# Novel Time Domain Multi-class SVMs for Landmark Detection

*Rahul Chitturi, Mark Hasegawa Johnson**

University of Texas at Dallas, USA
rahul.ch@crss.utdallas.edu

* University of Illinois at Urbana Champaign, USA
jhasegaw@uiuc.edu

10.21437/Interspeech.2006-105

## Abstract

The training of precise speech recognition models depends on accurate segmentation of the phonemes in a training corpus. Segmentation is typically performed using HMMs, but recent speech recognition work suggests that the transient acoustic features characteristic of manner-class phoneme boundaries (landmarks) may be more precisely localized using acoustic classifiers specifically designed for the task of landmark detection. This paper makes an empirical exploration of new features which suit Landmark Detection and the application of Multi-class SVMs that are capable of improving the time alignment of phoneme boundaries proposed by Binary SVMs and HMM-based speech recognizer. On a standard benchmark data set (A database of Telugu - Official Indian Language, spoken by 75 million people), we achieve a new state-of-the-art performance, reducing RMS phone boundary alignment error from 32ms to 22ms.

**Index Terms**: Landmark, Multi Class SVM, Time Domain flatness measure, Segmentation

## 1 Introduction

Accurately segmented training data improves the precision of both speech recognition and speech synthesis models. Phonetic segmentation has also been proposed as part of a lattice rescoring algorithm for multipass speech recognition [1]. Since manual segmentation of speech is time consuming and unrealistic in most conditions, various approaches on automatic speech segmentation have been proposed [2, 3], most typically including forced alignment of an HMM-based sentence model [4,5] observing standard speech recognition features such as energy, LPCC, MFCC, and PLP.

In 1998, Niyogi proposed the use of support vector machines (SVMs) to detect stop consonant bursts [6]. Instead of observing MFCCs once per 10ms, Niyogi's SVMs observed a much smaller feature vector (energy, lowpass energy, and spectral flatness) once per millisecond. Niyogi's SVMs significantly outperformed an HMM-based speech recognizer in this task [7]. Other authors have since used Niyogi's method to develop complete first-pass [8] and second-pass [1] speech recognition engines. As we proposed the time domain landmark detection techniques, we chose the features on the similar lines as Niyogi. We came up with a new feature which we named as Time Domain Flatness Measure gave the maximum performance for this problem.

Multi-class SVMs are usually implemented by combining several two-class SVMs. The one-versus-all method using winner-takes-all strategy and the one-versus-one method implemented by max-wins voting are popularly used for this purpose. In this paper we give empirical evidence to show that these methods are inferior to another one-versus-one method. The implementation of the binary SVMs was normal one-versus-one SVMs and so it is not explained in detail due to space constraints.

## 2 Related work

The usage of Multiclass SVMs in speech was introduced in 2002, by *Salomon et al*. for phoneme classification [9]. Recently in 2005, *Yin An-ron et al.* used this Multi Class SVM as an extension of binary classifier using ECOC (Hadamard Error-Correcting Output Code [10]. *Jorge Bernal-Chaves et al.* used the Multi class SVMs for isolated word recognition [11]. In 2006, *Andrew Hatch et al.* have used the same technique for Speaker Recognition [12]

In this article, we extend the usage of Multi-class SVMs for Phoneme segmentation with HMM segmentation system as the baseline. We also explore new features which suit Landmark Detection.

## 3 Database Description

All our experiments were conducted on the Telugu (Official Indian Language- spoken by 75 million people) database, which has over 600 sentences collected from a native speaker and contains substantial coarticulatory effects. This database was collected at a 16 KHz sampling frequency and a single channel, using a headset. This database has well defined transcriptions corresponding to the speech data and even the phone-level alignment was done manually. For the purpose of exact evaluation of our algorithms, we used the manual phone-alignments, but we don't assume that these are necessary.

