DIGIT RECOGNITION WITH STOCHASTIC PERCEPTUAL SPEECH MODELS

Nelson Morgan [‡,*] *Su-Lin Wu* [‡,*], *and Hervé Bourlard* [‡,†],

[‡] International Computer Science Institute, Berkeley, CA
[†] Faculté Polytechnique de Mons, Mons, Belgium
[*] University of California at Berkeley, Berkeley, CA
Emails: morgan, sulin, bourlard@icsi.berkeley.edu

ABSTRACT

We have recently developed a statistical model of speech that focuses statistical modeling power on phonetic transitions. These are the perceptually-dominant and information-rich portions of the speech signal, which may also be the parts of the speech signal with a better chance to withstand adverse acoustical conditions. We describe here some of the concepts, along with some preliminary experiments on digit recognition. These experiments show that the new models, when used in combination with our more standard models, can significantly improve performance in the presence of noise.

1. BACKGROUND

In [5] we reported the development of a statistical model of speech that incorporates some simple temporal properties of speech perception. The primary goal of this theoretical development was to avoid a number of current constraining assumptions for statistical speech recognition systems, particularly the model of speech as a sequence of stationary segments consisting of uncorrelated acoustic vectors. In the new model, speech was viewed from the perceiving side as a sequence of Auditory Events (**Avents**), which are elementary decisions that occur at some point when the spectrum and amplitude are rapidly changing (as in [3]). **Avents** are presumed to occur about once per phone boundary, and thus are modeled as being separated by relatively stationary periods (ca. 50-150 ms). The statistical model uses these **Avents** as fundamental building blocks for words and utterances, separated by states corresponding to the more stationary regions. In order to focus the statistical power on the rapidly-changing portions of the time series, all of the stationary regions are tied to the same non-**Avent** class. Markov-like recognition models use Avents as time-asynchronous observations. Discriminant models are trained to distinguish among all classes (including the non-Avent class). In the full embedded procedure, the training data is automatically aligned using dynamic programming, and the discriminant system (e.g., a neural network) is trained on the new segmentation. These two steps are iterated, as discussed in [2], and are guaranteed to converge to a local minimum of the probability of error (on the training set). This process should focus modeling power on the perceptually-dominant and information-rich portions of the speech signal, which may also be the parts of the speech signal with a better chance to withstand adverse acoustical conditions. We named this new framework the Stochastic Perceptual Auditory-event-based (Avent) Model, or SPAM.

Figure 1 shows a SPAM for the word "six". Note that all of the states with self-loops are unlabeled, representing non-Avent frames. The intervening states last only for one frame, and correspond to particular Avents.

We refer the reader to [5] for theoretical background on this approach. SPAM recognition is based on a computation of global posteriors based on the following local acoustic probabilities:

$$p(q_\ell^n | q_k^{n-\Delta(n)}, \Delta(n), X_{n-d}^{n+c}), \quad \left\{ \begin{array}{l} \forall \ell = 0, 1, \ldots, K \\ \forall k = 1, 2, \ldots, K \end{array} \right\} \quad (1)$$

where q_k^n refers to **Avent** q_k occurring at time n, q_0 refers to the non-Avent state, $n - \Delta(n)$ corresponds to the previous time index for which an **Avent** had been perceived, (i.e., the last time index $n - \Delta(n)$ for which a $q_k^{n-\Delta(n)}$ was perceived with $k \neq 0$), and X_{n-d}^{n+c} is a sub-sequence of acoustic vectors that is local to the current vector x_n, extending d frames into the past and c frames into the future. In principle, this probability can be estimated by a neural network trained with targets associated with Avent labels (or Avent probabilities, as in [1]), and with inputs representing the previous Avent, the time back to that Avent, and a local window of acoustic vectors. For the purposes of this paper, we have only implemented the system with acoustic inputs, so that we are actually evaluating

