

# Phonetic Segment Rescoring using SVMs

**Yeojin Kim**

2013 Beckman Institute  
University of Illinois at Urbana-Champaign  
[yeojin@uiuc.edu](mailto:yeojin@uiuc.edu)

**Mark Hasegawa-Johnson**

Electronics, Computer Engineering  
University of Illinois at Urbana-Champaign  
[jhasegaw@uiuc.edu](mailto:jhasegaw@uiuc.edu)

## Abstract

We used SVM to rescore the output of an HMM speech recognizer. We focused on confusable phone pairs to improve recognition rates and used the confusion matrix in order to choose confusable pairs. We performed experiments using parallel SVMs to determine which frames are most useful for rescoring in each context.

## 1 Introduction

Hidden Markov Models (HMM) are widely used in speech recognition because of their ability to do automatic time alignment. A fundamental limit of HMM is the trade-off between complexity and accuracy of Gaussian Mixture models (GMM) (Rabiner, 1989). Although GMM has good performance in terms of expressing speech signals, an accurate GM model of the speech PDF requires several dozen mixture elements, therefore requiring hundreds of hours of speech data to train.

A Support Vector Machine (SVM) produces an optimal decision hyperplane for binary classification (Burgess, 1998). However, it has a disadvantage; it does not consider time alignment. Combining HMM with time property and SVM providing an optimal decision hyperplane is expected to improve speech recognition error rates.

Ganapathiraju and Picone (2000) described the use of SVM within the framework of HMM based speech recognition and estimated a warping function that maps SVM distances to posterior probabilities. Their approach is to divide the segment into three regions in a set ratio and construct a composite vector from the mean vectors of the

three regions. 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.

However, it remains difficult to discriminate confusable phone pairs. A small number of confusable pairs account for most phone recognition errors. It is hard to reduce the error rate using segmental features without additional information revealing the characteristics of each phone. This paper suggests a way to choose the most confusable phones by use of the confusion matrix, and to find the most useful frames which include distinguishing features for these pairs. Our approach is to introduce neighbor frames from context segments to the SVM. The result was better than using only segmental features.

One remarkable point is that speech segments estimated by HMM are unreliable since the range of a segment is varied by hypothesis. This is a critical problem because high performance could not be expected if features as input to a decision function are not correct. The landmark time based speech segmentation (M. Hasegawa-Johnson et al., 2004) provides more reliable segments for SVM. We used this methodology to consolidate our integrated HMM/SVM framework.

## 2 Confusion Network

HTK produces a lattice from training data and the lattice is scored to select the correct phones (Young et al., 2002). The lattice can be “pinched” and rescored using posterior probability in order to further reduce word error rate (Mangu et al., 2000).

Phone errors in the first pass recognizer output can be summarized in a confusion matrix. So we use the confusion matrix in order to choose confusable pairs. Many phone errors involve phones

which are in the same phonetic group and have almost the same duration. If we can limit the number of comparable phones, it is possible to train an SVM for each pair of compared phones. The confusion matrix is a way to limit the population of phones rationally. If we can discriminate these pairs correctly, we can rescore the lattice, yielding higher recognition rates. This paper introduces SVMs as the discriminant functions to classify confusable phones in the lattice output of a speech recognizer.

### 3 Support Vector Machines

SVM is a machine learning method to perform pattern recognition between two classes by finding a decision surface that has maximum distance from the closest points in the training set, which are termed support vectors. If we label the training data  $\{x_i, y_i\}, i = 1, \dots, L, y_i \in \{-1, 1\}, x_i \in R^d$ , the decision function has the form

$$f(x) = \sum_{i=1}^L \alpha_i y_i x_i \cdot x + b \quad (1)$$

where the coefficients  $\alpha_i$  and the  $b$  are the solutions of the quadratic programming problem and  $L$  is the number of support vectors. The vectors  $x_i$ , called support vectors satisfy the following function,

$$y_i(w \cdot x_i + b) = 1 \quad \forall i \quad (2)$$

where  $w$  is the optimal hyperplane to separate two classes.

