ADVANCES IN HYBRID SVM/HMM SPEECH RECOGNITION[1]

A. Ganapathiraju

Conversational Computing Corporation
Redmond, Washington
aganapathiraju@conversay.com

J. Hamaker and J. Picone

Institute for Signal and Information Processing
Department for Electrical and Computer Engineering
Mississippi State University, Mississippi State, MS
{hamaker, picone}@isip.msstate.edu

ABSTRACT

In this paper, we describe our continuing work on the development of a hybrid SVM/HMM speech recognition system. The original hybrid system was evaluated on the OGI Alphadigits corpus and performed at 11.0% WER, compared to 11.9% for a triphone mixture-Gaussian HMM system. In a new set of experiments reported here, the hybrid system performs at 10.6% WER on the Alphadigits task using a simple score combination mechanism. On a large-vocabulary task, SWITCHBOARD, the hybrid system improves the performance over the baseline HMM-based system from 41.6% to 40.6% WER. This is the first time SVMs have been applied to a complex large-vocabulary task. Several oracle experiments are discussed which demonstrate the potential benefit of this approach over traditional HMM systems.

1. INTRODUCTION

Support vector machines (SVM) have had significant success on several classification tasks over the past few years [1]. Recently, they have been used as the core classifiers in speech and speaker recognition systems [2-5]. These systems perform at levels comparable if not better than traditional HMM-based systems. Through the introduction of the hybrid system in [2], we provided insight into many issues we face when transitioning from an HMM framework to an SVM framework. These include the application of temporal constraints to the static support vector classifier, generation of a posterior probability from the binary support vector classifier and balancing the need for a robust training set with computational requirements. However, this system was not used for large vocabulary tasks like SWITCHBOARD (SWB) [6] in which good acoustic modeling is critical.

In this paper we present results of a hybrid system on SWB. We also discuss the improvements achieved on the OGI Alphadigits task [7]. A key contribution of this paper is the construction of an oracle experiment that demonstrates the potential benefit of an SVM classifier compared to Gaussian probability distributions.

2. HYBRID ASR SYSTEM

Figure 1 shows a proposed hybrid architecture that can exploit the relative strengths of the traditional HMM approach and SVM classifiers. An important issue that had to be addressed in this hybrid system is the fact that SVMs output a distance measure, while the Viterbi decoding algorithm typically uses likelihoods or posterior probabilities. A simple approach to solving this problem of posterior estimation is to assume that the posterior takes the form of a sigmoid,

$$p(y|f) = \frac{1}{1 + \exp(Af + B)} \qquad (1)$$

and to directly estimate the sigmoid. In order to avoid severe bias in the distances for the training data, the free parameters, A and B are estimated on a cross-validation set. Once we have the posteriors, we can replace the Gaussians in the HMM system with the SVM classifiers.

Given the nature of the SVM classifier, the obvious way to introduce this technology into a speech recognition system

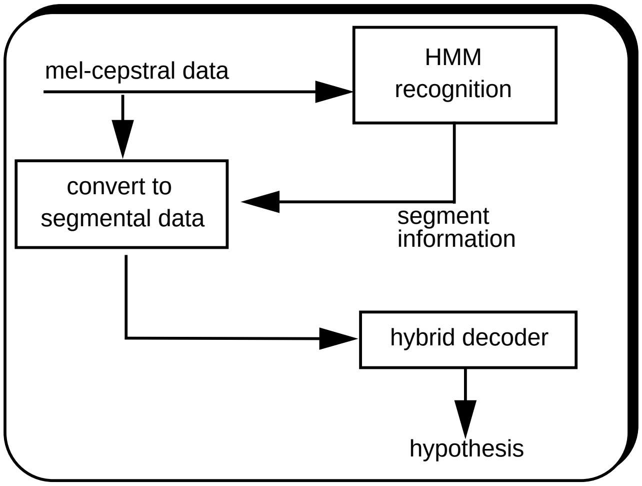

Figure 1: A hybrid system architecture

[1] This material is based upon work supported by the National Science Foundation under Grant No. IIS0095940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

would be to train the classifiers on frame level data and use them as the classifier in each state of an HMM. Since each classifier is trained as a one-vs-all classifier, the amount of training data is significant, and beyond current computational resources given the O(N^2) complexity of SVM training [8]. Hence, we chose to use segment-level data for our experiments.