$$p(q_\ell^n | X_{n-d}^{n+c}). \quad (2)$$

2. EXPERIMENTAL METHODS

Theoretical work has continued in the form of new training procedures for transition-based systems, and is reported elsewhere (see [1]). This summary is, however, intended to report work in progress in incorporating SPAMs in a word recognition system, namely one for classifying digits and simple control words spoken in isolation over the telephone. The data base, originally recorded at Bellcore, is one that we have used in the past for experimentation with new front ends, and due to the variability of hand sets and speaker pronunciation over the telephone is reasonably difficult despite its limited vocabulary. Our current best score on this test set for jackknifed tests is about 1.4% error with our best phone-based system, which consists of a neural network to estimate probabilities for a context-independent hidden Markov model with a single density per phone.

In a series of experiments that we performed over the last few months, we replaced phones as the basic units with Avents, and we trained multilayer perceptrons to

Figure 1: SPAM for the digit "six".

discriminate between these units. While we tried several slight variants for the choice of Avent categories, we ended up with diphone-like units, that is one class per possible pair of phones that occurred in the lexicon of zero through nine, plus "oh", "no", and "yes". This was a total of 24 phones, and 45 diphone-like units. Note that the units differed from diphones in that a single frame was designated as the Avent frame (and modeled by a single state with no self-loop), and all others around each transition were labeled as non-Avents. This may seem counterintuitive, since the choice of the Avent frame was often arbitrary (although it came from the results of embedded Viterbi segmentations for a phone-based system). However, we viewed it as a reasonable first attempt. In later systems we intend to train frames with continuous probabilities of being a state transition as learned by the REMAP procedure [1].

As noted above, for the experiments reported in this paper we did not learn a conditional dependence between the previous Avent or the time between a current frame and the previous Avent, although these terms are proposed in the SPAM theory [5]. Therefore the training is quite comparable to the approach used in our standard hybrid HMM/neural network approach, except that our units are Avents, which means that most frames will be hypothesized to correspond to a class that will be tied over all phonetic units. In practice, because of the difficulty of learning Avents when most frames are not labeled as such, we actually trained two networks separately: one to distinguish between Avents and non-Avents, and the other to distinguish between Avents. The former net was not trained on all frames, but rather on an equal number of Avent and non-Avent frames. Thus this net was not trained to give the true posterior probability of having an Avent, but rather an approximation that appeared to give us a better overall performance than a network that was trained on all the frames. The other network was trained in the usual fashion (i.e., using all frames with categories to be classified, resulting in a posterior estimator), except that it only learned to generate probabilities for frames pre-classified as Avents. Each net was a multilayer perceptron with a single hidden layer containing 100 sigmoidal units. The input for each was 9 feature vectors (the current frame and 4 from the immediate past and future).

We mention in passing that the network training was done in stages to provide a bias based on a greater number of frames: first, the phone network was used to initialize a net trained to recognize onset frames that were labeled by the context-independent phone class. This latter net was then used to initialize the net that discriminated between Avents which for the purposes of this experiment can be viewed as left-context-dependent phonetic onsets.

A second net was trained for phone discrimination according to our established hybrid procedure [2], using 200 hidden units and the same input features and training data. We gave this network the apparent advantage of having twice the number of hidden units because we wanted to have a comparable number of overall parameters with the two techniques, and for the SPAM case the size of Avent/non-Avent discriminator was close to that of the Avent classification net (since most parameters were in the input-to-hidden connections).

noise condition	phones	Avents	Combined
clean	1.8%	3.6%	1.6%
noisy	10.9%	10.6%	7.7%

Table 1: Error rates, isolated digits plus "oh", "no", and "yes", recorded over public-switched telephone network; noisy case includes artificially added car noise for a 10 dB SNR.

