector computations, we compute temporal difference feature vectors ("delta" features) for each primary representations, using a 5-frame window to compute differences. The resulting feature vector has 30 elements: 8 spectral-shape elements, 5 onset elements, 2 summary autocorrelogram elements, for a total of 15 primary elements, together with 15 delta features.

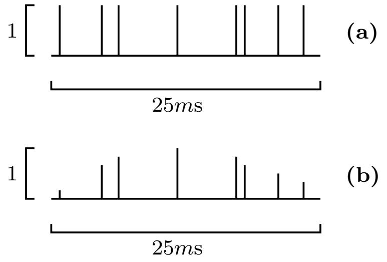


Figure 9. Graphic description of algorithm for converting the asynchronous event-list representation into uniformly sampled frames. A 25ms series of unit-height events **(a)** is scaled by a Hamming window **(b)**. The heights of the scaled events are added to form the frame value.

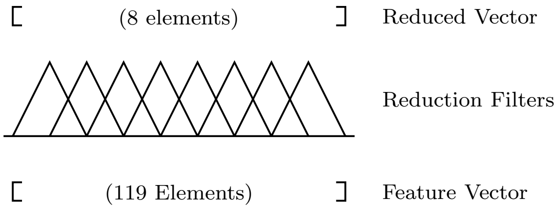


Figure 10. Graphic description of algorithm for subsampling full 119-element representations into a reduced feature vector.

5. SPEECH RECOGNITION ARCHITECTURES

In modern speech recognition architectures, word recognition from a sequence of feature vectors is a two-step process. In the first step, a pattern classifier maps the sequence of feature vectors into a sequence of predictions of the spoken phoneme in progress. In the second step, a vocabulary of permitted words is introduced (a lexicon), expressed as probabilistic state machines (Hidden Markov Models, abbreviated as HMMs). The sequence of phoneme probabilities is then mapped into a sequence of words from the lexicon, using dynamic programming.

While state models for a lexicons are typically crafted by hand, the features-to-phonemes pattern classifier is trained automatically, using a large database of example words. One popular classifier for this task uses a linear mixture of multivariate Gaussian functions to map the feature vector into the probability a particular phoneme is in progress. A complete mixture model has several types of parameters: each multi-variate Gaussian function has a mean vector and a diagonal covariance matrix, and weighting parameters control the contribution of each function in the mixture.

The choice of a diagonal covariance matrix reduces the number of covariance parameters for an N-element feature vector from N^2 to N. This choice enables the liberal application of multiple Gaussians to model probability space, using an acceptably small number of parameters. Low-parameter models require less training data for effective parameter estimation, and often improve generalization properties.

Choosing a diagonal covariance matrix is warranted if the off-diagonal matrix elements are small; this matrix property will be true of the elements of the feature vector are uncorrelated. However, as noted in the last section, the elements of the feature vector we generate from our multi-representation system are indeed correlated.

An alternative approach for mapping a feature vector into a phoneme probability vector is to use a multi-layer perceptron (MLP) architecture, trained with the backpropagation algorithm. This approach, as described in (Bourlard and Morgan, 1994), is more tolerant of feature vectors whose elements are correlated. The speech recognition results we present in this paper all use this MLP-based recognizer.

The input to the neural-network classifier is the feature vector of the current frame, as well as feature vectors from the 4 previous frames and the 4 upcoming frames for context. The net has a single hidden layer; unless otherwise indicated we use 200 hidden units in the system. There are 56 network outputs, associated with the 56 most-common phonemes; these outputs are the inputs to a dynamic programming module that performs word recognition.

In our speech recognition experiments, we used a database of 200 adults speaking 13 English words in isolation (the digits, including both “oh” and “zero”, plus “yes” and “no”), for a total of 2600 utterances. The database, supplied by Bellcore, was recorded over the U.S. public telephone network: the recordings typically have good signal-to-noise ratios, but display the limited bandwidth typical of the telephone network. Most of our experiments used this database directly; in some experiments, we added recorded noise from the interior of a running automobile to the speech, at a level resulting in a 10dB signal-to-noise ratio.

The small size of the database results in a significant variance of recognition performance, depending on the particular words chosen to be in the training set and the test set. To counter this problem, we divide the database into 4 segments, each consisting of 650 utterances by 50 speakers. We then train up 4 different recognizers using a different selection of three segments for training data, while testing on the fourth segment. Note that a recognizer is never tested using utterances used for its training. In this paper, we report the error scores for each of these four recognizers, along with the averaged score of the four recognizers.