## 4 Baseline System

The goal of the algorithms proposed in this paper is to refine the phoneme boundary times proposed by the HMM, in order to reduce their RMS error with reference to a "ground truth" manual phonetic transcription. Though HMMs are known to work better for tri-phone models, since we are concerned only with the time boundaries of a particular phone, monophone based HMM segmentation system is done. The 39-element feature vector includes 12 MFCCs plus sum-squared signal energy, with delta and delta-delta coefficients appended. There are a total of 48 phones as shown in Fig. 1. Each phone is modeled by 3 emitting states, with 3 Gaussians per state. HMMs were initialized flat (to the same mean and variance), then HMMs were trained using embedded re-estimation for 15 iterations. Finally, forced alignment was done to get the phone boundaries. This system achieved an RMS phone boundary alignment error of 32ms

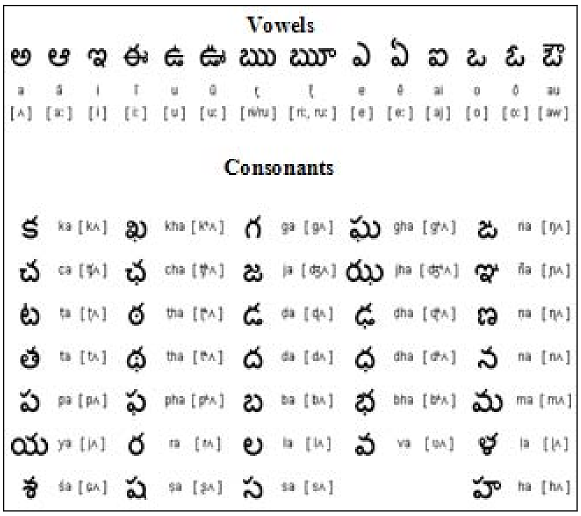


Figure 1- Phones in Telugu

## 5 Multi-class SVMs

In this section, we briefly review the implementations of all multiclass methods that will be studied in this paper. For a given multiclass problem, *M* will denote the number of classes and $\omega_i$, i=1,…, M. will denote the *M* classes. If the output of each binary classifier can be interpreted as the posterior probability of the positive class, Hastie and Tibshirani [13] suggested a *pairwise coupling* strategy for combining the probabilistic outputs of all the one-versus-one binary classifiers to obtain estimates of the posterior probabilities $p_i$=Prob ($\omega_i$ |$x$), i=1… M. After these are estimated, the PWC strategy assigns the example under consideration to the class with the largest $p_i$.

The actual problem formulation and procedure for doing this are as follows. The binary classifier *Cij* is trained taking the examples from $\omega_i$ as positive and the examples from $\omega_i$ as negative. Let us denote the probabilistic output of $C_{ij}$ as $r_{ij}$ = Prob($\omega_i \mid \omega_i\ or \omega_j$). To estimate the $p_i$'s, M(M-1)/2 auxiliary variables $\mu_{ij}$ 's which relate to the $p_i$'s are introduced: $\mu_{ij} = p_i / (p_i + p_j)$. $p_i$'s are then determined so that $\mu_{ij}$ 's are close to $r_{ij}$ 's in some sense. The Kullback-Leibler distance between $r_{ij}$ and $\mu_{ij}$ is chosen as the measurement of closeness:

$$l(p) = \sum_{i<j} n_{ij} \left(r_{ij} \log \frac{r_{ij}}{\mu_{ij}}\right) + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}} \quad (1)$$

where $n_{ij}$ is the number of examples in ($\omega_i \cup \omega_j$).

The associated score equations are:

$$\sum_{j \neq i} n_{ij} \mu_{ij} = \sum_{j \neq i} n_{ij} r_{ij}, i = 1,..,M \text{ Subject to} \sum_{k=1}^{M} p_k = 1 \quad (2)$$

The $p_i$'s are computed using the following iterative procedure 1.