In order to explore robustness to additive noise, we also experimented with adding automobile noise recorded over a cellular telephone, yielding a final SNR of 10 dB (in terms of average power ratio). Features used were JRASTA-PLP-8 cepstral coefficients 1-8, and their temporal derivatives, and the derivative only for the 0th coefficient (log energy). This is a feature set that was designed for robustness to additive and convolutional noise, but which sometimes increases error for "clean" or matching train and test conditions. We have hypothesized in the past that this increased error was at least partially due to the stationarity assumptions that were built in to the phone-based HMMs.

3. RESULTS AND DISCUSSION

The results on a test set (distinct from an another part of the corpus that we used for development) are shown in Table 1.

As shown in Table 1, the phone-based system, which has been optimized for a number of years, has about half the error rate of the new approach for the clean digits. However, for the noisy case, the performance of the two systems was comparable. Note that each hypothesized state path for a word typically had only about 5 frames that used distances (negative log probabilities) from a presumed Avent, as opposed to the state paths for a phone-based system that would be using class probabilities for each of the roughly 50 frames in a word.

Examination of the confusion matrices of both the phone-based system (Table 2) and the Avent-based system (Table 3) 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 phone-based recognition system seems to have more difficulty differentiating between "no", "oh" and "zero" in the presence of noise than the Avent-based system.

The apparent independence of the strengths and weak-

nesses of each system led us to experiment with blending the two systems. As this was an isolated word task, the likelihood could be calculated (with the Viterbi algorithm) for the most probable path through every word model. The word model likelihoods could then be rescored by combining these word probabilities for the two approaches. This is equivalent to what is now commonly done to rescore an N-best list for continuous speech recognition (BB&N ref), except that in the isolated word task N can consist of all possible hypotheses. Combination is done in the log domain, and the scaling factor for the Avent-based system was determined through experimentation on a development set. Within the range that we tried, the best scaling for the SPAM probabilities was found to be a value of 10. We note in passing that this number is roughly equal to the average number of phone emission probabilities used for every Avent probability. (The other frames are scored as non-Avents, and as such do not discriminate).

The third column of Table 1 ("Combined") gives the resulting score for the jackknifed test sets (over a total of 2600 test words and 200 speakers). The combination does not seem to have a strong effect for the clean case, but at least it does not hurt. On the other hand, there is a strong improvement for the noisy case; roughly 30% of the errors were eliminated by incorporating the preliminary SPAM in this way. Assuming a normal approximation to a binomial distribution for the errors, this is significant at $p < .01$. Table 4 shows the confusion matrix for this system.

Of course, the combined system uses twice the number of parameters as either system alone. To verify that the improvement in word error was not merely due to the larger number of parameters in the merged system, we trained a phone-based system using a neural network with twice the number of parameters. The resulting scores did not differ significantly from the scores from the smaller phone-based system, either for clean or noisy speech. Thus, we conclude that SPAM, even in its current limited form, seems to provide some further noise robustness when used in combination with a phone-based hybrid HMM/ANN system.

4. FUTURE WORK

There are a number of experiments that we plan to do to extend this result. It might well be that the improvements we see can only occur when SPAM approaches are combined with traditional ones, since the use of multiple maps from acoustics to words is likely to improve robustness as well. Nonetheless, we are just beginning to develop SPAM methodology. As noted earlier, we have yet to incorporate dependence on the previous Avent or on the time back to it, though the theory suggests that both are required. We have not yet experimented with Avent net sizes, and our choice of Avent categories was only the most obvious (essentially biphone classes). We had difficulty with training the Avent/non-Avent network, and were forced to subsample the data, leading to probabilities that were skewed from the true posteriors we wished to estimate. This problem could be greatly reduced using results from REMAP research [1], since many frames would have some nonzero probability of being an Avent. Finally, as yet we have not incorporated any new signal processing that might be helpful in estimating Avent probabilities; we are currently relying on RASTA-PLP. This certainly does already emphasize transitions, as noted previously, but we believe that more work can be done at the front end as well.