The HMM system is used to generate alignments at the phone level and each phone instance is treated as one segment. Figure 2 shows an example of how we construct a composite vector for a phone segment to avoid the problem of dealing with variable length segments. SVM classifiers in our hybrid system operate on such composite vectors.

For each phone, an SVM classifier was trained to discriminate between this phone and all other phones (one-vs-all classifier). In order to limit the number of samples (especially the out-of-class data) that is required by each classifier, a heuristic data selection process was used. Some of the important heuristics used included the requirement that the training set consists of equal amounts of within-class and out-of-class data. All within-class data available for a phone is by default part of the training set. The out-of-class data was randomly chosen such that one half of the out-of-class data came from phones that were phonetically similar to the phone of interest and one half came from all other phones. Balancing the data by similarity allowed for more data to be used during training.

For decoding, we get the segmentation information using a baseline HMM system — a cross-word triphone system with multiple Gaussian mixtures per state. Composite vectors are generated for each of the segments and posterior probabilities are hypothesized that are used to find the best word sequence using the Viterbi decoder. Assuming that we have already trained SVM classifiers for each phone in the model inventory, we generate N-best lists using a conventional HMM system. These N-best lists can be processed in two ways. The first possibility is to use the segmentation from the best hypothesis generated by the HMM system and rescore all other hypothesis using this segmentation. The lists can then be reordered and the new hypothesis can be chosen.

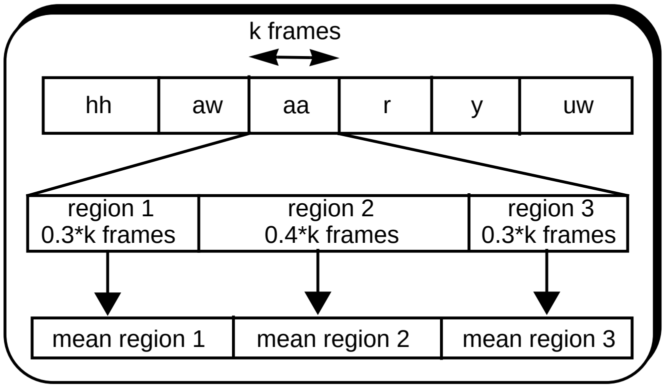

Figure 2: Example of a composite vector construction using a 3-4-3 proportion

Another possibility is to generate a model-level alignment for each of the hypotheses in the N-best list using the HMM system. Based on these alignments, a segmentation for each hypothesis is generated. The likelihood of the corresponding hypothesis is computed by using SVMs to classify each segment. Posterior probabilities are computed using the sigmoid approximation and are used to compute the utterance likelihood of each hypothesis in the N-best list. The N hypotheses can then be re-ranked using the new likelihoods and the best hypothesis can be chosen

The two approaches described above differ significantly in several ways. Using a single segmentation to reorder the N-best list makes the hybrid recognition process simpler. A single pass of rescoring a word-graph comprised of the N-best hypothesis is sufficient to complete the rescoring process. However this approach does not conform to the methodology used for training the SVM classifiers where segmentation generated based on HMM alignments are used to train the classifiers. The second approach of using a separate segmentation to compute the likelihood of each hypothesis in the N-best list is well-matched to the training paradigm. This approach also fits well with approaches where segment graphs are used for decoding [9]. However, it is very cumbersome and computationally expensive.

As a point of reference, we also produced results using a reference segmentation. These are a set of *oracle* experiments where the segmentations are produced by forced-alignments of the reference transcription. The results of these experiments provide a nice analysis tool as they give us a presumptive lower bound on the achievable error (the actual lower bound is the N-best list error rate, but it is a good assumption that we won’t do better than a system with perfect knowledge of the reference segmentation). We hypothesize that this form of oracle experiment isolates the segmentation issue from the recognition process and hence helps calibrate the absolute improvements provided by the SVM classifiers.

