Structured SVM for Speech Recognition Using Kernel Derivative Model

V. Malarmathi M.C.A [1], Dr. E. Chandra M.Sc, M.phil, Phd [2]

[1]*Research Scholar, Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore, India.*

[2] *Director, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore-32, India*

Abstract— **Kernels model-based is a superior technique for recognizing speech for huge amount of training data in continuous speech recognition technique. This study describes a specific model in this framework, kernel based structured support vector machine (SSVM).This illustrates that how it can be applied to average to large vocabulary recognition jobs. Here a high dimensional feature of derivative kernels and context – dependent models are used .This explores the previous work with combined form of generative and discriminative classifiers. The feature of derivative kernel is extracted. These extracted features can be segmented by using Viterbi-like scheme. Viterbi-like scheme is illustrated for obtaining optimal segmentation. At last a training algorithm can be included into large margin criterion. The performance of SSVM is estimated to aurora 2 speech recognition task. The experimental result presents kernel based SSVM provides better result over SSVM.**

Keywords— **StructuredSVM, kernel methods, *Log Linear Models, AURORA 2***

.

I. INTRODUCTION

A continuous speech recognition system process on speech in that word of the speech is connected together, that is not separated by any pauses. Since it has variety of effects, continuous speech is more complex to handle. Feature extraction and selection for continuous speech recognition is a tedious task. An additional factor which affects the speech is recognition. The recognition has an effect on the rate of speech. Several automatic speech recognition uses a generative model and hidden Markov models (HMM) employs as acoustic models. To produce sentence posterior the Bayes' rule of language model is utilized in prior and Likelihoods from these models were combined together. To form a model for a class label, hidden Markov models (HMM) for individual sub-sentence units can be combined. Even though these models produce successful result but underlying models are not correct in some case. This directs to interest in discriminative classifiers. Discriminative classifier directly models sentence posteriors or decision boundaries for a given set of features that is extracted from observation sequence. Various options have been modified to show how these features are extracted from observed term of sequence. These choices include parametric and non-parametric approaches, generative kernels and event detectors discussed in earlier work. The event detector uses multiple parallel streams of feature. This works at different levels of granularity such as word, multi-phonics and phone. The present applications of event detectors don't advance the underlying acoustic models. Therefore the recognition results from these models are used to derive the features. Additionally the problem associated with adapting features with speaker or noise conditions are not easy to handle. This can be handled by using features of Generative kernel. Features derived from Generative kernels from generative models have numerous advantages. An organized approach of adding new acoustic features has suggested first and higher order derivatives of the log-likelihood by using log likelihoods.

Then by use of model-based compensation or adaptation approaches, generative kernels can be adapted to the noise or speaker conditions.

In this work, Kernel based SSVM is used to recognize the speech for medium and large speech. The log likelihood kernel is adapted for feature space vector. This derived kernel feature has been segmented by Viterbi-like scheme. Large margin log linear model relation to training and decoding process has been analysed by using Structured SVM. Experimental results are presented on small and medium to large vocabulary Continuous speech recognition tasks such as: AURORA 2.

II. RELATED WORK

Almost all continuous speech recognition (CSR) systems use structured generative models, in the form of hidden Markov models (HMMs), as the acoustic models. A model for a class label HMMs can be formed for individual sub-sentence units can be simply combined together. To give out the sentence posterior it is based on Bayes' rule [1]. Likelihoods from these HMMs are combined advance. Two major models were used earlier for recognising speech

They are

1. Generative model
2. Discriminative model

Both of these models were used for recognising small and medium level of speech.

A feature based on generative models permits state-of-the-art noise robustness approaches and speaker adaptation for generative models to be used in [5].

For discriminative models three major choices need to be made:

(i) How to make use of the structure in continuous speech; (ii) the suitable training criterion (iii) and the model of the features to use. A number of features have been explored at the word level, model and frame [2], [4]. By using the conditional maximum likelihood (CML) [2], [3] criterion Discriminative models are frequently trained. To deal with this there has been interest in minimum Bayes' risk [6] criteria and large margin [4] for discriminative models.

Unstructured models namely support vector machines (SVMs) and logistic regression model, consider class labels which are independent and have no structure. The space of feasible classes becomes very large when applying these models to absolute utterance in CSR.One possible resolution to deal with this is to segment the continuous speech into words or sub-words Observation sequences which are similar to acoustic code-breaking [7].For each segment, logistic regression or multi-class can be proposed in the same way as isolated classification tasks [4], [5].

Thus, this approach has two problems. First, each segment is treated separately. Second, the classification is based on fixed segmentation. To deal with this task the following methods has been proposed.

.