5. CONCLUSIONS

Robustness to additive noise is a difficult problem for current speech recognition systems. In some cases good models can be developed for the interfering noise, but more generally it would be desirable to build systems that were inherently resistant to degradation from unknown additive non-speech sounds. Both systems that were reported here, i.e., phone-based and Avent-based, used J-RASTA processing [4], which in fact gave us significant robustness to the added noise due to its emphasis on transitional information. An earlier experiment using log-RASTA, which was less well-suited to the noisy situation, showed over three times the error rate. However, reducing the error further by front end signal processing alone may be quite difficult. In fact, while in principal the emphasis on spectral change can reduce the impact of constant or slowly-varying noise, it seems likely that modification of the statistical models to match this emphasis is important. This experiment is our first result that seems to confirm this expectation.

6. ACKNOWLEDGMENTS

Kristine Ma provided much assistance with the task infrastructure. Steve Greenberg continued to share his insights about the auditory perspective. Hynek Hermansky provided spiritual guidance. The work was partly sponsored by the Joint Services Electronics Program (JSEP) Contract No. F49620-93-C-0014, and the Office Naval Research, URI Grant no. N00014-92-J-1617. Su-Lin Wu was partially supported by a National Science Foundation Fellowship.

7. REFERENCES

[1] Bourlard, H., Konig, Y., & Morgan, N. (1994). "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities – Application to Transition-Based Connectionist Speech Recognition," *ICSI Technical Report TR-94-064, Intl. Computer Science Institute, Berkeley, CA.*

[2] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994

[3] S. Furui, On the Role of Spectral Transition for Speech Perception *J. Acoust. Soc. Am.* **80**, (4), 1016-1025, 1986

[4] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, and G. Tong: Integrating RASTA-PLP into speech recognition, In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Adelaine, Australia, 1994

[5] Morgan, N., Bourlard, H., Greenberg, S., & Hermansky, H. (1994). "Stochastic Perceptual Auditory-Event-Based Models for Speech Recognition," *Proc. Intl. Conf. on Spoken Language Processing* (Yokohama, Japan), pp. 1943-1946, September 18-22, 1994.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	125	9	19	14	4	2	20			1		4	2
yes		188	1			1	5	4		1			
zero		14	181					5					
oh	1			158		5	10		10	15		1	
nine	1	1			154		12	1	18		7	1	5
eight		1		1		196					1	1	
seven		1					188	9	1				1
six		3				1		196					
five					6		2		188		1		3
four			1				1	1	1	194			2
three	1		1			8	1	4			181	3	1
two			1			17	2	3		1	1	175	
one		1			1				3	1			194

Table 2: Confusion matrix from phone-based system, with 10db SNR. True words at left, recognized words at top.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	180	3	3	4	3		1					5	1
yes	7	186	1			1	1	1				2	1
zero	9	9	175		1	4						2	
oh	6			172		2	5		9	4	1		1
nine	27	1		3	157	1	1		3		2	3	2
eight			2	1		182		1			3	11	
seven	5					3	181	5		2	1	1	2
six	2					7	3	185		2		1	
five				9	12				173	2			4
four	2								1	193			4
three	3	1			1	6	1	1			176	10	1
two	3		3		1	7	2			3	1	180	
one	3			3	6				3	1			184

Table 3: Confusion matrix from Avent-based system, with 10db SNR. True words at left, recognized words at top.

	no	yes	zero	oh	nine	eight	seven	six	five	four	three	two	one
no	168	4	4	7	4	2	5	1		1		3	1
yes	1	193					2	2				1	1
zero	2	12	185			1							
oh	3			168		5	5		9	9			1
nine	6	1			172		5		7		5	1	3
eight	1			1		194					1	3	
seven	1	2					189	5			1		2
six						3	2	194		1			
five				1	6		1		187		1		4
four							1	1	1	194			3
three	1		1			6	1	2			183	5	1
two			1		1	11	1	2		1	1	182	
one	1			2	1				5				191

Table 4: Confusion matrix from combined SPAM-phone system, with 10db SNR. True words at left, recognized words at top.