Let $\overline{p}_i = 2 \sum_j r_{ij} / k(k-1)$. Hastie and Tibshirani [13] showed that the multi-category classification based on *~pi*'s is identical to that based on the $p_i$'s obtained from pairwise coupling. However, *~pi*'s are inferior to the $p_i$'s as estimates of posteriori probabilities. Also, log-likelihood values play an important role in the tuning of hyper parameters. So, it is always better to use the $p_i$'s as estimates of posteriori probabilities. Kernel logistic regression (KLR) has a direct probabilistic interpretation built into its model and its output is the positive class posterior probability. Thus KLR can be directly used as the binary classification method in the PWC implementation. We will refer to this multiclass method as PWC KLR.

1. Start from an initial guess of $p_i$'s and the corresponding $\mu_{ij}$ 's
2. Repeat{i=1,…,M…} until convergence
   - $p_i \leftarrow p_i \cdot \dfrac{\sum_{j \neq i} n_{ij} r_{ij}}{\sum n_{ij} \mu_{ij}}$
   - renormalize the $p_i$'s
   - recompute $\mu_{ij}$ 's

Procedure 1

The output of an SVM, however, is not a probabilistic value, but an un-calibrated distance measurement of an example to the separating hyperplane in the feature space. Platt [14] proposed a method to map the output of an SVM into the positive class posterior probability by applying a sigmoid function to the SVM output:

$$\Pr ob(\omega_1 \mid x) = \frac{1}{1 + e^{Af+B}} \quad (3)$$

where *f* is the output of the SVM associated with example **x**. The parameters *A* and *B* can be determined by minimizing the negative log-likelihood (NLL) function of the validation data. A pseudo-code for determining *A* and *B* is also given in [14]; To distinguish from the usual SVM, we refer to the combination of SVM together with the sigmoid function mentioned above as PSVM. The multiclass method that uses Platt's probabilities together with PWC strategy will be referred to as PWC PSVM.

## 6 Phone Boundary Refinement using SVMs

Landmarks are the boundaries between phones of different manner class. For refining the landmarks, the phonetic feature theory provides a hierarchical framework [6] and support vector machines (SVMs) provide the methodology for combining the speech knowledge [15]. The success of SVMs has been demonstrated for the problem of detection of stop consonants [7]. This article proposes the use of heterogeneous classifiers, including SVMs, in a second-pass speech segmentation system. First segmentation is done using HMM techniques and then the segmented data is sent to the phone boundary refiner.

### 6.1 Refinement Using Multi-Class SVMs

Previous landmark-based speech recognition systems [7, 8, 1] used SVMs as a kind of nonlinear filter. In these nonlinear-filtering methods, an SVM is applied sequentially to each frame of the input speech, with context taken into account using input memory buffers; if the SVM output exceeds a threshold in any frame, that frame is marked as a potential landmark. The nonlinear-filter application of SVMs was compared to that of a novel multi-class SVM-based phone boundary refinement algorithm, described in the remainder of this section. In all experimental tests, the multi-class SVM-based refinement algorithm outperformed the filter-style landmark detection algorithm.

Figure 2 shows a sample window which we use for extracting the features and the classes of the SVMs. The frame number in which the manually labeled phone boundary falls with respect to the HMM-proposed boundary is considered to be the target class. This can be understood by looking at the following sample window. In this, the target class for the given sample is 9-6 = 3. The feature vector is the vector of energies of the frames in the window, i.e., 8 frames before and 4 frames after the boundary proposed by the HMM. Frames for SVM analysis were 5ms in length.

```
                    man        HMM
                     |          |
F1  F2  F3  F4  F5  F6  F7  F8  F9  F10 F11 F12 F13
                     |          |
```

Figure 2- Sample window

We use the procedure2 for extracting the features and the corresponding classes.

