INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Properties of Stochastic Perceptual Auditory-event-based Models for Automatic Speech Recognition

Su-Lin Wu

TR-95-023

May 1995

Abstract

Recently, physiological and psychoacoustic studies have uncovered new evidence supporting the idea that human auditory processes focus on the transitions between spoken sounds rather than on the steady-state portions of spoken sounds for speech recognition. Stochastic Perceptual Auditory-event-based Models (SPAMs) were developed by Morgan, Bourlard, Hermansky and Greenberg to take this new evidence into account for word models in speech recognition by machines. This paper details our efforts to build a speech recognition system based on some of the properties of SPAMs. Although not all aspects of the complete SPAM theory have been implemented, we did find that fairly good recognition is possible with a system that concentrates almost exclusively on the transitions between speech sounds. Additionally, we found that such a system enhanced the more conventional phoneme-based system, which emphasized recognition of steady-state sounds. This blended system performed better than either system alone, especially in the case of noise-obscured speech.

1.0 Introduction

The current state of the art in automatic speech recognition is constrained by several underlying assumptions that are questionable from an auditory perspective. Stochastic Perceptual Auditory-event-based Models (SPAMs) were developed as an attempt to avoid one of these, the assumption that speech models are a sequence of stationary segments of uncorrelated acoustic vectors. Morgan, Bourlard, Hermansky and Greenberg hoped that SPAMs would prove to be more robust under adverse conditions (e.g. noisy speech) than conventional models [MBHG].

With conventional models, research into recognition has mostly focused on the steady-state regions of an utterance. Empirical evidence suggests that these regions are not as important perceptually as transition regions to the human ability to discriminate between words that sound alike, the syllables “baa” and “daa” for example. Psychoacoustic experiments by Drullman and, earlier, by Furui indicate that information needed for correct identification is largely contained in spectral transitions [DFP][Furui]. This led us to believe that the fundamental speech unit of recognition should emphasize the transitions between steady-state regions, rather than focussing almost exclusively on the steady-state regions themselves, as is usually the case.

SPAMs are a sequence of Auditory Events or *avents*, separated by relatively stationary periods, denoted in the model as *non-perceiving states*, also referred to in this paper as “non-transitioning states” or “nts”. These more slowly varying periods are around 50 to 150 milliseconds in duration and represent speech sounds such as the slowly changing portion of vowels. Avents are elementary auditory decisions, presumably made in response to rapid change in the speech spectrum and amplitude. Avents were designed to more closely represent the cues of human perception as researchers understand them [Greenberg]. In this study, avents are assumed to occur at the boundary between two phones and can be viewed as responses to left-context-dependent phonetic onsets.

We built a recognition system based on avents to validate the SPAM idea of recognizing speech by focussing on transitions [MBGHW]. We have not implemented several parts of the complete SPAM theory, most notably the dependency of an avent on previous avents or on the elapsed time between avents. Also, the REMAP[1] procedure, which is more ideally suited for recognition with SPAMs, has not been substituted in for the usual dynamic programming step.

2.0 Methods

We used a conjunction of a variety of techniques to implement a system that would recognize spoken words based on avents.

1. REMAP is a new approach [BKM] that could potentially provide soft (probabilistic) targets for the transition over a region around the estimated onset time.

2.1 The Digits+ Speech Corpus

The speech recognition task we chose is the Digits+ corpus available at ICSI. It is composed of 200 speakers saying the words “one” through “nine”, “zero”, “oh”, “no”, and “yes”. Each word was recorded in isolation over a clean telephone line at Bellcore. For the additive noise in these experiments, we used automotive sound that was recorded over a cellular telephone. Noise was randomly selected from this source and then added to the clean speech waveforms [Tong].

We chose this task over others available because of its small size and simplicity and because the speech group at ICSI has already had considerable experience with this corpus. With just thirteen words, training and testing times for each experiment were more manageable and less demanding on computing resources than larger, continuous speech tasks. Because each word is isolated, no grammar or natural language model is necessary. A recognition system based on conventional phone units had already been developed and optimized for performance by Kristine Ma [Ma]; this helped us make a realistic evaluation of the avent-based recognizer’s performance. The Digits+ task is a minimal task well suited for developing new speech recognition systems while still large and complex enough to allow general conclusions to be drawn from the results.

2.2 Hybrid Hidden Markov Model - Multilayer Perceptron System

The speech recognition system developed for this work is based on the hybrid Hidden Markov Model, Multilayer Perceptron system in use at ICSI [BM], illustrated below in Figure 1. Acoustic information is processed by a feature extraction method and the result is used as input to a neural network. A simple three-layer fully connected neural network is used to classify frames of features into speech sound units (Figure 2). The neural network produces a probability for each output for every time frame of speech. Dynamic programming operates on this output from the neural network and uses the knowledge in

Hidden Markov Models of the relevant vocabulary to determine which word best matches the input data.

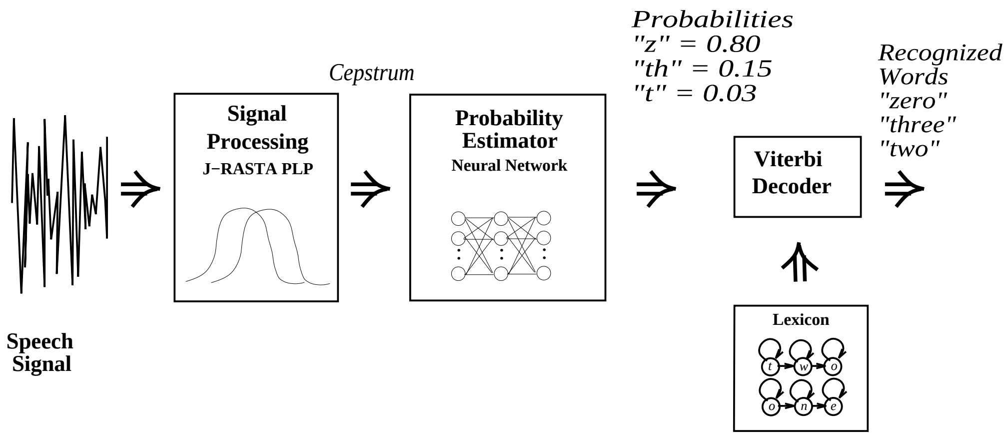

Figure 1: Hybrid HMM-MLP speech recognition system architecture.

