SVM BASED POSTERIOR PROBABILITIES FOR SYLLABLE CONFIDENCE ANNOTATION

Daniel Bolanos [1, 2], Wayne Ward [1], and Javier Tejedor[2]

[1] Center for Spoken Language Research, University of Colorado at Boulder, USA
[2] HCTLab-Escuela Politécnica Superior, Universidad Autónoma de Madrid, SPAIN
{bolanos, whw}@cslr.colorado.edu, e-mail: {javier.tejedor}@uam.es

ABSTRACT

In this paper we present a mechanism to incorporate support vector machine (SVM) based phone posterior estimates in the computation of posterior probabilities over syllable lattices. A continuous speech recognizer is used to generate a syllable lattice. Using the state alignment information associated to each syllable in the lattice, SVM-based posteriors are calculated for each phone and then combined to obtain syllable posterior probabilities. Finally, these probabilities are incorporated into the computation process of posterior probabilities over syllable graphs using the forward-backward algorithm.

Experimental results show that the SVM-based confidence measures computed over syllable lattices can substantially reduce the classification error rate of HMM-based state-of-the-art confidence measures.

Index Terms— Confidence annotation, machine learning, support vector machines, posterior probabilities, syllable lattices

1. INTRODUCTION

The rapid development of speech technology in the recent years has enabled the development of a wide variety of speech applications. Unfortunately speech recognition results are still far from perfect, which, for practical applications, requires the use of confidence annotation techniques that help in the detection of misrecognized words. In our previous work [1] we have introduced a children's speech reading tracker that makes use of a syllable rejection module to classify syllables in a reference string as correctly or incorrectly read. In this article we have tried to improve the performance of this rejection module by integrating SVM-based posterior estimates in the computation of posterior syllable probabilities over syllable lattices.

Recently, SVM-based classifiers have been successfully applied for N-best lists rescoring at the output of a conventional HMM decoder [2]. These classifiers produce posterior class probability estimates that can also be used to generate confidence annotation labels. Previous work has shown [3] that confidence measures based on word posterior probabilities estimated over word graphs outperform alternative confidence measures [4] such as acoustic stability and hypothesis density. In this article we try to incorporate syllable posterior probability estimates obtained from SVM classifiers into a posterior probability computation procedure over syllable lattices.

In section 2 we present three different SVM-based phonetic classifiers and show how they can be used to generate syllable posterior probabilities. In the next section we define three different confidence measures resulting from the integration of those syllable posteriors into the computation of posterior probabilities over syllable graphs. Finally, we compare the performance of the confidence measures proposed with an HMM-based confidence measure taken as baseline and present our conclusions.

2. SVM FOR SYLLABLE CONFIDENCE ESTIMATION

An SVM learns the decision boundary between samples belonging to two classes by mapping the training sample vectors into a higher dimensional space and then determining an optimal separating hyper-plane [5]. When SVMs are used in classification tasks for speech processing applications it is necessary to map the margin or distance they produce to a posterior class probability. This can be done by the use of a sigmoid [6], where the parameters A and B need to be estimated by cross-validation.

$$p(y=1 \mid f) = \frac{1}{1+e^{(Af+B)}} \tag{1}$$

In the context of syllable classification, this posterior probability can be used to express the probability that a sequence of speech frames belongs to a syllable class. In our case, due to the considerable number of syllables present in the speech corpora, we decided to use SVMs to calculate posterior phone probabilities and combine them to calculate syllable posterior probabilities. We do not attempt to model coarticulation.

2.1. Phonetic classification

We have proposed 3 different phonetic classifiers. The simplest one (2) consists of training an SVM classifier for each phone using speech features directly as training vectors. This way, the probability of a phone *ph* given the sequence of feature vectors $x_1^T = \{x_1, x_2, \ldots, x_T\}$ to which it is aligned, is estimated as the average of the posterior phone probabilities obtained from the SVM for each of its frames.

$$p_{frames}(ph \mid x_1^T) = \frac{1}{T}\sum_{t=1}^{T} p(ph \mid x_t) \tag{2}$$

In the second approach (3) we have tried to take advantage of the state alignment information produced by an HMM aligner to capture the time-varying structure of a phone, which is missing in the previous approach. We use the Sonic speech recognition system for the alignment [10]. Sonic uses 3-state HMM phone models. Feature vectors aligned to each state of a phone are

averaged and the resulting average vectors are concatenated to form a composite vector that is used to train the SVM classifier. The dimension of this vector is, consequently, three times the dimension of the original feature vectors. A similar approach was proposed in [2], but assigned a fixed percentage of the frames aligned with a phone to each state.