> *for all HMM phone boundaries* $\boldsymbol{l_i}$ *in the signal*
>     $\boldsymbol{C_i} = \boldsymbol{f_{ik}}$
>     *for all the frames* $\boldsymbol{f_{ij}}$ *of 5 ms in the window of (* $\boldsymbol{l_i}$ *– 40)ms to (*$\boldsymbol{l_i}$*+20) ms*
>         $\boldsymbol{F.V_{ij}} = energy(\boldsymbol{f_{ij}})$
>     *end*
> *end*

Procedure 2

where, $\boldsymbol{f_{ik}}$ is the frame number in which the manually labeled boundary falls with respect to the HMM boundary, $\boldsymbol{C_i}$ is the target class label of SVM for the token $\boldsymbol{l_i}$ *and* $\boldsymbol{F.V_i}$ is the feature vector for the token $\boldsymbol{l_i}$. Using these feature vectors and the corresponding class labels we can train the SVMs.

### 6.2 Acoustic Features Observed by the SVM

In order to get high accuracy for classification and segmentation, good features need to be selected that can model the temporal and spectral structures of audio [16]. In this paper we first test standard speech recognition features such as short time energy and MFCC. In addition to the standard features, various new features like spectral flatness and time domain flatness measures are analyzed and tested for SVM-based phone boundary refinement.

#### 6.2.1 Bandpass dE/dT

The energy of a speech signal passed through a number of different bandpass filters can be used to distinguish different manner classes [17]. We tested, as features, the time-domain derivatives of the following energies:

0- 400 Hz: useful to detect voicing
300 -1000 Hz: useful to detect sonority
1000 - 3000 Hz: distinguish nasals from glides
2000 -6000 Hz: detect frication energy
Full Band (no filtering): detect stop closures

#### 6.2.2 Spectral Flatness Measure

Temporal changes in the spectra of speech signals are believed to play an important role in human perception. Useful spectral shape information can be derived from very short temporal windows using the spectral flatness measure. Spectral flatness is calculated by the following formula [7]:

$$\int \log(S(f,t))df - \log(\int S(f,t)df) \quad (4)$$

#### 6.2.3 Time Domain Flatness (TDF) Measure

In analogy to Niyogi's spectral flatness measure, we define a "time domain flatness" measure. Energy as a function of time is flat during a quasi-static region, and non-flat at a landmark. Our intuitive notion of "time domain flatness" can be formalized as follows:

$$\int \log(E(t))dt - \log(\int E(t)dt) \quad (5)$$

where the integral is approximated by a sum over 7 sequential frames. Fig. 3 shows the plot of the flatness measure. Most of the actual landmarks are near the local minima of the TDF measure.

## 7 Experimental Results

This section compares the features and algorithms described in the previous sections. These experiments were conducted on the database that is described in Sec. 3. Multi-class phone boundary refinement SVMs were trained and tested using, individually and in combination, the features described in Sec. 6.2. Results are shown in Table 1. The second column represents the percentage of phones that have alignment error less than 5 ms. and similarly 10ms in the third column. Similarly in Table 2 we have different bands of deviation and the corresponding performance of algorithms. The most accurate SVM shows the Time Domain Flatness feature as the best feature.

Since the results vary with the speech database, and as there is no benchmark speech database for the Telugu Language, we have only compared our algorithms with the methods that are usually employed like HMMs, on the database that was described in Section 2. The train data set has nearly 500 sentences and the test data set has 100 sentences, each sentence having approximately 40-50 phonemes (10-15 words).