We used the neural network simulator BoB, written by Phil Kohn [Kohn], to train the various neural networks. Approximately 10% of each training set was reserved as a cross-validation set, so that the network would not overtrain and be unable to generalize for new inputs.

The word models for the phoneme-based system are conventional Hidden Markov Models [DKP] The word models for the avent-based system are also Hidden Markov Models, though the models in the full SPAM theory are not true HMMs. Viterbi decoding was used to find the HMM with the highest likelihood, given the observational vectors from the MLP.

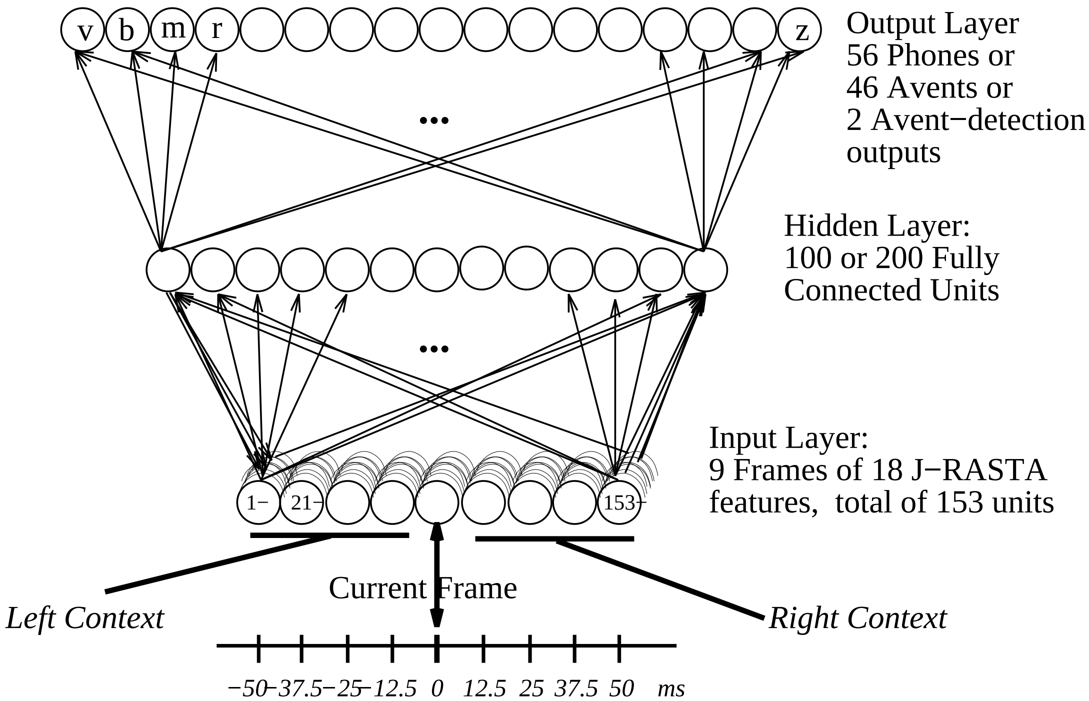

Figure 2: Neural network architecture.

2.3 Phone-based Recognition

As mentioned before, Kristine Ma implemented a phone-based recognition system for the Digits+ corpus. We duplicated her work because small changes, primarily to the phone set, were necessary to make the phone-based system comparable to our avent-based system.

2.3.1 Data

The stored waveforms of the Digits+ corpus were analyzed with the J-RASTA PLP feature-extraction process [KMHHT][Hermansky]. This produced eight features for every 25-millisecond frame of speech, where each frame overlaps 12.5 milliseconds of the next frame. Also, the process provides one value representing energy, and finally a “delta” feature for each of the aforementioned nine values. “Delta” features are approximate time derivatives. With deltas, each frame contains some information about change between it and its neighboring frames. Historically, delta features have improved recognition rates. In contrast, the energy feature actually handicaps recognition in experiments with realistic situations where the overall signal energy can vary considerably. For these experiments, 17 of the above 18 features were used, with the energy value left out.

There are approximately 54,733 frames of speech data in this database. Phonetic labels were generated through an automatic forced-alignment process. This is a procedure that uses dynamic programming to assign phonemes to frames given a fixed pronunciation order of sounds in the entire utterance. Because this is an automated procedure, the labels are not as accurate as hand-labels produced by human listeners. Spot-inspection of the labeled data shows that the onsets and offsets of the phones are often displaced by as much two or three frames.

Single pronunciation word models were constructed and the durations of the phones in the models were tuned with the automatic forced-alignment process mentioned above. A word model is illustrated below, in Figure 3, for comparison with the equivalent avent word model, to be discussed later.