III. PROPOSED WORK

A. *Joint Feature Space*

For medium and large continuous speech recognition task Joint feature space has been employed in earlier research. In general continuous speech recognition, hypothesized sentences are very large. To deal with this problem the labels are decomposed into phonetic and shared structure units. The given parameters to direct this work is as follows:

θ is an Alignment which decomposes the observation sequence into segments $O=\{O_{1|\theta},\ldots, O_{i|\theta},\ldots, O_{|W||\theta}\}$ with corresponding labels $w=\{w_{1,\ldots,}w_{i,\ldots,}w_{|w|}\}$

Where $O_{i|\theta}$ is the i^{th} segment connected to context-dependent phone label w_i.

The resulting Joint feature can be described as $\alpha=\begin{bmatrix}\alpha^{(v1)}\\ .\\ .\\ .\\ \alpha^{(v_M)}\\ \alpha^{1M}\end{bmatrix}$,

$$\emptyset\ (O,w,\theta) \triangleq \begin{bmatrix}\sum_{i=1}^{|w|}\delta(w_i - v1)\varphi(O_{i|\theta})\\ :\\ :\\ \sum_{i=1}^{|w|}\delta(w_i - v_M)\varphi(O_{i|\theta})\\ log\ P(w)\end{bmatrix} \rightarrow(1)$$

Where v1…….. v_M denotes all feasible triphones in dictionary

$\delta(w_i - vM)$ – Kronecker delta function

$\varphi(O_{i|\theta})$ - Feature vector extracted for segment $O_{i|\theta}$

$P(w)$ is the standard n-gram language probability model.

.**Derivative kernel feature**

In this Generative kernels extract features of generative models. The example for this scenario is the use of the log-likelihood kernels.

From observation sequence O log-likelihood of generative model calculated for class ωi is estimated by the following equation (2).

$\emptyset_b^0(O|\ \omega i) = [\ \log(p(O|\ \omega i))] \rightarrow(2)$

Where b represents base features.

The log likelihood of the forwarding class is added by **a** is as shown in equation (3)

$$\emptyset_a^0(O|\ \omega i)=\begin{bmatrix}\log(p(O|\omega 1))\\ log(p(O|\omega 2)\ \ldots\\ log(p(O|\ \omega k)\end{bmatrix}\rightarrow(3)$$

Conditional independence hypothesis of the generative models are inherited by features derived from the base (b) and added (a) of log likelihood kernels.

Features which are derived from derivative kernels have different conditional independence hypothesis.

Consider the following feature vector, the ρth order base derivative kernel is described in the equation (4)

$$\emptyset_a^\rho(O|\ \omega i)=\begin{bmatrix}\log(p(O|\omega i))\\ \nabla_\lambda\ log(p(O|\omega 2)\ \ldots\\ \nabla_\lambda^\rho log(p(O|\ \omega k)\end{bmatrix}\rightarrow(4)$$

The class log-likelihood of the feature vector can be corrected in equation (4) by using derivatives up to order ρ based on the generative model parameter.

Assume the first order derivatives with respect to θ^{jm} component

$\lambda_{jm} = \{\mu_{jm}, \Sigma_{jm}\}$

And output distribution:

$\nabla_{\lambda_{jm}} \log(p(O|)) = \sum_{t=1}^{T} P(\theta_t^{jm}|O)\ \nabla_{\lambda_{jm}} \log(p(o_t|\theta_t^{jm}))$

The above derivative is a function of $P(\theta_t^{jm}|O)$ component posterior probabilities, which based on complete observation sequence. This means that additional dependencies had been introduced into the features by the use of derivatives.

B. *Segmentation based on Viterbi-like scheme*

The derivative kernel feature described above is depending on specific segmentation θ_λ which is generated by using standard generative model HMM. The solution obtained from this segmentation can be fixed for both training and decoding. The following equation shows the solutiont of inference

w_α=arg max (w)$\alpha^T \emptyset(O,w;\ \theta_\lambda) \rightarrow(5)$

Whereθ_λ=arg max$P(\theta|w)p\ \lambda(O|\ \theta,w)\rightarrow(6)$

The above equation (6) can be solved by Viterbi algorithm while using HMM.

For generative model, θ_λ is the optimal segmentation and yields best segmentation result.

C. *Relationship with Log Linear Models*

Structured SVM can be viewed as large margin log linear model.

The posterior of the log linear model can be defined as follows

$$P\ (w|O;\ \theta_\alpha,\alpha)=\frac{exp\ (\alpha^T\ \emptyset(O,w;\theta_\alpha))}{\sum_{w'}\ exp\ (\alpha^T\ \emptyset(O,w';\theta_\alpha'))}\rightarrow(7)$$

Where θ_α is the best segmentation which maximizes posterior probability P (w|O; θ_α,α)

Decoding can be done with this log linear model. This can be expressed in following expression.

w_α=arg max P(w|O; θ_α, α)→(8).

