Multiclass Support Vector Machines for Articulatory Feature Classification

Brian Hutchinson and **Jianna Zhang**
Computer Science Department
Western Washington University
516 High Street, Bellingham, WA 98225-9062
{brian.hutchinson,jianna.zhang}@wwu.edu

This ongoing research project investigates articulatory feature (AF) classification using multiclass support vector machines (SVMs). SVMs are being constructed for each AF in the multi-valued feature set (Table 1), using speech data and annotation from the IFA Dutch "Open-Source" (van Son *et al.* 2001) and TIMIT English (Garofolo *et al.* 1993) corpora. The primary objective of this research is to assess the AF classification performance of different multiclass generalizations of the SVM, including one-versus-rest, one-versus-one, Decision Directed Acyclic Graph (DDAG), and direct methods for multiclass learning. Observing the successful application of SVMs to numerous classification problems (Bennett and Campbell 2000), it is hoped that multiclass SVMs will outperform existing state-of-the-art AF classifiers.

One of the most basic challenges for speech recognition and other spoken language systems is to accurately map data from the acoustic domain into the linguistic domain. Much speech processing research has approached this task by taking advantage of the correlation between phones, the basic units of speech sound, and their acoustic manifestation (intuitively, there is a range of sounds that humans would consider to be an "e"). The mapping of acoustic data to phones has been largely successful, and is used in many speech systems today. Despite its success, there are drawbacks to using phones as the point of entry from the acoustic to linguistic domains. Notably, the granularity of the "phonetic-segmental" model, in which speech is represented as a series of phones, makes it difficult to account for various subphone phenomena that affect performance on spontaneous speech.

Researchers have pursued an alternative approach to the acoustic-linguistic mapping through the use of articulatory modeling. This approach more directly exploits the intimate relation between articulation and acoustics: the state of one's speech articulators (e.g. vocal folds, tongue) uniquely determines the parameters of the acoustic speech signal. Unfortunately, while the mapping from articulator to acoustics is straightforward, the problem of recovering the state of the articulators from an acoustic speech representation, acoustic-to-articulatory inversion, poses a formidable challenge (Toutios and Margaritis 2003). Nevertheless, researchers have made much progress, and articulatory-based systems have been able to outperform phonetic-segmental systems, particularly on noisy and spontaneous speech (Chang 2002).

Feature	Values
Voicing	voiced,voiceless,sil
Manner	stop, fricative, nasal, glide, vowel, sil
Front-Back	back, backplus, front, frontplus, mid, nearfront, nil, sil
Rounding	rounded,unrounded,nil,sil
Place	alveolar, glottal, high, labial, labiodental, lateral, low, mid, palatal, velar, sil

Table 1: Articulatory Feature Set

Through the use of specific feature sets, researchers have successfully framed the AF classification problem into one solvable using standard SVM techniques. Juneja (2004) used SVMs to classify binary-valued AFs, while Toutios and Margaritis (2005) used support vector regression to recover articulatory trajectories (literal measurements of the articulators) from the speech signal. However, multi-valued AF classification has typically been conducted using multilayer perceptrons (MLPs) (Kirchhoff 1999; Chang 2002), and it is of interest to determine if classification results can be improved through the use of multiclass SVMs. The feature set defined for the current research consists of five AFs: voicing, manner, front-back, rounding, and place (Table 1). Using this linguistically meaningful, multi-valued set simplifies the integration of classification results into higher level linguistic systems (e.g. speech recognition), at the expense of somewhat abstracting away from the literal physiological measurements of articulation that are so closely tied to the acoustic signal, and with some additional computational burden.