In our recognition experiments, each word is modeled using a multi-state HMM; each state has a self-loop branch and a branch to the next state, with fixed transition probabilities of 0.5 for each branch. The model length varies with the number of phonemes in the word: “eight” is the shortest model, with 13 states, while “seven” is the longest model, with 18 states.

This isolated word database has been used for several previous speech recognition studies at ICSI, using the MLP-based recognizer in conjunction with two popular feature extraction systems, PLP and J-RASTA-PLP. These recognition studies, summarized in Figure 11, serve as a benchmark for comparison with recognition experiments using feature vectors derived from our multi-representation auditory model. Perceptual Linear Prediction (PLP) is a popular feature extraction system based on human perceptual data from psychophysics, that works well for speech recorded with high signal-to-noise ratios through benign transmission channels (Hermansky, 1990). The J-RASTA-PLP system (Hermansky and Morgan, 1994; Ma, 1995) is an enhanced version of PLP, designed to provide feature vectors relatively independent of noise mixed with the speech signal, as well as providing feature vectors independent of slowly-varying changes in the spectral properties of the speech transmission channel.

The first table entry in Figure 11 shows the performance of J-RASTA-PLP on the isolated word database, for the 4 data segmentations described above: both training and test data are the original “clean” database, without added car noise. The 1.8 percent average error score is comparable with current commercial systems used in voice-mail applications, using real-world telephony data: trade publications for interactive voice response telephony advise users to expect 3-5% scores for isolated digit recognizers in the field. Systems with error rates under 5% can work well in an application, if good error recovery strategies are available for the task.

Front End	Conditions	1	2	3	4	Average
J-RASTA-PLP	Clean	2.3	1.5	1.4	2.0	1.9
J-RASTA-PLP	Noisy	11.4	10.2	10.3	11.5	10.9
PLP	Noisy	42.8	37.1	40.8	49.1	42.4

Figure 11. Percent error for PLP-based front-ends (four database partitions).

The next two table entries in Figure 11 describe performance for recognizers that were trained "clean" utterances, but whose test utterances were mixed with automobile noise (10 dB signal-to-noise ratio), as described earlier. Note that in addition to being corrupted with noise, the test utterances were also novel: a recognizer is never tested using noisy utterances whose clean versions were a part of the training set. These table entries show the benefits of enhancing a feature extraction system to be robust to additive noise: the average error for PLP is 4 times greater than for J-RASTA-PLP. However, the absolute error of J-RASTA-PLP tested with noisy speech (10.9%) is marginal for use in an application.

6. SPEECH RECOGNITION EXPERIMENTS

We used the isolated-word database and MLP-based recognition system described in the previous section to evaluate the performance of the 30-element feature vector derived from our multi-representation system. Figure 12 summarizes the error performance of the system; all the scores in this table reflect "clean" testing and training data. The final line of Figure 12 shows recognition performance using the full 30-element feature vector. This 4.1% error rate is sufficiently low for many applications, although it is significantly larger than the J-RASTA-PLP's benchmark error rate (1.8%).

Features	Parameters	Hidden Units	1	2	3	4	Average
SS	65,586	326	6.6	6.9	5.4	8.0	6.7
SS + Auto	65,468	276	5.7	5.8	4.5	5.5	5.4
SS + Onset	65,531	225	4.9	5.1	4.3	4.9	4.8
SS + Auto + Onset	65,456	200	4.9	4.2	3.2	4.0	4.1

Figure 12. Percent error for feature vectors derived from auditory representations (four database partitions). Other fields show number of hidden units and number of parameters in the MLP classifier net. Code: SS = spectral shape features, Onset = onset features, Auto = autocorrelogram features.

Features	Total	9/5	oh/no	others
SS	6.7	1.4	1.0	4.4
SS + Auto	5.4	1.1	0.8	3.5
SS + Onset	4.8	1.0	0.7	3.1
SS + Auto + Onset	4.1	0.7	0.6	2.8

Figure 13. Error analysis of the recognition experiments in Figure 12 (averaged over partitions). Errors due to the two leading word confusions are listed (confusing "five" and "nine", and confusing "oh" and "no"), as well as the residual error.

