STOCHASTIC OBSERVATION HIDDEN MARKOV MODELS

Carl D. Mitchell [1] *Mary P. Harper* [2] *Leah H. Jamieson* [2]

[1] AT&T Bell Laboratories,
600 Mountain Ave., Murray Hill, NJ 07974, cdm@att.research.com

[2] School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907-1285, {lhj,harper}@ecn.purdue.edu

ABSTRACT

Hybrids that use a neural network to estimate the output probabilitiy for a hidden Markov model (HMM) word recognizer have been competitive with traditional HMM recognizers when both use monophone context. While traditional HMM recognizers can easily utilize more context (e.g., triphones) to achieve better results, the size of the task has made it impractical to use phonetic context directly in the neural network front end of a hybrid. In this paper, we suggest a simple method to incorporate more context by modeling the phone distributions obtained from the neural network. This allows the HMM to easily handle stochastic pronunciations as well as errors from the neural network phone recognizer. The re-estimation equations are derived for the new model. Results for the Resource Management task illustrate that SOHMM increases recognition accuracy for the cases of no grammar, unigram grammar, and word pair grammar.

1. INTRODUCTION

Recently, many hybrid techniques have been introduced that combine Markov models with other more discriminative methods of acoustic classification. Typically, a neural network first classifies the acoustic signal into an intermediate class (e.g., a phone label) [1, 3, 4, 5]. Next, a method of time alignment is used to map the intermediate symbols to words or phones. Robinson and Fallside [4, 5] have developed one of the most successful hybrids for both phone and word recognition. They use a recurrent neural network that estimates the output probabilities and a Markov model to represent the word network. In this paper, the stochastic observation HMM (SOHMM) is introduced as a method of recognizing words from phone candidates, as shown in figure 1. The word network processes the observation distributions as with Robinson and Fallside's approach, but the phonetic variations are modeled stochastically. As an extension to a standard HMM, the SOHMM models the distribution of the observations rather than the observations themselves. The SOHMM includes both the word matcher of [4, 5] and the standard HMM approach as special cases.

2. STOCHASTIC OBSERVATION HMM

With a standard HMM, the set of model parameters, λ, is estimated by maximizing the likelihood that the HMM output, V, will match the given observation sequence. Thus the goal is to maximize $P_\lambda(V=O)$. For the SOHMM, the observations are unknown, and the goal becomes maximizing the likelihood that the HMM's output will match the output of an externally defined stochastic process, i.e., $P_\lambda(V=Y)$.

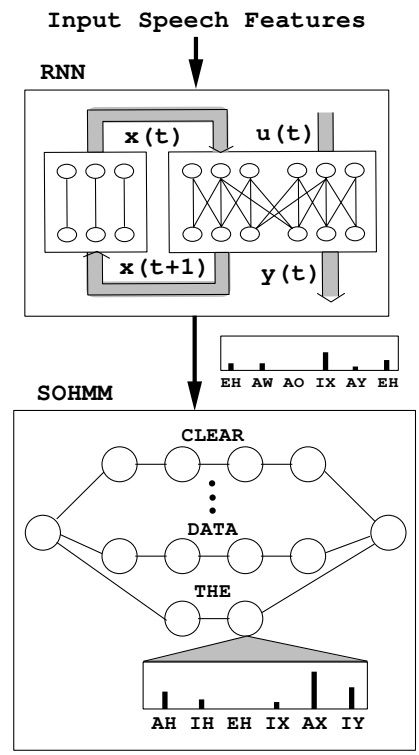

Figure 1. For each frame, the Recurrent Neural Network [4, 5] provides the phone probability distribution that is modeled by the SOHMM.

Let V be a discrete-time discrete-valued stochastic process representing the HMM output, and let Y be a discrete-time discrete-valued stochastic process that is defined external to the HMM. Assume that the distribution of y_τ for $t = 1, \ldots, T$ is given, and we want to determine the probability that the two stochastic processes are equivalent from $t = 1$ until $t = T$. Assume that y_τ and v_τ share the same domain. Let $S = \{s_0, s_1, \ldots, s_T\}$ denote a particular state sequence, and let $O = \{o_1, \ldots, o_T\}$ denote a particular HMM output sequence. For the stochastic observation case, O is referred to as the output sequence rather than the observation sequence since this sequence, like the state sequence S, is not directly observable. Instead, we "observe" the distribution of y_τ for each frame. As with the standard HMM, assume that v_τ is conditionally independent of v_t for $t \neq \tau$ given the state sequence. Also assume that y_τ is conditionally independent of y_t for $t \neq \tau$.