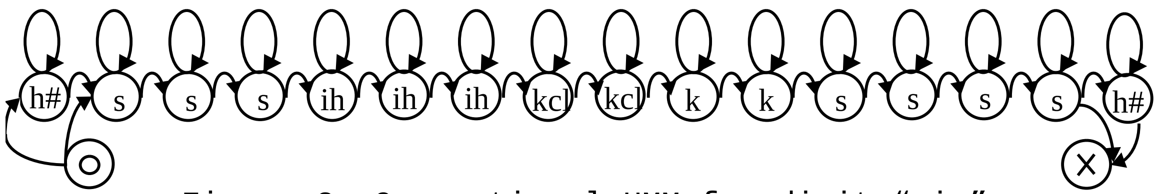

Figure 3: Conventional HMM for digit “six” for phone-based recognition systems.

These data and models were used to perform training and recognition.

In early experiments we tried to “bootstrap” the neural network with the much larger speech corpus NTIMIT before refining the training with the Digits+ data [Ma]. This improved the recognition performance of the system a modest amount; but not enough to warrant the additional time and computing resources necessary, so this technique was not pursued.

Four “jackknifed cuts” through the data were used, to smooth out anomalies due to the choice of training and test set. In the “jackknife” procedure, the Digits+ training set was divided into four equally sized portions. For each of the four “cuts”, one-fourth was reserved for testing and the remaining three-fourths composed the training set. In this way, all of the available data is eventually used as part of a test set [BM].

2.3.2 Training

Experiments showed that a 200-hidden unit multilayer perceptron neural network was the right size for the training data available. The neural network was trained on 1,950 of the 2,600 total number of words, where approximately a tenth of the training data was reserved as a cross-validation set. The remaining 650 words constituted the test set. A typical training progression is shown in Table 1. The final network weights are those that result from the epoch with the highest Cross-Validation Frame-level Correct percentage.

TABLE 1. 200 HU network trained on Digits+ data from training cut 1.

Epoch	Learning Rate	Training Frame-level Correct	Cross-Validation Frame-level Correct
0	0.008000	64.261566	72.367912
1	0.008000	83.212112	77.305862
2	0.008000	86.361115	79.020081
3	0.008000	88.133095	80.145828
4	0.008000	89.446999	79.339890
5	0.004000	89.198441	80.120247

2.3.3 Recognition

The trained network was used to classify each frame of the test set. For each frame, the neural network produced a probability for each phoneme the frame could represent. This probability is divided by priors (the frequencies of each output unit as calculated from the training set) to produce likelihoods [BM]. Using the HMM models for words, dynamic programming determined which word contained the highest likelihood path for each utterance in the test set. This output was scored and the error percentage was used for the comparisons in Section 3.

2.4 Avent-based Recognition

Our challenge was to generate suitable avent data and implement a system around the new sound unit that would effectively recognize the isolated words in the Digits+ corpus.

2.4.1 Data

We created training data for the avents from the training data for the phones. The phoneme labels were used to identify transitions from one phone to another in the waveforms. The frame just prior to the beginning of a new phone was automatically labelled as the transition. All other frames were labeled as “nts” for non-transitioning state, and mapped onto the non-perceiving state mentioned in SPAM literature [MBGH]. An example, the word “six” is shown in Figure 4. The waveform in the figure is labeled with phones, on the bottom row, and with the corresponding avent labels, on the upper row. The end of each sound is marked; the beginning is implicitly understood to be the end of the previous sound.

Out of the 54,733 frames of data available, 5,960 were labeled as transition frames, roughly 11%.

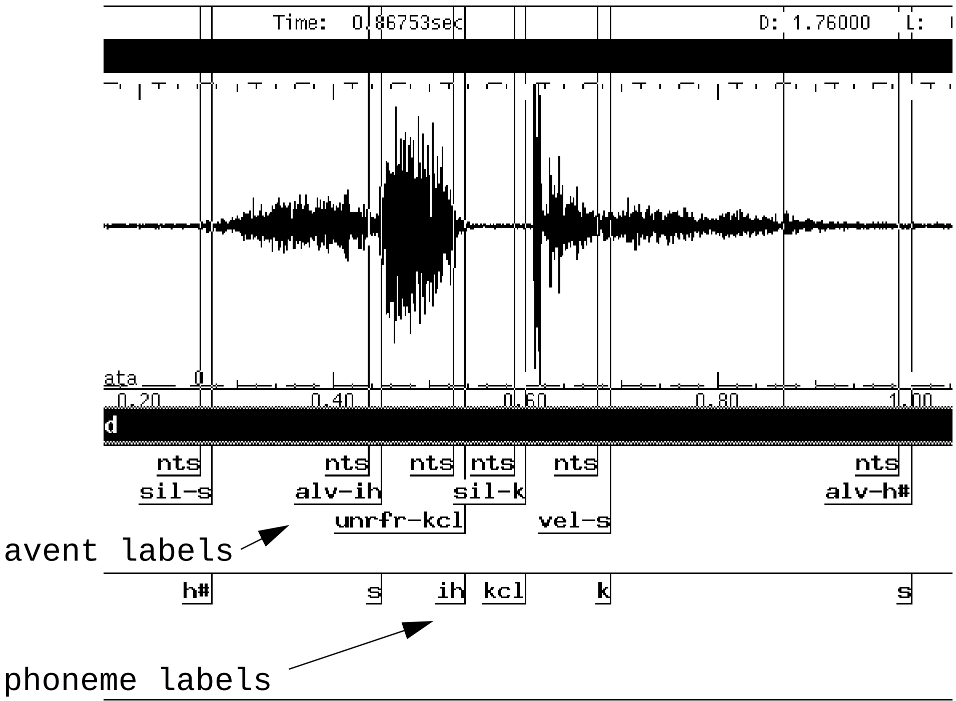

Figure 4: Waveform of “six” with avent labels and phoneme labels.

