Articulatory Manner Features Recognition with Linear and Polynomial Kernels

Jan Macek, Julie Carson-Berndsen

School of Computer Science and Informatics
University College Dublin
Belfield, Dublin, Ireland
{jan.macek, julie.berndsen}@ucd.ie

Abstract

A typical speech recognition system uses acoustic features to represent speech for its processing. Recently, articulatory features were introduced to serve the same purpose. They are motivated by linguistic knowledge and may therefore provide better or complementary representation of speech signal. We present research on recognition of such articulatory features by Support Vector Machines with three types of kernels—a linear kernel and two polynomial kernels. As input for recognizers we use standard set of Mel-frequency cepstral coefficients extended with values of formants and pitch of the speech signal. Performance is compared to recent results for the task based on other methods of machine learning. We conclude that for most of the articulatory features SVMs with a polynomial kernel give superior performance.

Razpoznavanje značilk artikulatornega načina z linearnimi in polinomskimi jedri

Tipičen sistem razpoznavanja govora uporablja pri procesiranju za predstavitev govora akustične značilke. V zadnjem času so se z istim namenom začele uporabljati tudi artikulatorne značilke. Uporabo leteh je motiviralo jezikoslovno znanje, zato lahko morda omogočajo boljšo ali komplementarno predstavitev govornega signala. V prispevku predstavljamo raziskavo o tem, kako z metodo podpornih vektorjev (MPV) razpoznavamo artikulatorne značilke s tremi vrstami jeder z linearnim jedrom in z dvema polinomskima jedroma. Kot vhodne podatke za razpoznavalnike uporabljamo standardno množico melodičnih frekvennih kepstralnih koeficientov, razširjenih z vrednostmi formantov in osnovnih period govornega signala. Kakovost izvedbe primerjamo z nedavnimi rezultati za isto nalogo na podlagi drugih metod strojnega učenja. Sklenemo z ugotovitvijo, da dajo za večino artikulatornih značilk polinomske MPV najboljše rezultate.

1. Introduction

Speech representations today are usually based on the acoustic information of the signal (Heřmanský, 1999). However, by relying only on this acoustic information, these speech representations seem to achieve only moderate success, especially, in adverse environments (noisy, out-of-task, out-of-vocabulary, etc). One of the ways to improve performance in such environments is to integrate linguistic knowledge as suggested in (Launay et al., 2002; Carson-Berndsen, 1998).

Articulatory features (AF) have been shown to improve word recognition accuracy under variable conditions of speech signal production. For example, in a multilingual environment, feature recognizers trained on data from different languages were shown to have the capability of improving the overall performance by ensemble recognizer or by crosslingual recognizer (Stüker et al., 2003). The AF representations have also been shown to perform well in noisy environment (Kirchhoff, 1999).

AF is thought to be a good compromise, offering better descriptions of the acoustic signal than phonemes yet still providing a linguistically interpretable symbolic annotation. Acoustic correlates of features have been described in the literature (Stevens, 2000; Stevens, 1980). The first detailed description of distinctive features (Jakobson et al., 1952) assumed that they had identifiable counterparts.

In this paper, Support Vector Machines (SVMs) with three types of kernels are presented for extraction articulatory features from the speech signal. The performance of the SVMs is compared among them and against referenced results of bagging that are reported as giving best results for this task among machine learning methods. We only refer to reported performance of Hidden Markov Models on this task (Kanokphara et al., 2006) where they do not provide good performance, apparently for the reasons of weaker probabilistic dependence between adjacent articulatory features in the speech signal. Our article extends the research reported in (Kanokphara et al., 2006) and (Macek et al., 2005). The SVM classifiers with variable kernels were run with the SVMLight implementation (Joachims, 1999).

Systematically, this paper is organized as follows. Section 2. explains the details of the experimental paradigm used in this paper, i.e. the corpus, the evaluation method and the feature table. Section 3. describes the general framework of support vector machines and Section 4. presents the results of experiments with SVM-based AF extraction. Finally, discussion and conclusions are presented in Section 5.

2. Experimental Setup

2.1. The Corpus

In the experiments we used the standard TIMIT corpus (Garofolo et al., 1993) consisting of 6300 sentences, 10 sentences spoken by each of 630 speakers, of which 462 are in the training set and 168 are in the testing set. There is no overlap between the training and testing sentences, except 2 SA sentences that were read by all speakers. The training set contains 4620 utterances and the testing set contains 1680 utterances. The core test set, which is the abridged version of the complete test set, consists of 192 utterances, 8 from each of 24 speakers. In this paper, the full training set with SA sentences is used as the training set while only the core test set without SA sentences is used as the test set.