To illustrate the relative contributions of the three representations in our system, we also trained recognizers using subsets of our 30-element feature vector, that contained only the elements from one or two of the representations in the system: the penultimate lines of Figure 12 show these scores. The number of hidden units for each recognizer was varied inversely with the number of elements in the reduced feature vector, to yield an MLP with approximately 65,000 parameters for each experiment.

A comparison of the different recognizers in Figure 12 shows the effectiveness of combining multiple representations of speech. Adding features from two additional representations (the onset features and the autocorrelation features) to the primary spectral-shape features decreases the average error by 61%.

Figure 13 shows an error analysis of the recognizers of Figure 12; for each recognizer, the percentage error attributed to the two most likely word confusions ("five" and "nine", and "no" and "oh") are shown, along with the residual error contributed by all other confusions. The addition of onset features and autocorrelogram features improves the recognition performance for all three categories of confusions.

In Figure 13, note that for five/nine confusions, the error improvements for adding onset features (1.4% to 1.1%, a 0.3% improvement) and for adding autocorrelation features (1.4% to 1.0%, a 0.4% improvement) add to equal the error improvement for adding both onset and autocorrelation features to spectral shape features (1.4% to 0.7%, a 0.7% improvement). This linear addition suggests the statistical independence of the information added by the onset and autocorrelation features for disambiguating "five" and "nine". Conversely, the table shows the statistical dependence of the information added by the onset and autocorrelation features for disambiguating words in the "others" category. If these features were statistically independent, an error rate of 2.2% (not 2.8%) would be expected for the "others" category for the full feature vector.

Figure 14 shows the error performance of the recognizers trained for Figure 12, when tested on the "noisy" utterances described in the last section. This table shows uniformly poor recognition results, comparable with the noisy recognition performance of PLP shown in Figure 11, and 5.5 times worse than the noisy recognition performance of J-RASTA-PLP. Although early studies of speech recognition in noisy conditions using auditory models reported encouraging results (Ghitza, 1998; Seneff, 1988), later studies found no significant noise robustness qualities for auditory models (Lippmann, 1995), and the data in Figure 14 confirms this finding.

Features	**Parameters**	**Hidden Units**	**1**	**2**	**3**	**4**	**Average**
SS	65,586	326	50	54	50	55	52
SS + Auto	65,468	276	53	55	51	55	54
SS + Onset	65,531	225	55	57	61	62	59
SS + Auto + Onset	65,456	200	57	57	59	62	59

Figure 14. Recognition results for noisy test data; automobile noise source, mixed with speech at signal-to-noise ratio of 10dB (NIST measurement method).

Figure 15 shows the effect of reducing the number of parameters in the MLP pattern classifier on error rate, for the full 30-element feature vector. The table compares recognizers with approximately 32,000 parameters (100 hidden units) and 16,000 parameters (50 hidden units) with the full 65,000 parameter recognition system. The effect of parameterization on error performance is particularly important for low-cost, low-power recognizer implementations.

The 200 speaker, 13-word isolated-word database consists of approximately 5 hours of speech. The analog processing circuits in the multi-representation system are not compensated for temperature drift; we omitted temperature compensation circuitry from our prototype system for simplicity. Ambient temperature variation in our laboratory over five hours results in a significant drift in auditory model responses.

To counter temperature drift problems during datataking for the experiments reported above, several steps were taken. Data was presented to the chip ordered by word: 200 speakers saying "1," followed by 200 speakers saying "2," ect. All chip parameters were recalibrated between each set of 200 utterances; this recalibration resets parameters with 5% accuracy. The complete dataset was taken several times, on different days of the week and different times of the day, and pilot recognition experiments guided the choice of the final dataset. Within the limits of our present hardware prototype, these error scores approximate the performance of a temperature-compensated multi-representation system.

For comparison purposes, Figure 16 shows recognition performance for the multi-representation system, if less care is taken to reduce temperature effects. In these experiments, data was presented to the system ordered by speaker, not by word: speaker 1 saying all 13 words, followed by speaker 2 saying all 13 words, ect. Recalibration of parameters occured every 10 speakers. The scores in this table reflect "clean" testing and training data; the final line of Figure 12 is reproduced in Figure 16 to provide a direct comparison between the two datasets.