As mentioned previously, automatic labelling procedures are not perfectly accurate; we found avent labels to be as much as two or three frames off from where a human labeler would place them. This may not be a critical inaccuracy, because the neural network used is provided with nine frames of context, enough to include the relevant transition.

We created a new, modified lexicon, again based on the word models for the phoneme-based recognition system. An example is shown in Figure 5, for comparison to the word model for the phoneme-based system in Figure 3. Although not detailed in the illustration, a minimum duration requirement was specified in the avent-word models, because this was found experimentally to improve performance.

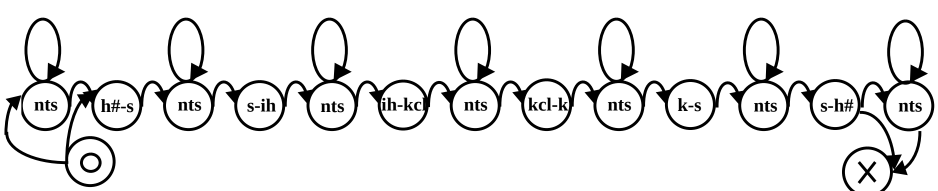

Figure 5: SPAM for the digit “six”.

2.4.2 Training

We discovered early in this study that a single net could not be trained to classify avents if the nts output was included, because the nts output encompassed 90% of every utterance. The neural network would tend to classify every frame as an nts frame. We tried subsampling the nts output label heavily in order to reduce the number of nts labels used for training to approximately the quantity available for the other avent labels. However, there were too few frames of the nts output used for training for the network to generalize and become proficient at detecting the nts frames. We resolved this by training two networks, one that was trained only on the avent labels where a transition was occurring and one network that was trained only to distinguish between the nts frames and avents in general, where all the avents were grouped into one category. Essentially, the first network is an avent-classifier and the second network is an avent-detector. It was still necessary, in this second network, to subsample the nts frames, but this formulation allowed more of the data to be used.

After some experimentation, we found that 100 hidden units was a good size for the hidden layer in the avent-detecting and avent-classifying networks. Note that the number of parameters in these two networks together is approximately equal to the number of parameters in the 200-hidden unit neural network used for phoneme classification in the earlier section.

The avent-classifying network was trained on about 10% of the data. While it learned to classify transitions reasonably well, we made several efforts to improve its abilities. Initially we used NTIMIT to "bootstrap" this network as well, and again the result didn't justify the expense in time and computing resources.

More effective was the technique of using a phoneme-trained neural network of approximately the same size to bootstrap the avent network. We trained a 100-hidden unit network on conventional phoneme labels and then trained it further on the beginnings of the phonemes, where the network received as input only the first frame of every phoneme. In this way we slowly retargeted the network for transitions by training on onsets. Then the network weights were used to initialize an avent network; for each avent, the weights associated with the phone corresponding to the right half of the avent were copied into the weights for that avent. The network was then trained further on avent data. This gradual retargeting procedure resulted in reducing the network's classification error significantly. The word-error rate from using this network was about a third lower than the word-error rate achieved by the neural networks we trained without this gradual retargeting process.

An example training progression for the whole avent system is shown in Tables 2, 3, 4, and 5. The percentages shown provide information about the network training, but the values do not necessarily correlate to the digit recognition ability of the complete system.

TABLE 2. 100 HU network trained on Digits+ phoneme data from training cut 1.

Epoch	Learning Rate	Training Frame-Level Correct	Cross-Validation Frame-Level Correct
0	0.008000	60.949879	70.423439
1	0.008000	80.699776	76.013817
2	0.008000	84.218147	78.150185
3	0.008000	86.243469	78.495590
4	0.004000	87.490852	78.802612

TABLE 3. 100 HU phone network trained additionally on onsets.

Epoch	Learning Rate	Training Frame-Level Correct	Cross-Validation Frame-Level Correct
0	0.008000	86.371872	82.573723
1	0.008000	88.783478	84.048256
2	0.008000	90.115738	85.790886
3	0.008000	91.158699	85.790886
4	0.004000	90.922348	85.924934

TABLE 4. 100 HU phone-onset network trained additionally on avents.

Epoch	Learning Rate	Training Frame-Level Correct	Cross-Validation Frame-Level Correct
0	0.008000	76.262451	74.213829
1	0.008000	83.797935	77.358482
2	0.008000	86.906738	77.735847
3	0.004000	87.497856	77.987419

TABLE 5. 100 HU network training on Digits+ nts data from training cut 1.

Epoch	Learning Rate	Training Frame-Level Correct	Cross-Validation Frame-Level Correct
0	0.008000	60.742035	63.672131
1	0.008000	68.765907	68.786880
2	0.008000	71.498398	69.967209
3	0.008000	73.438431	71.016388
4	0.008000	75.420219	71.278687
5	0.004000	75.002083	72.590157
6	0.002000	76.614220	72.590157

The outputs of the avent-detecting network and the avent-classifying network were combined as follows: the probability assigned by the neural network to the "avent-detected" output of the network in Table 5 was distributed over the avents, in proportion to the probability assigned to each avent by the avent-classifying network in Table 4. This combined output was used as input to the Viterbi decoder that then produced words. This system is illustrated in Figure 6 below.