Articulatory manner feature	Frequency in corpus	Phone (TIMIT transcription used)
approximant	8.12%	axr, r, w, y
closure	9.68%	bcl, dcl, gcl, kcl, pcl, tcl
flap	0.78%	dx, nx
fricative	16.47%	ch, dh, f, hh, hv, jh, s, sh, th, v, z, zh
lateral approx.	3.37%	el, l
nasal	5.72%	em, en, eng, m, n, ng, nx
stop	16.22%	b, bcl, d, dcl, g, gcl, k, kcl, p, pcl, q, t, tcl
vocalic	37.99%	aa, ae, ah, ao, aw, ax, ax-h, ay, eh, er, ey, ih, ix, iy, ow, oy, uh, uw, ux

Table 1: Assignment of articulatory manner feature classes to phonemes and their frequency in the TIMIT corpus

2.2. The Evaluation

The evaluation method used in this paper is a comparison of overall accuracy in terms of frame error rate (FER) together with recall, precision and F1-measure. FER is widely used for articulatory feature extraction evaluation (Chang et al., 2005). In our method the speech signal is represented as a sequence of numeric vectors where each vector represents speech in each time frame. Therefore, the AF extraction systems are evaluated on the frame level. Due to variable distribution of classes for each articulatory feature it is necessary to extend the performance measure of accuracy with the values of *precision*, *recall*, and *F1-measure* that are used in Tables 3, 4 and 5. The *precision* is defined as the ratio

$$\frac{\text{number of correctly classified instances of class } c}{\text{number of instances classified as class } c}$$

and the *recall* is defined as the ratio

$$\frac{\text{number of correctly classified instances of class } c}{\text{number of instances of class } c}$$

The trade-off between *precision* and *recall* is measured by the value of *F1-measure* defined as

$$\frac{2 * \mathit{precision} * \mathit{recall}}{\mathit{precision} + \mathit{recall}}.$$

All these measures analyse the performance for each class individually.

A true AF evaluation should compare between a reference (annotated at the feature level) and a hypothesized AF transcription. However, due to the cost and difficulty of corpus construction process, no feature annotated reference exists. In this paper, we directly convert reference annotations at the phone level into reference annotations at the articulatory feature level. These annotations lack some of the coarticulation information which would be typically found in references directly annotated at the articulatory feature level. However, this is the only resource available and it is widely accepted as the reference transcriptions for AF evaluation. Such transcription was done for the TIMIT corpus according to assignment of articulatory manner feature classes to phonemes presented in Table 1. Among the manner features we included both, closure and stop, where stop might be considered as a sequence of a closure and a burst. This allows us to see from performance of the respective classifiers if the simpler feature is more distinctive which is the case in our experiments.

Articulatory feature	Accuracy	Precision	Recall	F1-measure
−approximant	**93.12%**	0.947	0.98	0.963
+approximant		0.637	0.387	0.482
−closure	**94.41%**	0.957	0.982	0.970
+closure		0.763	0.567	0.651
−flap	**99.76%**	0.989	0.995	0.991
+flap		0.000	0.000	0.000
−fricative	**97.19%**	0.975	0.992	0.983
+fricative		0.956	0.865	0.908
−lateral approx.	**96.63%**	0.971	0.995	0.983
+lateral approx.		0.473	0.124	0.196
−nasal	**96.55%**	0.977	0.987	0.982
+nasal		0.636	0.504	0.562
−stop	**89.64%**	0.929	0.951	0.940
+stop		0.666	0.574	0.616
−vocalic	**89.22%**	0.907	0.907	0.906
+vocalic		0.873	0.874	0.873

Table 2: Accuracy Rates for Bagging with REP trees on TIMIT core test set for recognition of articulatory manner features

Articulatory feature	Accuracy	Precision	Recall	F1-measure
−approximant	**94.15%**	0.952	0.987	0.969
+approximant		0.569	0.251	0.348
−closure	**94.90%**	0.961	0.984	0.973
+closure		0.753	0.556	0.640
−flap	**99.76%**	0.998	1.000	0.999
+flap		0.000	0.000	0.000
−fricative	**93.84%**	0.952	0.974	0.963
+fricative		0.870	0.778	0.821
−lateral approx.	**96.77%**	0.968	1.000	0.984
+lateral approx.		0.000	0.000	0.000
−nasal	**96.30%**	0.973	0.988	0.981
+nasal		0.720	0.533	0.613
−stop	**89.86%**	0.903	0.990	0.945
+stop		0.795	0.276	0.410
−vocalic	**89.13%**	0.934	0.886	0.910
+vocalic		0.831	0.899	0.864

Table 3: Accuracy Rates for SVMs with linear kernel on TIMIT core test set for recognition of articulatory manner features

For reasons of further comparison we present in Table 2 the performance on the task of articulatory feature recognition for the method of bagging (Breiman, 1996) with reduced error pruned (REP) decision trees (Quinlan, 1987) that was reported to perform best among several machine learning techniques on the same data (Macek et al., 2005).