$$p_{segments}(ph \mid x_1^T) = p(ph \mid composite(x_1^T)) \quad (3)$$

The third approach (4) uses the phone temporal structure information while still using speech features directly as training vectors for the SVM classifier. The later consideration is important since, as we will see later, the averaging process carried out in the second approach prevents the reuse of SVM predictions across a lattice time frame. This procedure is more computationally expensive and significantly deteriorates the real time performance. Hence, three SVM classifiers are trained for each phone, each of them trained with the speech features from a different state.

$$p_{frames/states}(ph \mid x_1^T) = \frac{1}{3}\sum_{s=1}^{3}\frac{1}{T_s}\sum_{t=1}^{T_s} p(ph_s \mid x_t) \quad (4)$$

2.2. Syllable classification

As expressed in (5), syllable posterior probabilities are calculated by averaging phone posterior probabilities.

$$p(syl \mid x_1^T) = \frac{1}{N}\sum_{i=1}^{N} p(ph_i \mid x_{T_{i-1}+1}^{T_i}) \quad (5)$$

By using the three phonetic classifiers presented in section 2.1 as the phone posterior probability, we define the following posterior syllable probabilities: $p_{frames}(syl|x_1^T)$, $p_{segments}(syl|x_1^T)$ and $p_{frames/states}(syl|x_1^T)$, that can be used as syllable classifiers.

3. EXPERIMENTS FOR SYLLABLE CONFIDENCE ANOTATION

Experiments were carried out to evaluate the accuracy of the phonetic and syllable classifiers proposed.

3.1. Speech material

We present experimental results on the CU Read and Summarized Story Corpus [8]. We have selected speech belonging to first and second graders (a total of 171 and 57 speakers respectively) and partitioned it into a training set containing 9 hours of audio and a test set of about 2 hours of audio.

3.2. Training and parameter selection

For every speech segment present in the training set, 39-dimensional feature vectors, consisting of 12 Mel Frequency Cepstral Coefficients and energy plus first and second order derivatives, have been extracted. The children's speech corpora available is tagged at the word level only so phone boundaries are obtained using a Viterbi-based phonetic alignment against the transcriptions.

SVM classifiers are well suited for two-class separation tasks, however for n-class (n>2) separation tasks, like building a phonetic classifier, n SVM classifiers need to be trained. In this case we have selected a "one vs. all" approach in which up to three SVM classifiers [9] are trained for each of the 55 phonetic symbols used. For the training of each SVM, half of the data points (positive samples) belong to the actual class while the rest belong to the remaining classes (negative samples).

A radial basis function (RBF) kernel is used for which the parameters C (cost) and γ are estimated over the training set with a "grid-search" process using 5-fold cross validation.

3.3. Phonetic classification

The first experiment conducted evaluates the classification accuracy of the three phonetic classifiers proposed. For evaluating the classifiers we created a test set that contained 500 positive examples and a balanced set of 500 negative examples for each of the 55 phones used. In the test set, each instance is assigned a phone label (half of which are correct labels). For each of the classification algorithms, 55 classifiers (corresponding to one-vs-rest classifiers for each phone) were trained. For each instance, the classifier corresponding to the phone label for the instance is used to assign a probability of the phone given the data. This score is compared to previously trained thresholds (phone-dependent) to classify the phone occurrence as belonging to the phone class or not. Figure 1 shows the classification accuracy for each classification algorithm and every phone class.