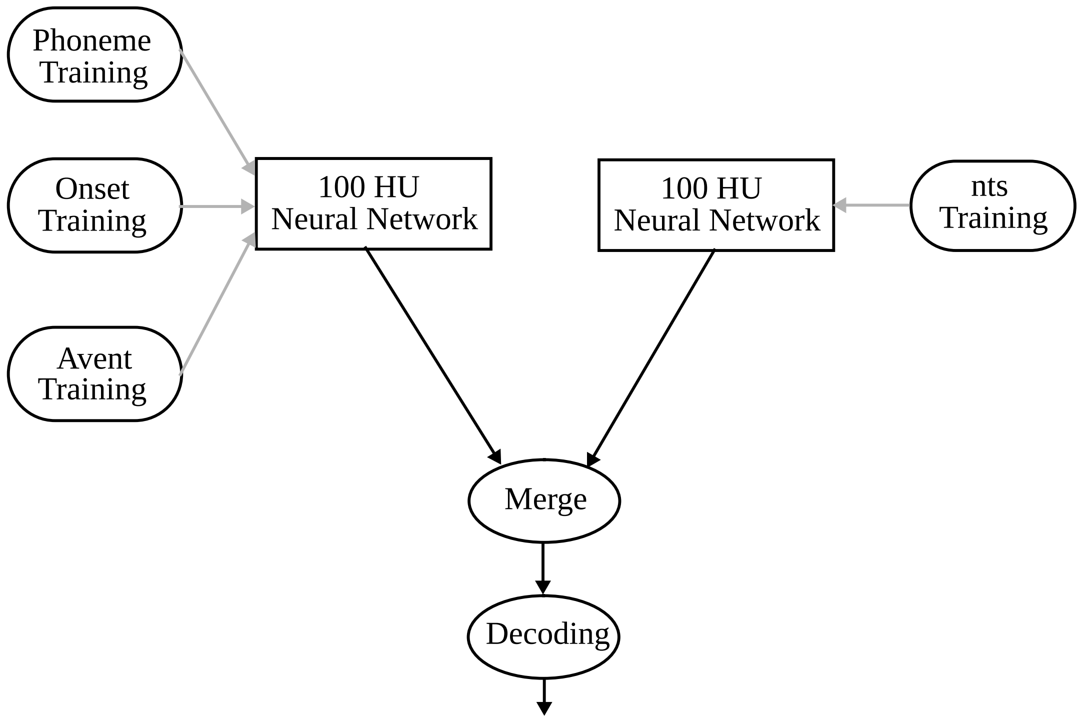

Figure 6: System overview of the avent-based recognition system showing training process.

2.4.3 Recognition

The word models for avent-based recognition are Hidden Markov Models where the avent states have no self-loop. That is, any path can stay in an avent state for at most one frame. Non-transitioning states do have self-loops, however.

Traditionally, priors are used to compensate for the neural network's tendency to favor the labels of frames it has seen more often, or, equivalently, to convert from posterior probabilities to data likelihoods (via Bayes' Rule [BM]). Division by priors was not fully implemented for the avent-based recognition system; that is, while avents-versus-nts training was done with an equilibrated training set (which should be equivalent to division by priors), we did not divide by priors for the avent classification categories. We believe that the system's performance would not be strongly affected by the utilization of priors because the training targets presented to the networks were reasonably balanced. An early experiment in which the merged avent probabilities were divided by priors produced some supporting evidence of this.

2.5 Combining Avents and Phonemes

Examination of the confusion matrices of both the phoneme-based system and the avent-based system showed that the errors and the types of errors that each system made were nearly orthogonal. We conjecture that this reflects the difference in the properties of the two recognition systems. For example, the phoneme-based recognition system seems to have more difficulty differentiating between “no”, “oh” and “zero” in the presence of noise than the avent-based system. Confusion matrices for our developmental experiments are shown in Tables 6, 7, 8, and 9.

TABLE 6. Confusion matrix[a] for 200 HU phoneme-based recognition system. Clean speech.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	45	1	1		1							1	1
yes		49											1
zero			50										
oh	1			47			2						
nine					49								1
eight						50							
seven							50						
six								49		1			
five					1				49				
four										50			
three											50		
two			1			2						47	
one													50

a. True words at left, recognized words at top

TABLE 7. Confusion matrix[a] for phone-based recognition system. Speech, 10db additive noise.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	34	3	2	2			4			1		3	1
yes		48				1	1						
zero		7	41					2					
oh	1			36		1	3		3	5		1	
nine					43		2	1	2		1		1
eight		1				49							
seven							48	2					
six						1		49					
five					2		1		47				
four										50			
three	1					5		1			41	1	1

TABLE 7. Confusion matrix[a] for phone-based recognition system. Speech, 10db additive noise.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
two						7	1	1				41	
one									1				49

a. True words at left, recognized words at top.

TABLE 8. Confusion matrix[a] for avent-based recognition system. Clean speech.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	48			1									1
yes		50											
zero			48			1						1	
oh				50									
nine	2				48								1
eight				1		49							
seven							50						
six	1							48		1			
five					1				49				
four				1						49			
three		1		1							45	2	1
two						1						49	
one					1								49

a. True words at left, recognized words along top.

TABLE 9. Confusion matrix[a] for avent-based recognition system. Speech, 10db additive noise.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	43	1	1	1	1		1					1	1
yes	2	46						1					1
zero	1	5	43			1							
oh	3			42			3		2				
nine	4	1			44								1
eight						48						2	
seven						1	45	2		1			1
six	1					5	1	42		1			
five					4				46				
four										50			
three	1	1				3					39	5	1
two	1				1	4				2		42	
one	1				3				2				44

a. True words at left, recognized words along top

The apparent independence of the strengths of each system led us to experiment with blending the two systems. From the dynamic programming (Viterbi) stage we can calculate a likelihood for the most probable path through every word, because our task only permitted isolated words. For every word in the Digits+ set we added the likelihood for a particular word from the phoneme-based system to the likelihood for the same word from the avent-based system. The likelihood value from the avent-based system is scaled by a constant factor to compensate for the difference in the observational values produced by the avent networks and the phoneme network. Then we selected the word with the best likelihood value over all for each utterance (Figure 7).