To extend to the case of nonlinear separating surfaces, each point  $x$  in the input space is mapped to a point  $z = \Phi(x)$  of a higher dimensional space where the data are separated by a hyperplane. The mapping of  $\Phi(\cdot)$  is subject to the condition that the dot product of two points in the feature space,  $\Phi(x) \cdot \Phi(y)$ , can be rewritten as a kernel function  $K(x, y)$ . Then the decision function has the form

$$f(x) = \sum_{i=1}^L \alpha_i y_i K(x_i, x) + b \quad (3)$$

Two useful families of kernel functions are the homogeneous polynomial kernel

$$K(x, y) = (x \cdot y)^d \quad (4)$$

and the radial basis function (RBF) kernel

$$K(x, y) = \exp\left(-\frac{1}{\sigma^2} \|x - y\|^2\right) \quad (5)$$

### 4 Parallel Support Vector Machines

We need to find some way to represent the context-dependence of each phone in the lattice. For example, consonants influence the spectral transition into the following phone; phonetic distinguishing information is not limited to frames covered by the consonant. Therefore, we can reduce error rates by including or excluding neighbor frames according to the context.

There are several ways to select frames for an SVM. We divided a phonetic segment into several parts in order to know which frames are most useful. Our experiments show that we can get lowest error rates when splitting a phonetic segment to several SVM classifiers rather than classifying the segment with one SVM. Each frame goes through the related SVM and the discriminant functions from the SVMs are summed and used to determine the segment's phoneme label. These parallel SVMs for a confusable pair may have different importance and the difference could be reflected by giving a proper weight to each SVM.

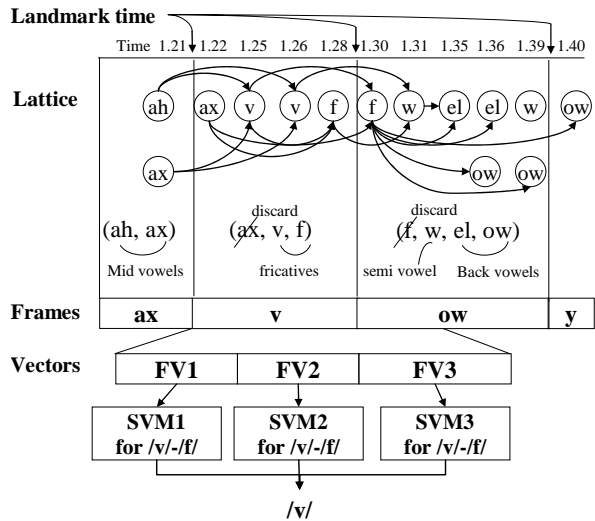


Figure 1. Parallel SVMs for a phone in the lattice

In our system, phoneme candidates in the lattice are first time-aligned to detected landmarks. As shown in (figure 1), some phone candidates may overlap more than one segment. For each segment, we first construct a candidate set by listing all phone candidates overlapping the segment. Then,

in order to decrease the number of compared phone, we delete from the candidate set any phones that were also considered in the previous segment. Based on the phones remaining in the candidate set, we construct an SVM vector, and call appropriate pair-wise SVMs.

## 5 Experiments

We performed experiments using landmark-based speech segments for the NTIMIT database (C. Jankowski et al, 1990). Cepstrograms with variable lengths were resampled in time in order to make a fixed length vector for an SVM. The vector length of each SVM is set to the mean segment length of each confusable phone pair. For construction of SVMs, we used a public SVM toolkit, SVMLight (T. Joachims, 2004).

