Distinctive Feature Based SVM Discriminant Features for Improvements to Phone Recognition on Telephone Band Speech

Sarah Borys and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
Urbana-Champaign, Urbana, IL 61801
{sborys, jhasegaw}@uiuc.edu

Abstract

Support vector machines (SVM's) can be trained to classify manner transtions between phones and to identify the place of articulation of any given phone with high accuracy. The discriminant outputs of these SVM's can be used as input features for a standard ASR system. There is a significant improvement in correctness and accuracy using these SVM discriminant features when compared to an MFCC based recognizer of equal parameters.

1. Introduction

A distinctive feature based approach to speech recognition is an approach that assumes a speech signal is comprised of articulatorily-motivated phonological features that have canonical acoustic properties. Stevens [1] proposes that the variability between acoustic correlates can be reduced if those acoustic correlates are examined by means of phonetic features. Using distinctive features and information about the human vocal tract, Keyser and Stevens [2] were able to devalope a hierarchical model of co-articulation and phonology. Stevens et al. [3] were also able to build a model based on distinctive feature heirachries to access words in a lexicon using measurements from key locations in the speech signal.

"Landmarks" are abrupt changes in the speech signal that correspond to a change in distinctive features. Juneja [4] reports that he is able to detect various landmarks using support vector machines (SVM's) with high accuracy. Nyogi and Burges [5] are also able to use SVM's to detect stop consonants with high accuracy using SVM's. SVM's are advantageous as landmark detectors because they are able to be trained on small data sets with large feature vectors to identify landmarks with high accuracy.

How best can distinctive features and SVM landmark detectors be integrated into speech recognition? In this paper, we attempt to integrate distinctive features and SVM's into ASR by (1) training a set of SVM's to distinguish between phonetic features, (2) using those SVM's to generate discriminant vectors for an entire database and (3) training an HMM based recognizer on these new discriminant features. This work extends from work done at the Johns Hopkins 2004 Summer Workshop [6].

2. Distinctive Features

Distinctive features [7] allow for an economical way of classifying phone segments and also allow for a better understanding of allophonic variation. Each phone can be classified by a unique set of binary valued (either positive (+) or negative (-)) distinctive features. There are two categories of distinctive features, articulator free and articulator bound/

An articulator free (manner) feature is a parameter of phonological structure that encodes a perceptually salient aspect of speech production. The five manner features we are primarily concerned with are *speech*, *continuant*, *sonorant*, *syllabic* and *consonantal*. The feature *silence* specfies whether a sound was created by the human vocal apparatus (-silence) or wheather it is silence or other ambient noise (+silence). *Continuant* describes the free airflow through the oral cavity. A phone that is +*continuant* is made with air flowing through the mouth. *Sonorant* determines how resonant a phone is. Sounds that are +*sonorant* are sounds that tend to resonant and have more acoustic energy than -*sonorant* sounds. +*Syllabic* sounds are those that can occur in the nucleus of a syllable. *Consonantal* determines if there is a narrow constriction in the oral cavity (+*consonantal*). Manner features allow for a heirarchy in which phones to be grouped into broad class categories such as vowels, glides, nasals, stops and fricatives.

An articulator bound (place) feature is a parameter that describes a physical, articulator dependent aspect of human speech production. Place features that can be sensibly defined for a phone are manner dependent, i.e. (most) different manner classes will have different place features.

Nasals can be charactarized by the features *alvelar*, *labial* and *velar*. *Alvelar* sounds are those that are made by pressing the tongue blade to the back of the alvelar ridge, as in the nasal /n/. *Labial* sounds are created by pressing the lips together. The sound /m/ is the labial nasal. Finally, *velar* sounds are made using the velum, the membrane that separates the mouth and the nose. The velar nasal is /ng/.

Stops are also characterized by the features *alvelar* (/t/, /d/), *labial* (/p/, /b/), *velar* (/k/, /g/). In addition, stops are also classified by the feature *voice*. Voiced sounds (+*voice*) are those that are made with the vibration of the vocal folds. The sounds /p/, /t/ and /k/ are unvoiced whereas /b/, /d/ and /g/ are voiced.

Fricatives can be described by the features *anterior*, *dental*, *labial*, *strident* and *voice*. *Anterior* fricatives are created in the paleto-alvelar region of the mouth, such as /s/. A phone with the feature *dental*, is realized using the teeth. The phone /th/ is an example of a dental fricative. *Strident* fricatives are those that have an obstacle placed in front of the constriction in the vocal tract, as in the phone /z/. (fricatives are +*continuant*, +*consonantal*). An example of a labial frictive is the sound /f/ and an example of a voiced fricative would be the sound /zh/.