We chose the constant scaling factor for the avent values by performing a series of experiments with a single training cut. We determined that a value of "10" was optimal for that training cut, and that is the value we have used for testing across all of the training cuts. We note that this number is roughly equal to the average number of phone emission probabilities used for every avent probability.

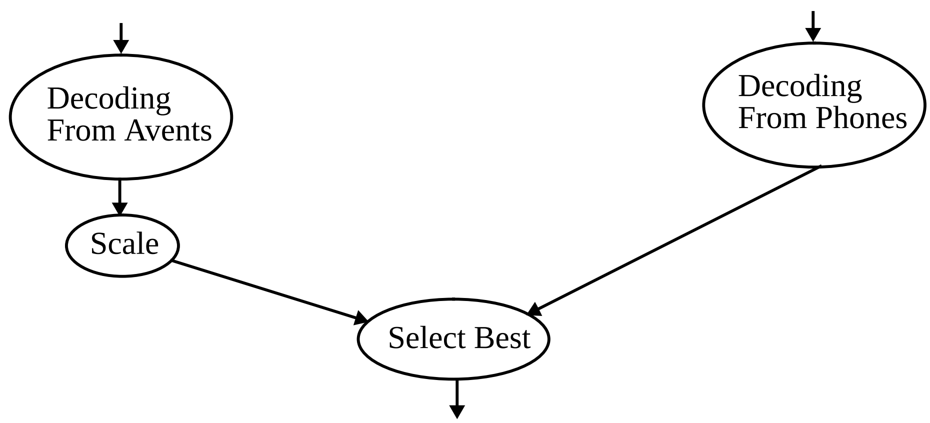

Figure 7: System overview of the blended recognition showing the merging of the values from the individual systems.

The blended result is significantly better for noisy speech then either system produces alone, while for clean speech the blended system is roughly equivalent to the phone-based system.

3.0 Results

The avent-based system, with merged output from the avent-detecting and avent-classifying network, recognized clean speech at roughly twice the error-rate of the phoneme-based system. For speech with 10db additive noise, the avent-based system performed about as well as the phoneme-based system. These numbers are summarized in Tables 10 and 11 below .

Table 10: Phone-based Recognition System Word-error Percentages

	1	2	3	4	Average
clean	2.3%	1.5%	1.4%	2.0%	1.8%
w/noise[a]	11.4%	10.2%	10.3%	11.5%	10.9%

a. with 10db additive noise.

Table 11: Avent-based System, Word-error

	1	2	3	4	Average
clean	2.9%	4.0%	3.1%	4.2%	3.6%
w/noise[a]	11.7%	9.4%	9.5%	11.8%	10.6%

a. with 10db additive noise.

The blended system, with both phoneme and avent likelihoods, produces better results than either system alone, as shown in the following table. However, only in the additive noise case is this a strong effect. Assuming that the distribution of correct answers is a binomial distribution and using a normal approximation to the binomial for calculation, we find that the above differences in the noisy speech case due to the blended system is significant with a p-value less than 0.01.

Table 12: Blended Avent-phone System, Word-error

	1	2	3	4	Average
clean	1.7%	0.9%	1.4%	2.3%	1.6%
w/noise[a]	9.4%	7.4%	5.7%	8.3%	7.7%

a. with 10db additive noise.

The blended system makes use of about twice the number of parameters as the 200-Hidden Unit phoneme-based recognition system. To verify that the improvement in the word-error scores was not due merely to the extra parameters, we trained a 400-HU network and performed phoneme-based recognition. The resulting scores did not differ significantly from the scores from the phoneme-based system with the 200-HU network for either clean speech or speech with 10db additive noise. This result indicates that the improvement in performance noted above probably comes from the different basis of the two systems rather than from the simple increase in the number of parameters.

4.0 Discussion

There are a number of open questions, not addressed by the experiments discussed here.

Transitions have been treated in this work as lasting exactly 12.5 milliseconds. This is a gross simplification. Hard targets such as these were necessary for practical considerations in pilot systems. More accurate would be the “soft” targets under development as part of the REMAP theory by Konig, Bourlard and Morgan [BKM]. Because these targets will more accurately portray transitions, we hope that they may improve the avent-based system.

Diphthongs are an open question with regards to this work. Diphthongs, in which a speaker glides from one vowel to another in a single syllable, are technically two distinct regions with a transition in the middle [Edwards]. For example, the diphthong “ay” in “nine” is composed of the sounds “aa” and “iy”. The issue is not so clear cut, however, in real speech. Linguists disagree on which sounds are diphthongs and which are not. Where exactly to put the transition from one vowel to another is also difficult, because the transition is often gradual and the second part is often entirely transitional in nature. In initial experiments, we tried putting the transition in the middle of the duration of each diphthong. To compare with the phone case, we also reduced diphthongs in the phoneme-based recognition system to their constituent vowel parts. This experiment had the effect that the phoneme-based system’s error rate almost doubled. Clearly, treating the diphthong as a single sound rather than as its two constituent sounds is important. We think that the large context window (nine frames) that the neural network has as input allows the network to be able to see the part of the spectrogram that changes in the middle of diphthongs, because the shift from one vowel sound part to another is likely to occur less than 4.5 x 12.5 milliseconds or less than 56 milliseconds from the onset of the diphthong.

When we modified the avent-based recognizer to treat diphthongs as single sounds, recognition performance improved modestly, not nearly as much as with the phoneme-based recognizer. It is not clear why the improvement should be different in the two cases. From examination of the confusion matrices, we tentatively conclude that the crude method of assigning a transition point to the exact middle of each diphthong is too inaccurate and further experimentation is needed here to determine a better labelling.