Features	**Parameters**	**Hidden Units**	**1**	**2**	**3**	**4**	**Average**
SS + Auto + Onset	65,456	200	4.9	4.2	3.2	4.0	4.1
SS + Auto + Onset	32,756	100	5.1	4.6	3.8	4.3	4.5
SS + Auto + Onset	16,406	50	5.7	6.8	4.6	6.6	5.9

Figure 15. Recognition results showing the effect of the number of parameters in the MLP-classifier on recognition results. Results are for clean testing data, using the full feature vector (SS+Onset+Auto).

Order	**Parameters**	**Hidden Units**	**1**	**2**	**3**	**4**	**Average**
By Word	65,456	200	4.9	4.2	3.2	4.0	4.1
By Speaker	65,456	200	6.6	4.9	6.0	7.7	6.3

Figure 16. Recognition results showing the effect of data presentation on recognition results. Results are for clean testing data, using the full feature vector.

7. DISCUSSION

The speech recognition performance of our multi-representation system, as shown in Figure 12 and Figure 14, is inferior to J-RASTA-PLP, both for clean and noisy test data. However, under high signal-to-noise conditions, the system provides adequate performance (4.1% error) for many isolated-word applications. For specialized applications where a micropower speech feature extractor is required, the signal processing technology used in our special-purpose analog-to-digital converter chip is a competitive option. For these applications, the remaining challenges include the micropower implementation of the rest of the recognition system, and the identification of end-user applications with sufficient market size to support the development effort.

Apart from the micropower niche, however, analog auditory models are currently uncompetitive with conventional front-end approaches for speech recognition applications. The success of auditory processing in biological systems, however, leaves us hopeful that a sustained research effort in using analog auditory models for speech recognition could result in recognition systems that perform significantly better than conventional front-end approaches. We see the following areas as important elements of such a research effort:

Improved Circuit Techniques.

The 4.1% error of the multi-representation system, for clean speech, is distinctly inferior to the 1.8% error for J-RASTA-PLP on the same task. In contrast, studies of software implementations of similar auditory models (Jackowoski *et al.*, 1995) typically show comparable performance in comparison with conventional front-ends. The shortcomings of our analog circuit implementation, including limited signal-to-noise ratio, limited dynamic range, and inaccuracy due to parameter variation and temperature-related drift, may play a role in this difference.

The circuit technologies that implement the signal processing datapath shown in Figure 2 date from the first silicon audition designs (Lyon and Mead, 1988). Several generations of improved circuits and algorithms for silicon audition have been published since these early designs, and research continues in several groups worldwide. Many of these improvements focus on signal-to-noise, dynamic range, and improving uniformity across cochlea channels. These improvements may directly translate to improvements in speech recognition scores, bringing silicon auditory models to the performance of their software counterparts.

Parameter drift due to inadequate temperature compensation is another area for improvement, the temperature compensation approach we use in our multi-representation is primitive (Lazzaro *et al.*, 1994), and parameter drift may be a significant source of recognition error, as Figure 16 suggests. Improvements in this area are straightforward, using techniques such as those described in (Vittoz, 1985).

Enhanced Auditory Models.

The cochlear model in our special-purpose analog-to-digital converter chip is an extreme simplification of physiological cochlear processing; software-based auditory models used in other speech recognition studies share most of these simplifications. Key physiological cochlear response characteristics, including synchrony suppression, rate suppression, and temporal masking, are absent from these models; many auditory theorists believe these characteristics underlie the robust coding of speech in the presence of noise in biological auditory systems. Physiological cochleas are deeply non-linear, and exhibit characteristics consistent with extensive channel-specific automatic gain control: the auditory models used in speech recognition experiments to date do not correctly model these characteristics. We believe that more accurate cochlear models are an important part of future research in using auditory representations for speech recognition.

In addition to improving the cochlear models, cleaner implementations of the computations underlying the secondary representations in our system (correlation and temporal adaptation) would add considerable robustness to these representations. Also of interest is the addition of other secondary representations, in particular models of neural maps that code for temporal offsets, amplitude modulation, frequency modulation, and quick temporal sequences typical of the voiced-onset transition in speech. If multi-microphone recordings of speech databases are available, binaural representations are another possible enhancement to the system.

Adapting Robust Techniques to Auditory Models.

A variety of techniques for robust feature extraction in noisy environments have been developed for use with conventional front-ends for speech recognition. Adapting these techniques to function with auditory representations is a promising avenue of research.