The glides, /h/, /l/, /r/, /w/, and /y/ are unique in that they are each articulated with a different region of the oral cavity. Therefore, our place features for glides are simply *h*, *l*, *r*, *w* and *y*.

Vowels are defined by the features *advanced tongue root*, *front*, *high* , *low*, *reduced*, *round* and *tense*. The features *front*, *low* and *high* describe the tongue tip position during production of the vowel. A vowel with the feature *advanced tongue root* is produced with a widened pharynx (ey vs. ih). *Round* describes if there is lip rounding during vowel production. The vowel /uw/ is +*round*. *Tense* vowels, like /aa/, are usually longer in duaration, have a higher pitch and higher tongue position than lax (non-tense) vowels, such as /uh/. Vowels that are *reduced* are generally unstressed, such as the schwa /ax/.

3. The Corpus

The NTIMIT corpus [8] was used for the experiments described in section 5. NTIMIT was constructed by filtering the utterances from the original phonetically rich TIMIT database [9] through telephone channels and redigitizing them. The NTIMIT utterances been time aligned with the original TIMIT data allowing for the use of the detailed TIMIT phonetic transcriptions with the NTIMIT data.

4. Support Vector Machines

4.1. Background

The following subsection gives a brief background of support vector machines (SVM's). A more detailed explanation can be found in [10].

Given a set of N binary observations $y_i \in \{-1, 1\}$ each associated with a feature vector $\mathbf{x_i} \in R^n$ drawn from an unknown probability distribution $P(\mathbf{x_i}, y_i)$ we wish to find the optimal mapping function $f(\mathbf{x_i}, \alpha)$, where α is a set of adjustable model parameters. We define the expected error as

$$R(\alpha) = \frac{1}{2}\int |y - f(\mathbf{x}, \alpha)|\, dP(\mathbf{x}, y) \tag{1}$$

and the empherical error as

$$R_{emp}(\alpha) = \frac{1}{2N}\sum_{i=1}^{l} |y_i - f(\mathbf{x_i}, \alpha)| \tag{2}$$

where the quantity $\frac{1}{2}|y_i - f(\mathbf{x_i}, \alpha)|$ is referred to as the loss and can be equal only to 0 or 1 for binary classification. For a loss taking on these values, with probability $1 - \eta$, an upper bound for $R(\alpha)$ can be defined as follows

$$R(\alpha) \le R_{emp}(\alpha) + \frac{\sqrt{d(\log(\frac{2N}{d} + 1) - \log(\frac{\eta}{4}))}}{N} \tag{3}$$

where η is defined such that $0 \le \eta \le 1$ and d is a positive integer called the Vapnik Chervonenkis (VC) dimension.

We assume now that our data are separable, that is, a hyperplane can be drawn that completely isolates one observation set from another. It can be further assumed that the data are seperable with " margin" $\frac{2}{||\mathbf{w}||}$ (the distance between the clouds $y_i = 1$ and $y_i = -1$ is $\frac{2}{||\mathbf{w}||}$) under the following conditions. If the vector $||\mathbf{w}||$ is defined to be a vector normal to the hyperplane and $\frac{|b|}{||\mathbf{w}||}$ is the perpendicular distance from the hyperplane to the origin, where $||\mathbf{w}||$ is the Euclidean norm of $\mathbf{w}$, then the grouping conditions for all observable examples in the data set are

$$y_i(\mathbf{x_i} \cdot \mathbf{w} + b) - 1 \ge 0 \;\; \forall\, i \tag{4}$$

The ith constraint is satisfied if and only if the ith example is correctly classified by the hyperplane.

Therefore, if $R_{emp}(\alpha) = 0$, $R(\alpha)$ can be minimized by minimizing $||\mathbf{x}||$ subject to constraints i. For each constraint, a Lagrange multiplier, α_i, is introduced. We minimize

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l}\alpha_i y_i(\mathbf{x_i} \cdot \mathbf{w} + b) + \sum_{i=1}^{l}\alpha_i \tag{5}$$

To find the optimal L, the Lagrangian must be minimized with respect to $\mathbf{w}$ and b. Setting $\frac{dL}{d\mathbf{w}} = 0$ and $\frac{dL}{db} = 0$, we get

$$\begin{aligned} \sum_i \alpha_i y_i \mathbf{x_i} &= \mathbf{w} \\ \sum_i \alpha_i y_i &= 0 \end{aligned} \tag{6}$$

Inserting equation 6 into equation 5 yields:

$$L = \sum_i a_i + \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{7}$$

which satisfies all conditions and therefore, provides an optimal Lagrangian.