3. EXPERIMENTS AND RESULTS

The hybrid architecture has been benchmarked on the OGI Alphadigit corpus that has a vocabulary of 36 words [7]. We used 29 phones to represent the pronunciations of the words, and therefore trained 29 SVM classifiers. The baseline HMM system was trained on 39-dimensional feature vectors comprised of 12 cepstral coefficients, energy, delta and acceleration coefficients. Twelve Gaussian mixture components per state were used. The training set had 50,000 sentences averaging 6 words a sentence. This baseline system delivered a word error rate (WER) of 11.9% on the test set. The test set was an open-loop speaker independent set with 3329 sentences.

The system also has been evaluated on SWB [6]. The training set consists of 114,441 utterances while the development test set consists of 2,427 utterances. These utterances have an average length of six words and an average duration of two seconds. The test set vocabulary is approximately 22,000 words while the training set vocabulary has over 80,000 words. A 42-phone set has been used for this task. The baseline HMM system was trained on 60 hours data from 2,998 conversation sides. The input features were mel-cepstral coefficients which had been normalized to have a zero-mean and unit variance. Twelve mixture components per state were used. This baseline system has a WER of 41.6% on the development test set.

Table 1 compares the performance of the hybrid system with that of the baseline system on various word-classes of the Alphadigits data. These word classes have been found to comprise the major error modalities for the dataset. It is interesting to note that the two systems handle the word classes with varying degree of accuracy. This observation prompted us to explore the idea of combining the acoustic likelihoods from the two system to obtain the final hypothesis.

We explored a system combination scheme where the word likelihood score from the SVM system was combined with the word-likelihood score from the HMM baseline according to

$$likelihood = \text{SVM score} + \frac{\text{HMM Score}}{\text{norm factor}} \quad . \qquad (2)$$

Data Class	HMM (%WER)	SVM (%WER)	HMM+SVM (%WER)
a-set	13.5	11.5	11.1
e-set	23.1	22.4	20.6
digits	5.1	6.4	4.7
alphabets	15.1	14.3	13.3
nasals	12.1	12.9	12.0
plosives	22.6	21.0	18.9
Overall	**11.9**	**11.8**	**10.6**

Table 1: Comparison of performance of the HMM and SVM systems in isolation and in combination as a function of prominent word classes in the alphadigits vocabulary.

This method does require the estimation of another free parameter to normalize the respective scores. As the normalization factor increases, the likelihood is dominated by the SVM hypothesis. Likewise, as the normalization factor decreases, the HMM score dominates. Using this approach we improve the overall system performance to 10.6% WER as compared to the baseline of 11.9%. Interestingly, this score combination scheme shows gains for every error modality explored in this recognition task.

For the SWB task, we use 10-best lists with a list error rate of 29.5% for all experiments. For this experiment we used a segmentation derived from the HMM’s hypothesis to rescore the N-best list. This hybrid setup does improve performance over the baseline, albeit only marginally — WER of 40.6% compared to a baseline of 41.6%.

We then compared and contrasted the effect of using oracle segmentations and transcriptions in the hybrid system. This is an important exercise for gaining further insights into the effect of these features on system performance. On the Alphadigits task, using the reference segmentations improves the performance of the hybrid system from 11.0% to 7.0% WER (compared to a baseline of 11.9% WER). On the SWB task, the reference segmentation improves the performance of the system from 40.6% to 36.1%. This demonstrates that the SVM system is capable of locking onto the correct segmentation.

We conducted another set of experiments to determine the effect of the richness of N-best lists on the performance of the hybrid system. The N-best list error rate was artificially reduced to 0% by adding the reference to the original 10-best lists. Rescoring these new N-best lists using the corresponding segmentations result in error rates of 9.1% WER and 38.1% on Alphadigits and SWB respectively. This improvement corresponds to a 30% relative improvement in performance on the Alphadigits task. On this task, the HMM system does not improve performance over the baseline even when the reference (or correct) transcription is added to the N-best list.

This result indicates that SVMs do a better job than HMMs when they are exposed to accurate segmentations. Unfortunately, the current hybrid approach does not allow the SVM to be trained in a way in which it is exposed to alternate segmentations. Hence, the SVM doesn’t learn to discriminate between alternate segmentations. We hypothesize that this is the reason that introduction of the correct segmentation has such a big impact on performance for the SVM.