One popular method of speech enhancement in noise is spectral subtraction (Boll, 1979). In this approach, a spectral model of the background noise in the recent past is generated, and subtracted from the current input. Another method of speech enhancement in noise, the J-RASTA-PLP system (Hermansky and Morgan, 1994; Ma, 1995), uses information about the temporal properties of speech to filter speech signal from background noise. Both approaches could be used in conjunction with auditory representations.

Closing the Representation-Recognizer Gap.

As Figure 8 summarizes, auditory representations are a poor match to current speech recognition systems. This paper makes no significant contribution towards closing this "representation-recognizer gap". Our straightforward approach of collapsing the spike-based, high-dimensionality auditory representations (Hamming windows and gross sub-sampling) destroys most of the unique coding aspects of the auditory representation. Apart from choosing an MLP-based pattern classifier, no advances in recognition algorithms were made to help close the gap from the recognition side.

We believe that making significant contributions to closing this gap, both by modifying core speech recognizer technology, and by developing enhanced methods of distilling information from high-dimensional, adaptively-sampled representations, is essential to significantly improve speech recognition performance of auditory models.

Several research groups have done initial work on changing core speech recognition technology to be more amenable to auditory representations. These methods take different approaches to the problem; one recent publication uses the visual scene analysis concept of occlusion as a starting point (Cooke *et al.*, 1994), while other recent work is motivated by the importance transient information in the speech signal (Morgan *et al.*, 1994). Attacking the problem from the representation side, research in mapping in high-dimensional spaces into low-dimensional features has been recently applied to cochlear models (Intrator, 1993).

8. SUMMARY

In this paper, we have evaluated the suitability of analog implementations of auditory models, using an empirical approach: we integrated a multi-representation analog auditory model with a speech recognition system, and measured the performance of the system on a speaker-independent, telephone-quality 13-word recognition task. The performance of the system is adequate for many applications, but inferior to conventional approaches for front-end processing. In addition, the auditory models show no advantages for robust speech recognition applications.

Acknowledgments

We thank Nelson Morgan for providing the speech databases, the speech-recognition software, and the computational resources for this study; we also acknowledge his extensive advice on MLP design and training. We thank Su-lin Wu for her extensive help in the recognition studies; we also thank the entire ICSI speech recognition group for suggestions and software maintenance. Thanks to Richard Duda, Steve Greenberg, Hynek Hermansky, Richard Lippmann, Richard Lyon, Carver Mead, Malcolm Slaney, and the members of the Physics of Computation laboratory at the California Institute of Technology for advice and suggestions on auditory modeling. We used the HTK Toolkit Version 1.4 for preliminary speech recognition studies preceding this work. Research funded by the Office of Naval Research (URI-N00014-92-J-1672).

References

Bhadkamkar, N. A. (1994). Binaural source localizer chip using subthreshold analog CMOS. *1994 IEEE International Conference on Neural Networks,* **3**, 1866-1870.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *Trans. on ASSP*, **ASSP27**:2, 113-120.

Bourlard, H. and Morgan, N. (1994). *Connectionist speech recognition : a hybrid approach.* Boston, Mass: Kluwer Academic Publishers.

Brown, G. J. and Cooke, M. P. (1994). Computational auditory scene analysis. *Computer Speech and Language*, **8:**4, 297-336.

Chandrakasan, A. P. and Brodersen, R. W. (1995). *Low power digital CMOS design.* Boston, Mass: Kluwer Academic Publishers.

Coggins, R., Jabri, M., Flower, B., Pickard, S. (1995). A hybrid analog and digital VLSI neural network for intracardiac morphology classification. *IEEE Journal Solid State Circuits,* **30:**5, 542-550.

Colomes, C., Lever, M., Rault, J. B., and Dehery, Y. F. (1995). A perceptual model applied to audio bit-rate reduction. *J. Audio Eng. Soc.,* **43:**4, 233-239.

Cooke, M., Beet, S., and Crawford, M. (1993). *Visual Representations of Speech Signals,* New York: Wiley.

Cooke, M., Green, P., Crawford, M. (1994). Handling missing data in speech recognition. *1994 International Conference on Spoken Language Processing,* **3**, 1555-1558.

Ghitza, O. (1988). Temporal non-place information in the auditory nerve firing patterns as a front-end for speech recognition in a noisy environment. *Journal of Phonetics,* **16:**1, 109-123.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal Acoustical Society of America,* **87:**4, 1738-1752.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions of Speech and Audio Processing,* **2**:4, 578-589.