In addition to the output independence assumption, the SOHMM also requires that the output process be conditionally independent of the external random process, i.e., $P(V|Y) = P(V)$. This is a poor assumption for modeling the output from a frame-level phonetic classifier (such as a neural network). Clearly the acoustic class for a frame of speech is not independent of the phonotactic context (the SOHMM), but such an assumption is still more accommodating than the alternative, i.e., a Markov model word network. In section 4, a number of experiments are reported that support this assertion. The SOHMM allows pronunciation variations and phone classification errors to be modeled stochastically.

The derivation of the SOHMM re-estimation equations is very similar to the corresponding derivation for a standard HMM. The SOHMM forward and backward probabilities are defined as follows:

$$\alpha_t(j) = Pr(v_1 = y_1, \ldots, v_t = y_t, x_t = j|\lambda)$$
$$= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{i,j} \sum_{k=1}^{M} b_j(k) P(y_t = k)$$

$$\beta_t(i) = Pr(v_{t+1} = y_{t+1}, \ldots, v_T = y_T, x_t = i|\lambda)$$
$$= \sum_{j=1}^{N} a_{i,j} \beta_{t+1}(j) \sum_{k=1}^{M} b_j(k) P(y_{t+1} = k)$$

Using the inequality $\log\{x\} \leq x - 1$, it is easy to show that maximizing the likelihood $P_{\lambda'}(V=Y)$ as a function of λ' is equivalent to maximizing the following auxiliary equation as a function of the new SOHMM parameters, λ':

$$Q(\lambda, \lambda') = \sum_S \sum_O P_\lambda(X=S, V=O, Y=O)$$
$$\times \log\{P_{\lambda'}(X=S, V=O, Y=O)\}$$

where

$$P_\lambda(X=S, V=O, Y=O) = \prod_{t=1}^{T} a_{s_{t-1}, s_t} \prod_{t=1}^{T} b_{s_t}(o_t) \prod_{t=1}^{T} P(y_t = o_t)$$

Now $Q(\lambda, \lambda')$ can be written in terms of the forward-backward counts:

$$Q(\lambda, \lambda') = P_\lambda(V=Y) \sum_{i=1}^{N} \sum_{j=1}^{N} \log\{a_{i,j}\} C_a^\lambda(i,j)$$
$$+ P_\lambda(V=Y) \sum_{j=1}^{N} \sum_{k=1}^{M} \log\{b_{j,k}\} C_b^\lambda(j,k)$$
$$+ \sum_S \sum_O \sum_{t=1}^{T} P_\lambda(X=S, V=O, Y=O) \log\{P(y_t = o_t)\}$$

where

$$C_a^\lambda(i,j) = \frac{1}{P_\lambda(V=Y)} \sum_{t=1}^{T} \sum_O \sum_{\substack{\text{all S} \ni \\ s_{t-1} = i \\ s_t = j}} P_\lambda(X=S, V=O, Y=O)$$
$$= \sum_{t=1}^{T} \alpha_t(i) a_{ij} \beta_{t+1}(j) \sum_{k=1}^{M} b_j(k) P(y_t = k)$$

and

$$C_b^\lambda(j,k) = \frac{1}{P_\lambda(V=Y)} \sum_{t=1}^{T} \sum_{\substack{\text{all O} \ni \\ o_t = k}} \sum_{\substack{\text{all S} \ni \\ s_t = j}} P_\lambda(X=S, V=O, Y=O)$$
$$= \sum_{t=1}^{T} \alpha_t(i) a_{ij} \beta_{t+1}(j) b_j(k) P(y_t = k)$$

$Q(\lambda, \lambda')$ can be maximized using Lagrange multipliers, yielding the following re-estimation equations:

$$a_{i,j} = \frac{C_a^\lambda(i,j)}{\sum_{n=1}^{N} C_a^\lambda(i,n)}, \quad j = 1, \ldots, N$$

$$b_{j,k} = \frac{C_b^\lambda(j,k)}{\sum_{m=1}^{M} C_b^\lambda(j,m)}, \quad k = 1, \ldots, M$$

The proof that re-estimation of the SOHMM leads to a higher likelihood is very similar to the corresponding proof for a traditional HMM.

3. INTERPRETATION AS A MIXTURE MODEL

One interpretation of the SOHMM is as a mixture of probability mass functions where the mixture coefficients are determined by a random process that is external to the HMM. In order for the condition $V = Y$ to hold true, only those paths where $v = y$ are considered, as shown in boldface in figure 2.