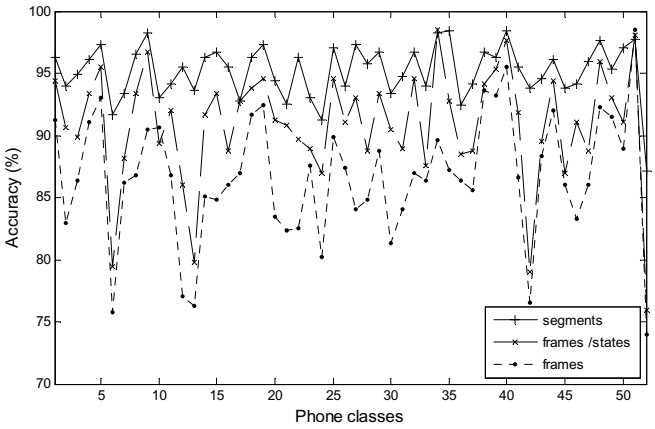

Figure 1. Phonetic classification accuracy for the three classifiers proposed

In table 1 we summarize the average classification results across the whole phonetic symbol set for the proposed classifiers. As expected, the segment-based approach yields the best classification accuracy, followed by the frame-based approach that makes use of state alignment information. An interesting detail observed is that the trained threshold used in the decision making process varies greatly between phones. This suggests that score distribution information for each phone could be incorporated in the syllable classifier for improved accuracy.

Approach	Average accuracy (%)
$P_{frames}(ph\|X)$	86.67
$P_{segmentss}(ph\|X)$	95.20
$P_{frames/states}(ph\|X)$	90.87

Table 1. Average classification accuracy across the phonetic symbol set

3.4. Best single path confidence annotation

In this experiment we compare the classification accuracy of the three syllable classifiers proposed in section 2.2. The experiment consists of running a decoding process using Sonic [10]. The single best scoring path is then annotated with posterior syllable probability estimates. In this experiment, the confidence annotation uses only the state alignment information of the best single path and no lattice information or language model probability is used. The Figure 3 shows a comparison of the syllable classifiers using Detection-Error Tradeoff curves that contain a plot of the false acceptance rate over the false rejection rate.

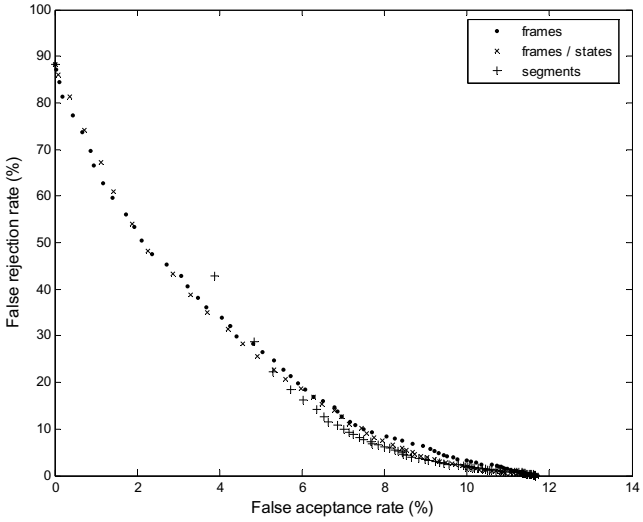

Figure 2. DET curves for lattice based posterior probabilities.

Surprisingly, despite the observed difference in classification performance of the three phonetic classifiers used as a basis for the confidence annotation, the curves depicted are very similar. This can be observed especially in the most relevant area of the graph, i.e. for FRR < 10. However, the segments-based classifier yields a slightly better accuracy for syllable confidence annotation.

4. COMPUTATION OF SYLLABLE POSTERIOR PROBABILITIES OVER SYLLABLE LATTICES

In this section we discuss the mechanism for incorporating the SVM based phone posterior estimates presented in section 2 in the computation of posterior probabilities over syllable lattices.

Initially we describe briefly the typical computation procedure of syllable posterior probabilities over syllable lattices. The posterior probability $p([syl;s,e]|X)$ for a syllable can be calculated as defined in (6) by summing up the posterior probabilities of all paths in the lattice of length M which contain the hypothesis $[syl;s,e]$. $[syl;s,e]$ is the syllable starting at time s and ending at time e, and $X=\{x_1,\ldots,x_T\}$ is the acoustic observation sequence against which it is aligned.

$$p([syl;s,e]\,|\,x_1^T) = \sum_{\substack{[syl;s,e]_1^M:\\ \exists n\in\{1,\ldots,M\}:\\ [syl_n;s_n,e_n]=[syl;s,e]}} \frac{\prod_{m=1}^{M} p(x_{s_m}^{e_m}\,|\,syl_m)^{\alpha}\, p(syl_m\,|\,syl_1^{m-1})^{\beta}}{p(x_1^T)} \tag{6}$$

Typically these posterior probabilities are calculated very efficiently over syllable graphs using the forward-backward algorithm as described, in the case of words, in [3] and [7]. This algorithm considers edges in the graph as HMM-like states, where emission probabilities are the HMM acoustic models scores and transition probabilities between links are obtained from the language model used.