Intrator, N. (1993). Combining exploratory projection pursuit and projection pursuit regression with application to neural networks. *Neural Computation,* **5**, 443-455.

Jackowoski, C. R., Vo, H. D. H., Lippmann, R. P. (1995). A comparison of signal processing front ends for automatic word recognition. *IEEE Transactions of Speech and Audio Processing,* **3**:4, 286-293.

Lazzaro, J. P. and Mead C. (1989a). Silicon models of pitch perception. *Proc. Natl. Acad. Sci. USA,* **86**, 9597-9601.

Lazzaro, J. P. and Mead C. (1989b). Silicon models of auditory localization. *Neural Computation,* **1**, 47-57.

Lazzaro, J. P. and Mead, C. (1989c). Circuit models of sensory transduction in the cochlea. In *Analog VLSI Implementations of Neural Networks,* Mead, Ismail, (eds.), Norwell, MA: Kluwer, 85-101.

Lazzaro, J. P (1991). A silicon model of an auditory neural representation of spectral shape. *IEEE Journal Solid State Circuits,* **26**, 772-777.

Lazzaro, J. P. (1992). Temporal adaptation in a silicon auditory nerve. In Moody, J., Hanson, S., Lippmann, R. (eds), *Advances in Neural Information Processing Systems 4.* San Mateo, CA: Morgan Kaufmann Publishers.

Lazzaro, J. P., Wawrzynek, J., Mahowald., M., Sivilotti, M., and Gillespie, D. (1993). Silicon auditory processors as computer peripherals. *IEEE Journal of*

Neural Networks **4**:3, 523-528.

Lazzaro, J. P., Wawrzynek, J., and Kramer, A (1994). Systems technologies for silicon auditory models. *IEEE Micro*, **14**:3, 7-15.

Lazzaro, J. P. and Wawrzynek, J. (1995a). A multi-sender asynchronous extension to the address-event protocol. In Dally, W. J., Poulton, J. W., Ishii, A. T. (eds), *16th Conference on Advanced Research in VLSI,* 158-169.

Lazzaro, J. P. and Wawrzynek, J. (1995b). Silicon models for auditory scene analysis. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E. (eds) *Advances in Neural Information Processing Systems 8,* Cambridge,Mass: MIT Press.

Liu, W., Andreou, A., and Goldstein, M. (1992). Voiced-speech representation by an analog silicon model of the auditory periphery. *IEEE Transactions of Neural Networks* **3**:3, 477-487

Lyon, R. F., and Mead, C. (1988). An analog electronic cochlea. *IEEE Trans. Acoust., Speech, Signal Processing* **36**, 1119-1134.

Lyon, R. F. (1991). CCD correlators for auditory models. *IEEE Asilomar Conference on Signals, Systems, and Computers*, 1991, 785-789.

Ma, K. W. (1995). Applying Large Vocabulary Hybrid HMM-MLP Methods to Telephone Recognition of Digits and Natural Numbers. *International Computer Science Institute Technical Report,* TR-95-024.

Morgan, N., Bourlard, H., Greenberg, S., and Hermansky, H. (1994). Stochastic perceptual auditory-event-based models for speech recognition. *1994 International Conference on Spoken Language Processing,* **4**, 1943-1946.

Sachs, M. B. and Young. E. D. (1980). Effects of nonlinearities on speech encoding in the auditory nerve. *J. Acoust. Soc. Am,* **68:**3, 858-875.

van Schaik, A., Fragniere, E., and Vittoz, E. (1995). Improved silicon cochlea using compatible lateral bipolar transistors. In Tourestzky, D. *et al.*, (eds) *Advances in Neural Information Processing Systems 8,* Cambridge,Mass: MIT Press.

Seneff, S. (1988). A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing. *Journal of Phonetics,* **16:**1, 55-76.

Vittoz, E. A. (1985). The design of high-performance analog circuits on CMOS chips. *IEEE Journal Solid State Circuits,* **20:**3, 657-665.

Watts, L., Kerns, D. A., Lyon, R. F., and Mead, C. (1992). Improved Implementation of the Silicon Cochlea. *IEEE Journal Solid State Circuits,* **27:**5, 692-700.

Woodland, P. C., Odell, J. J., Valtchev, V., and Young, S. J. (1994). Large vocabulary continuous speech recognition using HTK. *ICASSP-94,* **2**, II/125-8.