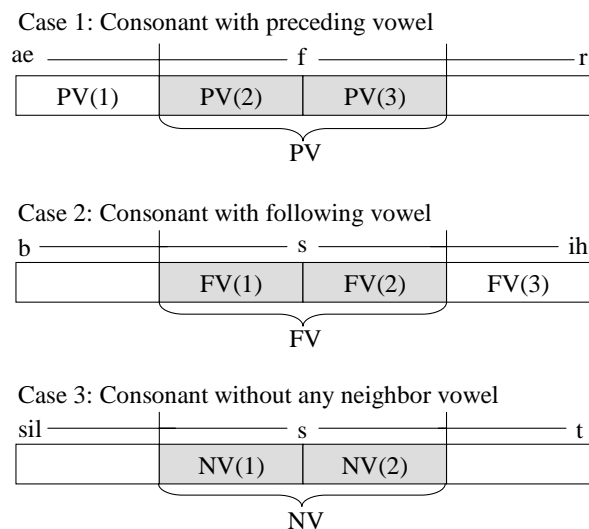


Figure 2. Splitting phonetic segment for SVM

We considered three cases for consonants (figure 2). The first is the case of consonant with preceding vowel (PV) and the second is the case of consonant with following vowel (FV). Finally, the third is the case of consonant without any neighbor vowel (NV). We selected RBF kernel and got the optimal parameters of SVM by experiments. Tables 1-3 show results of experiments distinguishing /f/ and /s/; /f/ and /s/ were the most confused pair in the output of our LPCC-HMM NTIMIT phone recognizer. Table 2 shows that we can get better results by using frames of the vowel part, FV(3). The parallel SVMs covering PV or NV part have the lowest error rate in table 1 and table 3.

The total error rate is 18.24% when the total sample number is 7269 (/f/: 963, /s/: 6306). This is 2.34% lower than non-context error rate, 20.58%.

	PV(1)	PV(2)	PV(3)	PV
/f/	0.3227	0.2173	0.3323	0.1518
/s/	0.4001	0.3236	0.1897	0.2448
Avg.	0.3891	0.3086	0.2093	0.2317
	PV(1+2)	PV(1+3)	PV(2+3)	PV(1+2+3)
/f/	0.2444	0.2460	0.1693	0.2029
/s/	0.3515	0.2367	0.2051	0.2516
Avg.	0.3363	0.2380	0.2001	0.2448

Table 1. Error rates of SVM for discrimination of /f/ and /s/ with preceding vowel (Case 1)

	FV(1)	FV(2)	FV(3)	FV
/f/	0.2368	0.3722	0.2632	0.2556
/s/	0.3096	0.1324	0.2500	0.1850
Avg.	0.2970	0.1738	0.2523	0.1972
	FV(1+2)	FV(1+3)	FV(2+3)	FV(1+2+3)
/f/	0.2256	0.1842	0.2632	0.1917
/s/	0.1959	0.2445	0.1865	0.1795
Total	0.2010	0.2341	0.1997	0.1816

Table 2. Error rates of SVM for discrimination of /f/ and /s/ with following vowel (Case 2)

	NV(1)	NV(2)	NV(1+2)	NV
/f/	0.3380	0.2113	0.2394	0.1831
/s/	0.1791	0.1455	0.1161	0.1243
Avg.	0.1878	0.1491	0.1229	0.1275

Table 3. Error rates of SVM for the discrimination of /f/ and /s/ without vowel (Case 3)

## 6 Conclusions

At present, we are constructing parallel SVMs for a large number of confusable pairs selected from the HMM confusion matrix. We will conduct experiments to choose the most useful frames for vowels, not only consonants and such as constructed parallel SVMs will be tested in the lattice, generated with N-best hypothesis. Our future work is to incorporate these parallel SVMs for lattice rescoring. For this, a methodology for multi-class SVM will be also needed to discriminate more than two confusable phones.

## References

- C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 2(2): 955-974, 1998.

- A. Ganapathiraju and J. Picone, "Hybrid SVM/HMM architectures for speech recognition," Neural Information Processing Systems, 2000.
- M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez and T. Wang, "Landmark-Based Speech Recognition: Report of The 2004 Johns Hopkins Summer Workshop", ICASSP, 2005.
- C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", ICASSP, 1990.
- T. Joachims, SVMlight: Support Vector Machine, <http://svmlight.joachims.org/>, University of Dortmund, Feb. 2004.
- L. Mangu, E. Brill and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," Computer, Speech and Language, 14(4): 373-400, 2000.
- L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, Feb. 1989.
- S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Vlatchev and P. Woodland, "The HTK Book (for HTK version 3.2.1)," 2002.