In another area, we noticed that the neural network had significantly less success learning to classify the initial fricatives at the start of “three” and “zero” than it did learning other avents. In cross-validation frame-level performance (during training), the “silence-th” and “silence-z” avents are correctly recognized in about 30% of the frames, whereas the average is about 80%. Inspection of a few of the waveforms suggests that those words are sometimes spoken with initial low energy, and sometimes with a high-energy burst. The low energy onset could be more difficult for the network to learn to recognize. This experience suggests that it may be important to distinguish steady-state speech from steady-state silence, or implement multiple pronunciation avent models. An initial experiment in this area was inconclusive. It is likely that more training examples are needed.

One avenue of attacking these problems with the avent-based system is to try a forced alignment to see if these problems are exacerbated by poor labels. Early in our experiments, we tried a trial run of the forced alignment process, but the performance of the resulting system actually deteriorated. This result is inconclusive; more effort is needed in this area.

5.0 Summary and Conclusions

This work is a pilot study of the value of focussing recognition algorithms on the transitions between speech sounds, rather than on the steady-state regions. We found that a recognition system based on avents performs almost as well as a phoneme-based system, particularly with noisy speech, even though the avent-based system we implemented is non-optimal in many ways. The recognition system based on transitions makes different errors from the phoneme-based system and was used to improve the recognition performance of this conventional system. The result is a significant improvement in performance rate in the case of noisy speech, more than either system can achieve alone.

We conclude from these experiments that decision states based on transitions are a viable unit of speech recognition with desirable properties not present in the conventional sound unit, phonemes. The relative success of the combination of the two systems suggests that the currently simplified version of the avents used in the recognizer may not be rich enough to improve performance levels by itself; it needs assistance from the system based on steady-states. This work merely scratches the surface of the many issues in this area.

6.0 Future Work

Because only 10% of the training data is classified as avent frames, a natural next step is to experiment with dropping unnecessary frames. This has the effect of speeding up training and testing by a scalar factor. It may also facilitate our training of the avent-detecting neural network by reducing the imbalance between the avent-detected and nts classes more meaningfully then by simple random subsampling. This an idea has been mentioned in published literature at least since 1978 [TD]. In several more recent papers, variable frame rate analysis techniques are able to eliminate 50% of the total number of frames without significant loss in performance [MB][LV]. Some of these papers mention that when low percentages of frames were eliminated, the performance of the recognizer increased. This supports the idea that conventional speech recognition systems focus unduly on the steady-state regions, because eliminating them sometimes improves performance.

The criteria for dropping a frame in the Le Cerf and Van Compernolle method is the following: if the norm of the derivatives of the features in a frame is less than some threshold, the frame is eliminated. We plan to use this criterion to conduct our own analogous experiment with our avent-based system. The inaccurate labelling presents a problem in that the frame-dropping process might result in important labels being lost. A forced alignment is necessary and might correct the problem by relabelling the training data. We will train the same avent-based and phone-based recognition systems on the filtered data. Although much of this methodology is in place, we have no results to report as yet.

Avents clearly show considerable potential for modeling human speech perception better than steady-state sounds units, in several ways. Much additional research will be necessary to effectively utilize the avent idea.

7.0 Acknowledgments

Thanks to Dan Jurafsky for the artwork in Figures 1 and 2. Also, many thanks to Kristine Ma who provided much of the speech infrastructure for these experiments and valuable help with ICSI's speech software. Steve Greenberg generously shared a wealth of knowledge with the author, which the author appreciates. John Wawrzynek contributed much advice and helpful reviews of this paper. Morgan provided the background, impetus and guidance for this work and made his experience and ideas readily available throughout, for which the author is immensely grateful.

The author thanks ICSI for continuing support.

This paper is based upon work supported under a National Science Foundation Graduate Research Fellowship. Any opinions, finding, conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

This work was sponsored, in part, under the Joint Services Electronics Program, Contract Number F49620-93-C-0014 and in part by the Office of Naval Research, URI Grant No. N00014-92-J-1617.

8.0 Appendix

TABLE 13. Phoneme Label Set[a]

Index	Phone	Broad	Example	Description
0	b	lab	buy	voiced bilabial stop
1	d	alv	dog	voiced alveolar stop
2	g	vel	goat	voiced velar stop
3	p	lab	pie	voiceless bilabial stop
4	t	alv	tom	voiceless alveolar stop
5	k	vel	cat	voiceless velar stop
6	dx	alv	writer	alveolar flap
7	bcl	sil		voiced bilabial stop closure
8	dcl	sil		voiced alveolar stop closure
9	gcl	sil		voiced velar stop closure
10	pcl	sil		voiceless bilabial stop closure
11	tcl	sil		voiceless alveolar stop closure
12	kcl	sil		voiceless velar stop closure
13	jh	alv	gym	voiced palatal affricate
14	ch	alv	chase	voiceless palatal affricate
15	s	alv	so	voiceless alveolar sibilant
16	sh	alv	share	voiceless palatal sibilant
17	z	alv	zebra	voiced alveolar sibilant
18	zh	alv	pleasure	voiced palatal sibilant
19	f	lab	fin	voiceless labiodental fricative
20	th	lab	thigh	voiceless dental fricative
21	v	lab	vain	voiced labiodental fricative
22	dh	lab	the	voice dental fricative
23	m	lab	my	bilabial nasal
24	em	lab	bottom	bilabial nasal, syllabic allophonic variation
25	n	alv	not	alveolar nasal
26	nx	alv	dinner	nasal flap
27	ng	vel	sing	velar nasal
28	en	alv	button	velar nasal, syllabic allophonic variation
29	l	alv	limb	alveolar lateral
30	el	alv	bottle	alveolar lateral, syllabic allophonic variation
31	r	r	right	retroflex approximate
32	w	round	when	bilabial glide
33	y	unrfr	yet	palatal glide
34	hh	vel	hot	voiceless glottal fricative
35	hv	vel	ahead	voiced glottal fricative
36	iy	unrfr	feet	high front unrounded long