IV. EXPERIMENTAL RESULTS

This section presents the experimental result attained for Proposed SSVM. To demonstrate the proposed SSVMs the noise-corrupted corpus AURORA 2 was used.

A. *AURORA 2*

AURORA 2 is a connected and continuous digit string task. The vocabulary size is 12 that is one to nine, plus zero, oh and silence. Depends on the TIDIGITS database with artificial added noise, the utterances are from one to seven digits long. To train the acoustic generative models (HMMs) the 8440 clean mixed-gender training utterances were utilized. Three test sets were used .They are set A, B and C. For set A and B total of 8 noise conditions was added i.e., 4 for each and 5 for different SNR (Signal to Noise ratio). For set C the distortion was added.

A series of configuration was compared to compute the needs of the Kernel based SSVMs framework. The word error rate accuracy has been evaluated to the proposed research. The following table describes the accuracy and average value obtained for Joint based feature SSVM and log-linear kernel based SSVM.

Model	Criterion	Test Set (WER%)			Avg
		A	B	C	
SSVM	LM (n-slack)	7.8	7.3	8.0	7.7
Kernel based SSVM	LM (n-slack)	7.0	6.4	7.5	7.0

.

. Table 1: AURORA2 WER% result for SSVM and Kernel based SSVM

Table 1 describes the word error rate accuracy for SSVM and Kernel based SSVM. The Kernel based SSVM gives better accuracy when compared with Structured SSVM. The graphical depiction of proposed SSVM can be demonstrated in the following graph with its corresponding estimated values in the following table2 respectively.

SNR (dB)	Avg of test sets A, B and C	
	SSVM	Kernel SSVM
20	1.2	0.8
15	1.8	1.4
10	3.4	2.9
05	8.5	8.3
00	23.5	20

Table 2: Aurora 2 result for SSVM and Kernel SSVM

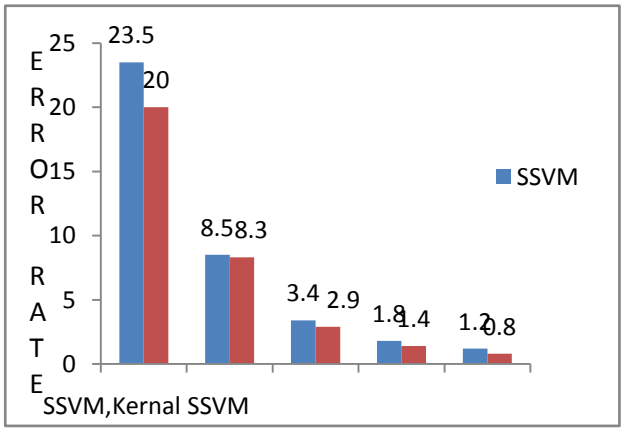

Figure 1: ssvm verses kernel ssvm

The above graph in Figure 1 describes the result obtained for Joint feature SSVM and Kernel based SSVM. The Signal to noise ratio of 0 to 20 dB (decibel) has been compared on the vertical axis. The estimated value for SSVM and kernel based SSVM is shown in the table 2. The result shows that the kernel based SSVM gives better performance when compare with existing work.

V. CONCLUSION

This research has described a structured SVM framework which is suitable for medium to large vocabulary speech recognition. Various theoretical and practical extensions to previous research on small vocabulary task are reported. This research has described a continuous discriminative model based on log-linear derivative kernels. This method is suitable for noise-robust medium/large vocabulary speech recognition. Using model-based techniques noise/speaker conditions has been adapted by using the generative models. The features had been extracted from observed sequence by using the adapted model. Viterbi-like scheme segments the extracted features. Structured SVM can be viewed by Large margin log linear model. The relationship with this model has been addressed in the proposed research. The performance of the proposed work was evaluated on noise corrupted task namely AURORA 2 with three test sets. Future work can be extended by adding Active learning with SSVM for predicting the speech.

REFERENCES

[1] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundat. Trends Signal Process.*,p.2007.

[2] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU*, 2009.

[3] M. Layton and M. Gales, "Augmented statistical models for speech recognition," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 129–132.

[4] S.-X.Zhang, A.Ragni, andM.J.F.Gales,"Structured log linear models for noise robust speech recognition," *IEEE Signal Process.Lett.*, vol. 17, no. 11, pp. 945–948, Nov. 2010.

[5] M. J. F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Comput. Speech Lang.*,vol. 24, no. 4, pp. 648–662, 2010.

[6] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition," Ph.D. dissertation, Cambridge Univ., , 2004.

[7] V. Venkataramani, S. Chakrabartty, and W. Byrne, "Support vector machines for segmental minimum Bayes risk decoding of continuous speech," in *Proc. ASRU*, 2003..