We have proposed an alternative computation procedure (7) where the HMM acoustic scores for each syllable are substituted by the posterior syllable probabilities produced by the SVM syllable classifiers defined in section 2. We realize that this is replacing a quantity that represents P(observations | syllable) with a quantity that is a direct estimation of the posterior probability P(syllable | observations). We believe that these posterior syllable probabilities, given the equality assumption for the prior class probability made in the construction of the SVM classifiers that produce them, can still be effectively combined with language model probabilities in the computation of posterior probabilities over syllable graphs.

$$p([syl;s,e]\,|\,x_1^T) = \sum_{\substack{[syl;s,e]_1^M:\\ \exists n\in\{1,\ldots,M\}:\\ [syl_n;s_n,e_n]=[syl;s,e]}} \frac{\prod_{m=1}^{M} p([syl_m;s_m,e_m]\,|\,x_{s_m}^{e_m})^{\alpha}\, p(syl_m\,|\,syl_1^{m-1})^{\beta}}{p(x_1^T)} \tag{7}$$

In (6) and (7) α represents the acoustic score scaling factor while β represents the language model probability scaling factor. These parameters are necessary to compensate the different dynamic range of acoustic and language model scores, and need to be estimated over a cross-validation set independent from the test set. However, previous work [3] has demonstrated that posterior probabilities calculated as (6) or (7) do not produce satisfactory results. The reason is that the fixed starting and ending time frames of a hypothesis syllable strongly determine the paths involved in the calculation of the forward-backward probabilities. Usually, syllable hypotheses with similar starting and ending time frames represent the same syllable; it therefore makes sense to consider the summation of the posterior probabilities of these syllables as a confidence measure. For this reason we have used a confidence measure (8) proposed in [3] for which the posterior probability accumulation process is carried out over all the time frames of the hypotheses under consideration. After the accumulation process is done, the highest probability value is selected as a measure of confidence.

$$C([syl;s,e]) = \max_{e_{\max}\in\{s,\ldots,e\}} \sum_{[syl;s',e']:s'\le e_{\max}\le e'} p([syl;s',e']\,|\,x_1^T) \tag{8}$$

By substituting in (7) the three posterior syllable probabilities defined in section 2.2 and applying the probability accumulation process defined in (8) we define the following respective confidence measures: $C_{SVMframes}[syl;s,e]$, $C_{SVMsegments}[syl;s,e]$ and $C_{SVMframes/states}[syl;s,e]$. The performance of these confidence measures will be evaluated in section 5 against a baseline confidence measure computed combining expressions (6) and (8) and referenced in the following as $C_{HMM}[syl;s,e]$.

5. EXPERIMENTS FOR LATTICE BASED POSTERIOR PROBABILITIES

In this experiment we compare the performance of the three SVM-based confidence measures proposed in section 4 against an HMM-based one, also described in section 4, which constitutes the baseline. All these confidence measures make use of lattice information so all of them are applied after an initial step of syllable lattices generation using the SONIC continuous speech recognizer. For all of them the scaling factors α and β has been trained over a development set different than the test set. The experimental set-up is the same as described in section 3.

In addition to the use of DET curves, the metric selected to evaluate these confidence measures is the classification error rate (CER) defined as the number of incorrectly assigned tags (which comprises false acceptations and false rejections) divided by the total number of recognized syllables. In figure 3 we show the DET curves of the confidence measures $C_{SVMframes}[syl;s,e]$ and $C_{HMM}[syl;s,e]$ (the baseline). The reason for not depicting the other two SVM-based confidence measures ($C_{SVMsegments}[syl;s,e]$ and $C_{SVMframes/states}[syl;s,e]$) is that their respective DET curves almost completely overlap with the $C_{SVMframes}[syl;s,e]$ in the graph so cant be distinguished. However, the best confidence error rates (CER) for all of them are shown in Table 2.