TABLE 13. Phoneme Label Set[a]

Index	Phone	Broad	Example	Description
37	ih	unrfr	fit	high front unrounded short or lax
38	eh	unrfr	pet	mid front unrounded short or lax
39	ey	diph	fate	eh -> iy
40	ae	unrfr	fat	low front unrounded
41	aa	unrbk	father	low back unrounded
42	aw	diph	how	aa -> uw
43	ay	diph	pie	aa -> iy
44	ah	unrbk	but	mid central unrounded stressed
45	ao	unrbk	caught	low back rounded
46	oy	diph	boy	ao -> iy
47	ow	round	boat	disputably a dipthong, -> uh
48	uh	unrbk	book	high back rounded
49	uw	unrbk	boot	high back rounded short or lax
50	er	r	bird	rhotacized mid central vowel
51	axr	r	butter	rhotacized mid central short vowel
52	ax	unrbk	about	mid reduced
53	ix	unrfr	debit	high reduced
54	h#	sil	(silence)	silence
55	q	sil	‘oh	glottal stop

a. Adapted from material from CSLU’s Labelling Guide [LM] and work by Gary Tajchman.

TABLE 14. Avent Label Set for Digits+

Index	Label	Index	Label
0	ntst	23	v-ix
1	h#-n	24	ix-n
2	n-ow	25	s-ih
3	ow-h#	26	ih-kcl
4	h#-y	27	kcl-k
5	y-eh	28	k-s
6	eh-s	29	h#-f
7	s-h#	30	f-ay
8	h#-z	31	ay-v
9	z-ih	32	v-h#
10	ih-r	33	f-ao
11	r-ow	34	ao-r
12	h#-q	35	r-h#
13	q-ow	36	h#-th
14	n-ay	37	th-r

TABLE 14. Avent Label Set for Digits+

Index	Label	Index	Label
15	ay-n	38	r-iy
16	n-h#	39	iy-h#
17	h#-ey	40	h#-t
18	ey-tcl	41	t-uw
19	tcl-h#	42	uw-h#
20	h#-s	43	h#-w
21	s-eh	44	w-ah
22	eh-v	45	ah-n

9.0 References

[BM] H. Bourlard, and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[BKM] H. Bourlard, Y. Konig, and N. Morgan, “REMAP: Recursive Estimation and Maximization of A posteriori Probabilities”, ICSI Technical Report TR-94-064, 1994.

[DPH] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals, Macmillan Publishing Company*, New York 1993.

[DFP] R. Drullman, J. Festen, and R. Plomp, “Effect of temporal smearing on speech reception”, Journal of the Acoustical Society of America, 95 (2), February 1994.

[Edwards] H. Edwards, *Applied Phonetics: The Sounds of American English*, Singular Publishing Group, San Diego 1992.

[Furui] S. Furui, “On the Role of Spectral Transition for Speech Perception”, Journal of the Acoustical Society of America, 80, (4), pages 1016-1025, 1986.

[Greenberg] S. Greenberg, “Auditory Processing of Speech”, *Principles of Experimental Phonetics*, Chapter 10, N. Lass (editor), St. Louis: Mosby.

[KMHHT] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, G. Tong, “Integrating RASTA-PLP into Speech Recognition”, IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1994, Adelaide, South Australia, pages I421-I424.

[Hermansky]Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis of Speech”, Journal of the Acoustical Society of America, April 1990, Pages 1738-1752.

[Kohn] Phil Kohn at ICSI. Private communication.

[LV] P. Le Cerf and D. Van Compernolle, “A New Variable Frame Rate Analysis Method for Speech Recognition”, IEEE Signal Processing Letters, December 1994, pages 185-187.

[LM] T. Lander and S. T. Metzler, “The CSLU Labeling Guide”, Center for Spoken Language Understanding, Oregon Graduate Institute, For Labeled Data Release 2.0, February 2, 1994.

[Ma] K. Ma, “Applying Large Vocabulary Hybrid HMM-MLP Methods to Telephone Recognition of Digits and Natural Numbers”, Masters Thesis, UC Berkeley, Forthcoming in Spring 1995.

[MB] H. Murveit and R. Brodersen, “An Integrated-Circuit-Based Speech Recognition System”, IEEE Transactions on Acoustics, Speech, and Signal Processing, December 1986, pages 1465-1472.

[MBGH] N. Morgan, H. Bourlard, S. Greenberg, H. Hermansky, “Stochastic Perceptual Auditory-Event-Based Models for Speech Recognition”, Proceedings ICSLP, pages 1943-46, Yokohama, Japan, 1994.

[MBGHW] N. Morgan, H. Bourlard, S. Greenberg, H. Hermansky, and S.-L. Wu, “Stochastic Perceptual Models of Speech”, IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Detroit, Michigan, 1995..

[TD] C. C. Tappert and S. K. Das, “Memory and Time Improvements in a Dynamic Programming Algorithm for Matching Speech Patterns”, IEEE Transactions on Acoustics, Speech, and Signal Processing, December 1978, pages 583-586.

[Tong] G. Tong, “Combating Additive Noise and Spectral Distortion in Speech Recognition Systems with JAH-RASTA”, Masters Thesis, UC Berkeley, Spring 1994.