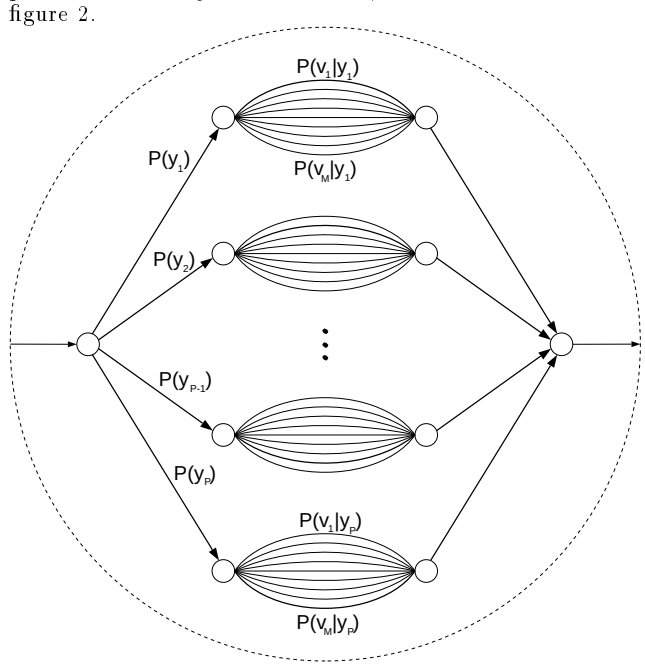

Figure 2. A state in the SOHMM.

The arcs exiting the leftmost node shown in figure 2 represent the mixture coeficients. All of the paths going into the rightmost node are summed during recognition. Note that $P(v|y) = P(v)$ given the assumption stated in the last section.

4. EXPERIMENTS

Experiments were conducted which used the SOHMM in conjunction with Robinson and Fallside's recurrent neural network phone recognizer [4, 5] on the Resource Management task. The neural network outputs the probability of each of the 68 labels in the phone inventory for each 16 ms frame of speech. Then the modified Viterbi algorithm is used to find the most likely path through the SOHMM given the phone distributions.

The training set consists of 2800 sentences. Each of the test subsets (Feb89, Oct89, Feb91, and Sep92) consists of 300 sentences, so the total number of test sentences is 1200. There is no overlap between the 40 speakers in the test set and the 280 speakers in the training set. For all of the experiments in this section, homophone confusions (e.g., "to" and "two") are not counted as errors.

4.1. No Grammar

Table 1 shows the results of the SOHMM word matcher when testing on Resource Management with a zero-gram grammar. With a zero-gram grammar, all words are equally likely and thus the test set perplexity is equal to the number of words in the vocabulary (992). SOHMM-0 refers to the model used by Robinson, i.e., the SOHMM output probabilities are 0 or 1, depending on the phonetic spelling stored in the pronunciation dictionary [4, 5]. SOHMM-6 uses stochastic pronunciations that were learned over 6 iterations of Baum-Welch re-estimation using the formulas given in section 2. In SOHMM-12, the HMM output probabilities are trained over 12 iterations. As can be seen from table 1, both the SOHMM-6 model and the SOHMM-12 model achieve slightly better %correct and about a 2% improvement in %accuracy compared to the Markov model word matcher, SOHMM-0.

4.2. Unigram Grammar

Table 2 shows the test set perplexity and results for the Feb89, Oct89, Feb91, and Sep92 test sets using a unigram grammar. After 12 iterations of training (SOHMM-12), it can be seen that the flexibility of the stochastic output HMM adds about 1.5% to %correct and nearly 3% to %accuracy, compared to the Markov model word matcher (SOHMM-0).

4.3. Word Pair Grammar

Table 3 shows the test set perplexity and results for the Feb89, Oct89, Feb91, and Sep92 test sets using a word pair grammar. Relative to SOHMM-0, SOHMM-12 increases %correct and %accuracy by 0.2% and 0.65%, respectively.

4.4. Duration

Three different types of duration models are shown in table 4: "No Dur", "Dur 30", and "Min Dur." "No Dur" refers to an implicit duration model where there was no minimum or maximum placed on the number of frames associated with a state. "Dur 30" denotes an explicit duration model where the maximum allowed duration was 30 frames [2]. Ferguson's nonparametric probability mass function was used for the explicit model for each state. "Min Dur" is also an implicit duration model, except that in this case paths with durations less than a prescribed minimum are not considered. Each phone is constrained so that its duration exceeds a user defined minimum, which ranges from one frame (16 ms) for plosives to as many as seven frames for some vowels.

The experimental results provided in tables 1, 2, and 3 are based on the minimum duration model, "Min Dur" [5]. Table 4 illustrates the usefulness of the minimum duration constraint for the SOHMM. The explicit duration model does not perform as well as the minimum duration model because of the inappropriateness of the output independence assumption.