In most cases, the data cannot be separated by a hyperplane and are therefore non-separable. In the case of non-separable data, equation 7 is no longer true so minimizing 5 is no longer equivalent to minimizing $||\mathbf{w}||^2$. Non-separable data can be accounted for by defining the positive slack variable, ξ_i $(1 \le i \le L)$ and rewriting equation 4 as follows

$$\begin{aligned} \mathbf{x_i} \cdot \mathbf{w} + b &\ge 1 - \xi_i \; for \; y_i = 1 \\ \mathbf{x_i} \cdot \mathbf{w} + b &\le -1 + \xi_i \; for \; y_i = -1 \\ \xi_i &\ge 0 \end{aligned} \tag{8}$$

For a classification error to occur, ξ_i must exceed unity. An upper bound on the number of training errors can be determined to be $R(\alpha) \le \sum_i \xi_i$. Thus, equation 3 becomes

$$L = \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i - \sum_i \alpha_i y_i(\mathbf{x_i} \cdot \mathbf{w}_b) + \sum_i \alpha_i \tag{9}$$

where C is the error penalty. The optimal Lagrangian must satisfy the following conditions.

$$\begin{aligned} 0 \le \;\; \alpha_i \;\; &\le C \\ \mathbf{w} &= \sum_i \alpha_i y_i \mathbf{x_i} \\ \sum_i a_i y_i &= 0 \end{aligned} \tag{10}$$

The optimal Lagrangian is given by the equation

$$L = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{11}$$

A non-planar separatrix can be constructed by using $K(\mathbf{x_i}, \mathbf{x_j})$ in place of $\mathbf{x_i}$ and $\mathbf{x_j}$ in the equations above, where $K(\mathbf{x_i}, \mathbf{x_j})$ is a positive definite "kernel" function. Aside from the linear kernel, other common kernels are the radial basis function (RBF), polynomials and sigmoid functions. In our experiments, the RBF kernel based SVM's had the highest accuracies. The RBF kernel is given by the equation

$$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-\gamma|\mathbf{x_i} - \mathbf{x_j}|^2} \tag{12}$$

SVM	%ACC	SVM	%ACC
-+Silence	92.1	+-Silence	91.6
-+Continuant	79.7	+-Continuant	81.1
-+Sonorant	86.4	+-Sonorant	91.1
-+Syllabic	88.6	+-Syllabic	78.5
-+Consonantal	78.1	+-Consonantal	73.1

Table 1: Accuracies of the manner change SVM's.

SVM	%ACC	SVM	%ACC
Adv. Tongue Root	86.0	Low	87.6
Alvelar (Nasal)	78.9	R	87.5
Alvelar (Stop)	85.9	Reduced	84.5
Anterior	89.1	Round	98.6
Dental	87.7	Strident	83.5
Front	86.2	Tense	76.8
H	93.8	Velar (Nasal)	95.3
High	91.0	Velar (Stop)	89.8
L	84.7	Voice	74.7
Labial (Fricative)	82.8	W	89.2
Labial (Nasal)	84.5	Y	92.8
Labial (Stop)	88.5		

Table 2: Accuracies of the place detection SVM's.

4.2. SVM Landmark Detection

A set of 33 RBF kernel SVM's were trained to make distinctions between 10 manner transitions and 23 place features. These distinctions are listed in tables 3 and 4 respectively.

Each manner change SVM was trained using 6500 examples from each class. Each feature vector contains 11 concatinated 25ms windows (with a 10ms skip between each window) of "information" about the speech signal. The first window in each vector is sampled at 50ms before the landmark, the 6th window is sampled at the landmark time and the final window is sampled 50ms after the landmark has occured.

Place classification SVM's were trained using the maximum number of available samples per class. Each feature vector containes information in 7 25ms windows starting at the landmark frame and extending forward in time. The only exception is the SVM trained to distinguish velar nasals from non-velars. This SVM has a feature vector with a format identical to that of the manner change SVM's.

The "information" included in a feature vector (both manner and place) was MFCC's, delta coeficients, acceleration coefficients, formants [12] and APS [13]. Given the included acoustic information, manner-change SVM's (and the velar nasal detector) were trained on a feature vecotr of size 1023 and place distinction SVM's were trained on a feature vector of size 651. Results for the manner change and place classification SVM's are listed in tables 1 and 2 respectively.