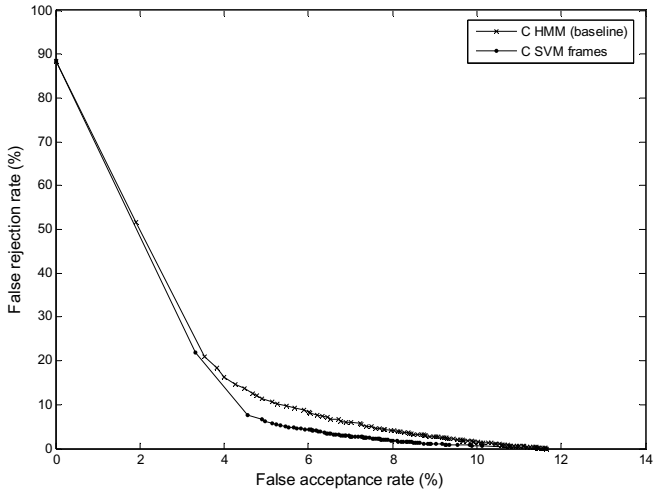

Figure 3. Detection-error tradeoff curves for lattice based posterior probabilities.

Confidence Measure	CER	Relative error reduction (%)
$C_{HMM}[syl;s,e]$ (baseline)	11.43	
$C_{SVMframes}[syl;s,e]$	9.68	15.16
$C_{SVMsegments}[syl;s,e]$	9.63	15.60
$C_{SVMframes/states}[syl;s,e]$	9.35	18.05

Table 2. Confidence error rates and relative error reduction respect to the baseline for the different confidence measures proposed.

As can be seen in Figure 3, the SVM-based confidence measures clearly outperform the HMM-based one used as baseline. In particular, the CER of the SVM-based approaches is at least 15% better than the baseline. Another interesting point is that, despite the considerable differences in classification accuracy observed in the phonetic classifiers (see Table 1) in which these confidence measures rely, their CER is very similar. Considering this similarity, in the context of a real world application, the $C_{SVMframes}[syl;s,e]$ is the most interesting one because the SVM-predictions at the frame level can be shared during the calculation of posterior class probabilities of overlapping phones in the lattice. Note that in the $C_{SVMframes}[syl;s,e]$ confidence measure computation is also possible to share a good number of SVM-predictions (the amount of predictions reused strongly depends on the lattice density). However for the computation of $C_{SVMsegments}[syl;s,e]$, due to the averaging process necessary for creating the composite vectors, no SVM-predictions can be reused so the real time performance deteriorates significantly when the lattice density is relatively high.

6. CONCLUSIONS AND FUTURE WORK

An effective way to incorporate SVM-based posterior probabilities in the computation of posterior probabilities over syllable graphs has been introduced. The new confidence measures presented clearly outperform existing ones in the experiments carried out. Moreover, these confidence measures can be used not only for rejection tasks but for lattice rescoring. Future work will be focused to the use of these confidence measures not only for building a syllable rejection module as part of our children's speech reading tracker but for increasing the information available in the algorithm we are currently using for aligning syllable lattices against multiple pronunciations graphs of syllables.

7. REFERENCES

[1] D. Bolanos, W. Ward, S. Van Vuuren, J. Garrido, "Syllable Lattices as a Basis for a Children's Speech Reading Tracker", in *InterSpeech*, Antwerp, Belgium 2007.

[2] Applications of Support Vector Machines to Speech Recognition A. Ganapathiraju, J. E. Hamaker, J. Picone, *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348-2355, 2004

[3] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288–298, March 2001.

[4] T. Kemp and T. Schaaf, "Estimating confidence using word latices," in *Proc. 5th Eur. Conf. Speech, Communication, Technology 1997*, Rhodes, Greece, Sept. 1997, pp. 827–830.

[5] Vapnick, V. T*he Nature of statistical Learning Theory.* Springer-Verlag, New York, 1995.

[6] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999.

[7] K. Hacioglu and W. Ward, "A Concept Graph Based Confidence Measure", in *International Conference of Acoustics, Speech, and Signal Processing*, Orlando-Florida, USA, 2002

[8] R. Cole and B. Pellom. University of Colorado read and summarized story corpus. Technical Report TR-CSLR-2006-03, University of Colorado, 2006.

[9] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001.

[10] B. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, CSLR, University of Colorado, March 2001.