5. APPLICATIONS OF THE SOHMM

The increased context used by the SOHMM improved %correct and %accurate for all of the grammars used in these experiments. Although the improvement comes at the cost of increasing the number of parameters, the SOHMM provides a simple way to utilize context for hybrid models that combine Markov models with a neural network classifier.

Since the SOHMM processes phone labels instead of acoustic observations, the training set can include sentences that were collected under different recording conditions. This could allow the SOHMM word matcher to benefit from a larger amount of training data than is available for training a set of acoustic models.

The two-stage approach can gracefully handle vocabulary words that were not seen while the acoustic models were trained. For example, the SOHMM is well-suited for name recognition. It is not practical to train an acoustic level HMM with all possible names. There could be tens of thousands of names for some tasks (e.g., library catalog access or directory assistance). A better approach would be to recognize the phones in a name (which will be well trained) and then model the phone recognition errors and pronunciation variations with a SOHMM.

6. CONCLUSION

This paper has presented the Stochastic Observation HMM. Key features of the SOHMM are the modeling of the distributions of the observations rather than the observations themselves, and the use of stochastic pronunciations. The results on the Resource Management task demonstrate that the SOHMM can improve word recognition accuracy. In no grammar, unigram grammar, and word pair grammar tests, the SOHMM with 12 iterations of training increased both percent correct and percent accuracy compared to a traditional Markov model word matcher. The most significant improvement occurred in the unigram case.

7. ACKNOWLEDGEMENT

This work would not have been possible without the recurrent neural network phone recognizer that was made available by Cambridge University [4, 5].

REFERENCES

[1] M. Fanty, P. Schmid, and R. Cole. City name recognition over the telephone. In *International Conference of Acoustics, Speech, and Signal Processing*, April 1993.

[2] J.D. Ferguson. Variable duration models for speech. In J.D. Ferguson, editor, *Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech*, pages 143–179, Princeton, NJ, 1980.

[3] N. Morgan and H. Bourland. Continuous speech recognition using multilayer perceptrons with hidden Markov models. In *International Conference of Acoustics, Speech, and Signal Processing*, April 1990.

[4] A. J. Robinson. Application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.

[5] A.J. Robinson and F. Fallside. A recurrent error propagation network speech recognition system. *Computer Speech & Language*, 5(3):257–274, July 1991.

Test Set	Feb89	Oct89	Feb91	Sep92	Total
Perplexity	992.0	992.0	992.0	992.0	992.0
SOHMM-0	81.61/78.84	80.37/77.42	81.88/78.10	75.26/70.54	79.77/76.22
SOHMM-6	82.19/80.63	80.40/78.87	81.88/79.59	76.16/73.19	80.15/78.07
SOHMM-12	82.43/80.67	80.85/79.14	81.92/79.39	76.32/73.43	80.38/78.16

Table 1: Results on the Resource Management corpus with a zero-gram grammar. Shown are %Correct/%Accurate for the SOHMM with 0, 6 and 12 iterations of training.

Test Set	Feb89	Oct89	Feb91	Sep92	Total
Perplexity	345.6	407.1	341.4	370.8	366.0
SOHMM-0	84.26/80.16	84.50/79.23	82.60/77.83	77.73/71.16	82.26/77.09
SOHMM-6	85.12/82.08	84.13/81.26	85.27/81.68	79.33/74.09	83.46/79.78
SOHMM-12	85.47/82.16	84.31/81.48	84.90/80.72	79.91/74.56	83.65/79.74

Table 2: Results on the Resource Management corpus with a unigram grammar. Shown are %Correct/%Accurate for the SOHMM with 0, 6, and 12 iterations of training.

Test Set	Feb89	Oct89	Feb91	Sep92	Total
Perplexity	≈ 60	≈ 60	≈ 60	≈ 60	≈ 60
SOHMM-0	95.67/94.92	94.93/94.19	95.37/94.32	91.75/90.07	94.43/93.38
SOHMM-12	96.02/95.74	94.82/94.52	95.25/94.52	92.34/91.32	94.61/94.03

Table 3: Results on the Resource Management corpus with a word pair grammar. Shown are %Correct/%Accurate for the SOHMM with 0 and 12 iterations of training.

Test Set	Feb89	Oct89	Feb91	Sep92	Total
Perplexity	345.6	407.1	341.4	370.8	366.0
No Dur	84.58/80.44	82.79/78.80	84.26/79.59	78.86/73.19	82.61/78.00
Dur30	85.24/81.57	83.64/80.03	84.54/80.27	78.90/73.15	83.08/78.76
Min Dur	85.47/82.16	84.31/81.48	84.90/80.72	79.91/74.56	83.65/79.74

Table 4: Results on the Resource Management corpus with a unigram grammar. Shown are %Correct/%Accurate for the SOHMM with different duration models.