Another set of experiments were run to quantify the absolute ceiling in performance improvements the SVM hybrid system can provide. This ceiling can be achieved when we use the hybrid system to rescore the N-best lists that include the reference transcription using the reference-based segmentation. Using this approach, the system gave a WER 3.3% on the Alphadigits task, and 5.8% on SWB. This huge improvement should not be mistaken for a real improvement for two reasons. First, we cannot guarantee that the reference segmentation is available at all times. Second, generating N-best lists with 0% WER is extremely difficult, if not impossible for conversational speech. This improvement should rather be viewed as a proof of concept that by using good segmentations to rescore good N-best lists, the hybrid system has a potential to improve performance significantly. The various results are summarized in Table 2.

4. SUMMARY

In this paper, we discussed the first application of a hybrid SVM/HMM system to a large vocabulary task — SWB. We have extended previously published work on issues related to the hybrid architecture and have explored the segmentation issue in greater detail via a set of oracle experiments. On the OGI Alphadigits task, the hybrid system achieves a WER of 10.6% compared to 11.9% achieved using a traditional HMM system. On the SWB task, the hybrid system achieves a word error rate of 40.6% compared to a baseline of 41.6%.

The results obtained in the experiments clearly indicate the classification power of SVMs and affirm the use of SVMs for acoustic modeling. The fact that the improvements are made on all classes of sounds (some being minimal pairs), indicates that the SVM classifiers are capable of classifying even extremely confusable data better than HMMs. Oracle experiments were performed using segmentations from correct transcriptions and N-best lists containing the reference hypothesis. These experiments show that given good segmentations, SVM classifiers do a significantly better job than traditional Gaussian classifiers.

5. REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA,1995.

[2] A. Ganapathiraju, et al., “Support Vector Machines for Speech Recognition,” *Proc. of the ICSLP*, pp. 2923-2926, Sydney, Australia, November 1998.

[3] A. Ganapathiraju, et al., "A Hybrid ASR System Using Support Vector Machines," *Proc. of the ICSLP*, vol. 4, pp. 504-507, Beijing, China, October 2000.

[4] A. Ganapathiraju and J. Picone, "Support Vector Machines For Automatic Data Cleanup," *Proc. of the ICSLP*, vol. 4, pp. 210-213, Beijing, China, October 2000.

[5] S. Fine, et al., “Hybrid GMM/SVM Approach to Speaker Identification”, *Proc. of the ICASSP*, Salt Lake City, Utah, USA, 2001.

[6] J. Godfrey, et al., “SWITCHBOARD: Telephone Speech Corpus for Research and Development,” *Proc. of the ICASSP*, vol. 1, pp. 517-520, San Francisco, California, USA, March 1992.

[7] R. Cole et al, “Alphadigit Corpus,” *http://www.cse.ogi.edu/CSLU/corpora/alphadigit*, Oregon Graduate Institute, 1997.

[8] E. Osuna, et. al. “An Improved Training Algorithm for Support Vector Machines,” *Proc. of the IEEE NNSP’97*, pp. 24-26, Amelia Island, USA, September 1997.

[9] J. Chang, *Near-Miss Modeling: A Segment-Based Approach to Speech Recognition*, Ph.D. dissertation, MIT Department of EECS, 1998.

S. No.	Information Source		HMM		Hybrid	
	Transcription	Segmentation	AD	SWB	AD	SWB
1	N-best	Hypothesis	11.9	41.6	11.0	40.6
2	N-best	N-best	12.0	42.3	11.8	42.1
3	N-best + Ref.	Reference	—	—	3.3	5.8
4	N-best + Ref.	N-best + Ref.	11.9	38.6	9.1	38.1

Table 2: Summary of recognition experiments using the baseline HMM system and the hybrid system on the Switchboard (SWB) and Alphadigits (AD) tasks. The two information sources that define the experimental setup are the transcriptions that need to be reordered and the segmentations that are fed to the hybrid system. N-best segmentation implies that each of the N segmentations were used to process the corresponding hypothesis in the N-best list.