The SVM is a binary classifier which has demonstrated excellent performance on a variety of classification tasks (Bennett and Campbell 2000). The learned decision boundary corresponds to the optimal separating hyperplane, which maximizes the margin between two sets of linearly separable vectors. Several variations of the SVM, designed to classify linearly inseparable data, are discussed in (Schölkopf and Smola 2002). Because the SVM can be trained and tested

using only inner products between input vectors, it is well-suited for the *kernel trick*. A kernel function K calculates

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

without explicitly applying the mapping ϕ, and at significantly reduced cost. The SVM can therefore perform efficient nonlinear classification using kernels, implicitly applying nonlinear map ϕ to the input space vectors to produce vectors in a more desirable (and often higher dimensional) feature space. The current research uses the popular Gaussian radial basis function kernel:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Several approaches have been proposed to generalize the binary SVM classifier to solve problems where there are $m > 2$ classes. The one-versus-rest approach trains m binary classifiers. Each classifier c_i learns the decision boundary between data points in class i and all other data points. The one-versus-one approach trains $(m-1)(m/2)$ classifiers, one for every distinct pair of classes. These approaches use a committee to decide the final classification verdict. A variation on the one-versus-one approach, DDAGs (Platt, Cristianini, and Shawe-Taylor 2000) determine the classification result by traversing a directed acyclic graph, eliminating one class at each decision node. Finally, there have been several methods proposed to directly optimize a multiclass SVM, though in general, these SVMs have been sub-optimal or slow to train (Schölkopf and Smola 2002). The current research assesses a computationally efficient direct method proposed by Crammer and Singer (2001). Multiclass SVM classification is still an active research area, and no single approach has proven optimal under all circumstances. Using the LIBSVM software (Chang and Lin 2001), each of these approaches will be compared to assess their appropriateness on the AF classification task.

The current research uses two phonetically transcribed speech corpora: the IFA Dutch Spoken Language Corpus, with 19,465 sentences from eight speakers, and the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, with 6,300 sentences from 630 speakers. Thus far, the research has been conducted using the IFA Corpus. Several subsets of the corpus have been defined, including those consisting entirely of vowel speech, spontaneous speech, and speech from a single speaker. The waveform data is split into 25 ms frames with a 10 ms offset. Each frame is processed to produce 12 mel-frequency cepstrum coefficients (MFCCs), augmented with log energy, first and second derivatives. To provide a temporal context, each input vector to the SVM consists not only of the coefficients for the given frame, but also those of prior and following frames. The phone labels corresponding to each frame are extracted from the phonetic annotation, and generate the AF labels via a rule-based lookup table (Table 2). Kernel and SVM parameters are set via grid search with 10-fold cross-validation. MLP classifiers, the traditional multi-valued AF classifier, serve as the performance baseline.

This research is ongoing, and experimental results will be available soon. The results will be compared to those of previous researchers (Kirchhoff 1999; Chang 2002), and will be communicated to the research community through additional publication. Further research is currently underway to perform cross-language phonetic transcription using articulatory features. For more information on this and related research projects visit http://studentweb.cs.wwu.edu/~hutchib2/research/.

Phone	Voice	Manner	...	Place
p	voiceless	stop	...	labial
n	voiced	nasal	...	alveolar
⋮	⋮	⋮		⋮
a	voiced	vowel	...	low

Table 2: Articulatory Feature Lookup Table

References

Bennett, K. P., and Campbell, C. 2000. Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations* 2(2):1–13.

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

Chang, S. 2002. *A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech Recognition*. Ph.D. diss., UC Berkeley.

Crammer, K., and Singer, Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* 2:265–292.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; Pallett, D. S.; and Dahlgren, N. L. 1993. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM.

Juneja, A. 2004. *Speech Recognition Based on Phonetic Features and Acoustic Landmarks*. Ph.D. diss., University of Maryland, College Park.

Kirchhoff, K. 1999. *Robust Speech Recognition Using Articulatory Information*. Ph.D. diss., Universität Bielefeld.

Platt, J.; Cristianini, N.; and Shawe-Taylor, J. 2000. Large margin DAGS for multiclass classification. In Solla, S.; Leen, T.; and Mueller, K.-R., eds., *Advances in Neural Information Processing Systems 12*, 547–553.

Schölkopf, B., and Smola, A. J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Toutios, A., and Margaritis, K. 2003. Acoustic-to-articulatory inversion of speech: A review. In *Proceedings of the International 12th TAINN*.

Toutios, A., and Margaritis, K. 2005. Mapping between the speech signal and articulatory trajectories. In *Proceedings of the 7th HERCMA*.

van Son, R.; Binnenpoorte, D.; van den Heuvel, H.; and Pols, L. C. 2001. The IFA Corpus: A phonemically segmented dutch "open source" speech database. In *Proceeedings EUROSPEECH*.