5. Experiments

5.1. The Baseline System

The baseline system was an HMM phone recognizer constructed using the Hidden Markov Toolkit (HTK) [11]. Each phone was modeled with a 5-state (3 emmitting states) HMM with 3 gaussian mixtures per state. Mixtures were vectors of length 33 (10 MFCC, 10 delta, 10 acceleration and 3 energy coefficients). This vector length was chosen to equalize the number of parameters between the baseline system and the SVM discriminant feature system.

-+Silence -+Continuant -+Sonorant -+Syllabic
+-Silence +-Continuant +-Sonorant +-Syllabic
-+Consonantal +-Consonantal

Table 3: A list of the manner transition features for which an SVM discriminant value was output and used in the feature vector for recognition. A "-+" indicates a transition from a segment with a property of *-feature* to a segment with a property of *+feature*. A "+-" indicates the reverse. For a detailed description of the meaning of these features, see section 2

5.2. The Distinctive Feature Based SVM System

The distinctive feature based SVM (DFSVM) system modeled phones using a 5-state (3 emmitting states) HMM with 3 mixtures per state. Features for these models consisted of a vector of discriminant values generated by the SVM's described in section 4. Each feature vector contained information about 10 different manner transition types (listed in table 3) and 23 different cues as to the place of articulation (listed in table 4). HMM's were trained using HTK [11].

Alvelar (nasals)
Alvelar (stops)
Anterior
Advanced Tongue Root
Dental
Front
H
High
L
Labial (fricatives)
Labial (nasals)
Labial (stops)
Low
R
Reduced
Round
Strident
Tense
Velar (nasals)
Velar (stops)
Voice
W
Y

Table 4: A list of the place features for which an SVM discriminant value was output and used in the feature vector for recognition. For a detailed description of the meaning of these features, see section 2

5.3. Results

The results for the baseline and DFSVM systems are given in table 5. The DFSVM system shows an improvements of 4.22% in correctness and 1.19% in accuracy.

6. Conclusion and Future Work

In this paper, we presented a method for integrating SVM's amd HMM's into a recognition system. We also presented a system based on this method which shows improvment over a baseline HMM only system.

The next step in our work will be to extend the distinctive feature set in our DFSVM recognizer. We will also attempt to improve the accuracy of our existing SVM's so that each

	%Corr	Acc
Baseline	38.00	36.06
DFSVM	42.22	37.25

Table 5: Correctness and accuracy for the baseline and distinctive feature based SVM phone recognition systems described in sections 5.1 and 5.2.

SVM performs with at least 80work will also be extended to the Switchboard corpus in the near future.

There is an alternative integration approach to the one that we have shown here. Perhaps instead of only using a single HMM recognizer, it would be advantageous to use the 10 manner change features listed in table 3 to first recognize only the linguistic class; nasal, stop, vowel, fricative, glice, silence, etc. The unidentified phone would then be processed by a second recognizer (based solely on place features) designed specifically to handle the given manner class.

7. References

[1] K. Stevens, *Relational properties as perceptual correlates of phonetic features*. International Conference of Phonetic Sciences, 1987. 352-355

[2] S. Keyser, N. Stevens, *Feature geometry and the vocal tract*. Journal of Phonetics, 1999.

[3] K. Stevens, S. Manual, S. Shattuck-Hufragel, S. Liu, *Implementation of a model for lexical access based on features* International Conference on Spoken Language Processing, 1992.

[4] A. Juneja, *Speech recognition using acoustic landmarks and binary phonetic feature classifiers*. PhD. Thesis Proposal, University of Maryland, 2003.

[5] P. Niyogi, C. Burges, *Detecting and implementing acoustic features by support vector machines*. University of Chicago Tech Report TR-2002-02.

[6] WS04 FINAL REPORT

[7] K. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1999.

[8] Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Ju dith Spitz. *NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database*. Proceedings of ICASSP-90, April 1990.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, *The DARPA TIMIT acoustic phonetic speech corpus*, NIST, 1993.

[10] C. Burges. *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery, Vol. 2, No. 2, 1998, pp 1-47.

[11] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland, *The HTK book*. Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, 2002.

[12] Y. Zheng and M. Hasegawa-Johnson. Formant tracking by mixture state particle filter. In *Proc. ICASSP*, 2004.

[13] N. Bitar. *Acoustic Analysis and Modelling of Speech Based on Phonetic Features*. PhD thesis, Boston University, 1998.