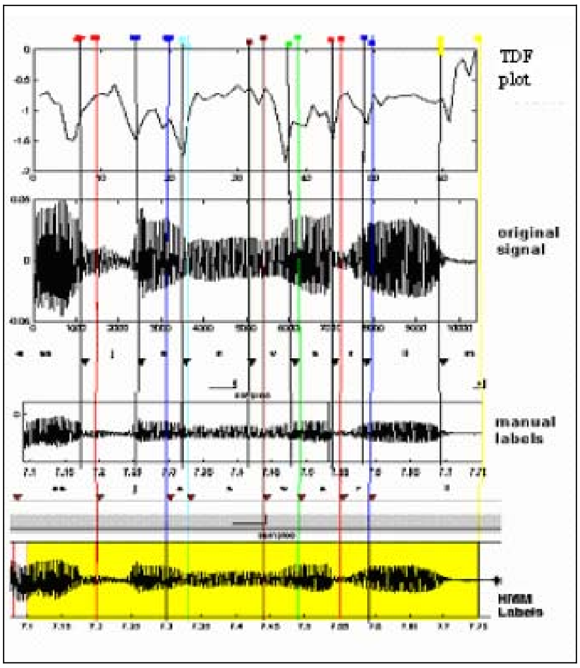


Figure 3- Plot of Time Domain Flatness measure

| **Feature** | **< 5ms** | **<10 ms** |
|---|---|---|
| Individual Bandpass dE/dt | 21% | 31% |
| Combined Bandpass dE/dt | 19% | 30% |
| Spectral Flatness | 18.5% | 34% |
| Time Domain Flatness | 24% | 36% |
| Time domain Flatness+ Bandpass Energies | 19% | 28% |

Table 1: Percentage of phone boundary alignment errors below threshold, as a function of the acoustic feature observed using Multi-class SVMs.

| **Deviation** | **% of phones using** | | |
|---|---|---|---|
| | **Multi-SVM** | **Bin-SVM** | **HMM** |
| <5ms | 24% | 8% | 3.82% |
| <10ms | 36.3% | 11.2% | 8.7% |
| <15ms | 47.2% | 21% | 18% |
| <20ms | 55.8% | 38% | 34.3% |
| Average | 22 ms | 26.1 ms | 32 ms |

Table 2: Overall percentage of phone boundary alignment errors below threshold with Multi class SVMs.

## 8 References

[1] M. Hasegawa-Johnson et al *Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop.* Technical report

[2] Y.-J. Kim, A. Conkie, "Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction," in: Proc. ICSLP2002, Denver, Colorado, 145- 148, Sept. 2002.

[3] Syrdal, AK, Hirschberg, J., McGory, J. and Beckman, M., "Automatic ToBI prediction and alignment to speed manual labeling of prosody," In Speech. Communication, vol. 33, 135-151.

[4] K. Tokuda, H. Zen, A.W. Black, ``An HMM-based speech synthesis system applied to English,'' *IEEE Speech Synthesis Workshop*, 2002.

[5] A. Sethy and S. Narayanan, "Refined speech segmentation for concatenative synthesis." *ICSLP*, 2002.

[6] P. Niyogi, "Distinctive Feature Detection Using Support Vector Machines." *ICASSP* pp 425-428, 1998.

[7] P. Niyogi, C. Burges, P. Ramesh, ``Distinctive Feature Detection using Support Vector Machines,'' *Proc. ICASSP*, 1999

[8] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines." *Proc. International Joint Conference on Neural Networks*, 2003

[9] *Jesper Salomon, Simon King* FRAMEWISE PHONE CLASSIFICATION USING SUPPORT VECTOR MACHINES – ICASSP 2002

[10]Yin An-rong, Xie Xiang, Kuang Jing-ming- *Using Hadamard ECOC in multi-class problems based on SVM* - Interspeech 2005

[11] *Jorge Bernal-Chaveset al* Multiclass SVM-Based isolated-digit recognition using a HMM-guided segmentation- NOLISP 2005

[12] *Andrew Hatch et al* Generalized Linear Kernels for One-Versus-All Classification: Application to Speaker Recognition –ICASSP 2006

[13] Hastie, T., Tibshirani, R. (1998) Classification by pairwise coupling. In Jordan, M.I., Kearns, M.J., Solla, A.S. (eds.), *Advances in Neural Information Processing Systems*, Vol. 10.

[14] Platt, J. (1999) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola, A.J., Bartlett, P., Sch¨olkopf, B., Schuurmans, D. (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press.

[15] P.Niyogi and M. M. Sondhi. "Detecting Stop Consonants in Continuous Speech" *Journal of the Acoustical Society of America.* 111, 1063 (2002)

[16] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator." *Proc ICASSP* pp 1331-1334, 1997

[17] S. Liu, "Landmark detection for distinctive-feature based speech recognition." *J. Acoustical Society of America* **100**(5):3417-3430, 1996