Articulatory feature	Accuracy	Precision	Recall	F1-measure
$-$approximant	**94.85%**	0.967	0.979	0.973
+approximant		0.606	0.497	0.546
$-$closure	**96.13%**	0.976	0.982	0.979
+closure		0.782	0.729	0.754
$-$flap	**99.76%**	0.998	1.000	0.999
+flap		0.000	0.000	0.000
$-$fricative	**95.10%**	0.958	0.984	0.970
+fricative		0.916	0.804	0.856
$-$lateral approx.	**97.44%**	0.977	0.997	0.987
+lateral approx.		0.772	0.294	0.426
$-$nasal	**97.94%**	0.985	0.993	0.989
+nasal		0.862	0.745	0.799
$-$stop	**92.35%**	0.949	0.965	0.957
+stop		0.726	0.642	0.682
$-$vocalic	**91.52%**	0.936	0.926	0.931
+vocalic		0.883	0.898	0.890

Table 4: Accuracy Rates for SVMs with polynomial kernel of order $d = 2$ on TIMIT core test set for recognition of articulatory manner features

Articulatory feature	Accuracy	Precision	Recall	F1-measure
$-$approximant	**94.96%**	0.969	0.977	0.973
+approximant		0.608	0.537	0.570
$-$closure	**96.30%**	0.977	0.982	0.980
+closure		0.789	0.745	0.766
$-$flap	**99.77%**	0.998	1.000	0.999
+flap		1.000	0.030	0.058
$-$fricative	**95.30%**	0.959	0.985	0.972
+fricative		0.922	0.810	0.862
$-$lateral approx.	**97.49%**	0.979	0.996	0.987
+lateral approx.		0.728	0.356	0.478
$-$nasal	**97.97%**	0.986	0.993	0.989
+nasal		0.856	0.757	0.803
$-$stop	**93.08%**	0.954	0.967	0.961
+stop		0.753	0.682	0.715
$-$vocalic	**91.73%**	0.935	0.931	0.933
+vocalic		0.890	0.895	0.893

Table 5: Accuracy Rates for SVMs with polynomial kernel of order $d = 3$ on TIMIT core test set for recognition of articulatory manner features

3. Support Vector Machines

Support Vector Machines learn separating hyperplanes to classify instances in the feature space that are mapped from the input space of the classified data. The mapping from input space to feature space is performed with application of a kernel on the feature space. The dimension of the feature space is typically much higher than that of the original input space. The term 'feature' in this context is of course distinct from articulatory feature.

For a binary classification task with data points $\mathbf{x}_i$ ($i = 1, \ldots, n$) and labels y_i we have the decision function $f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$. If the dataset is separable we can find a $\mathbf{w}$ such that the decision function will assign value $f(\mathbf{x}_i) = y_i$ for every i. As the sign is invariant to positive scaling of the expression inside of the sign, we can define canonical hyperplanes such that $\mathbf{w} \cdot \mathbf{x} + b = 1$ for the closest points on one side and $\mathbf{w} \cdot \mathbf{x} + b = -1$ for the closest points on the other side. The separating hyperplane is then defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$ and its normal is then $\mathbf{w}/\|\mathbf{w}\|_2$. The margin between the canonical hyperplanes can be found as a projection of distance between the two closest points on opposite sides ($\mathbf{x}_1$ and $\mathbf{x}_2$) on the normal of separating hyperplane. Since $\mathbf{w} \cdot \mathbf{x}_1 + b = 1$ and $\mathbf{w} \cdot \mathbf{x}_2 + b = -1$ the margin is $1/\|\mathbf{w}\|_2$.

The SVM approach to binary decision function learning is to maximize the margin $1/\|\mathbf{w}\|_2$ that is summarized in an optimization task formulation

$$\min g(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \textit{w.r.t.} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \text{ for all } i$$

and the learning task can be reduced to minimization of the primal lagrangian

$$L = \frac{1}{2}(\mathbf{w}^T \cdot \mathbf{w}) - \alpha_i(y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1),$$

where α_i are Lagrangian multipliers.

3.1. Kernels

From the description of support vector machines it is apparent that for a nonlinear problem it is not suitable to use a linear classifier. To make use of the beneficial properties of a linear SVM we need to map nonlinearly separable data into a space of typically higher dimensionality where linear separation of the data is possible. Thus we define a map from the input space $\mathbf{X}$ into feature space $\mathbf{H}$, $\Phi : \mathbf{X} \to \mathbf{H}$.

Although there is an infinite number of such mappings only some are suitable for practical application for computational complexity reasons. The kernel trick (Schölkopf and Smola, 2002) relieves from often exponential explosion of computations by introducing kernel k that is equivalent to the map Φ in that it holds $\langle \Phi(x), \Phi(x') \rangle = k(x, x')$, where $\langle ., . \rangle$ is dot product. This property holds for polynomial kernels that map input vector into the feature vector composed of ordered polynomial expansions, eg. for order $d = 2$ of the polynomial and 2-dimensional input space we have $\Phi : (x_1, x_2) \to (x_1^2, x_2^2, 2x_1x_2)$.

4. Experiments with SVMs for Articulatory Feature Recognition

We extracted 52 values for every frame of speech signal that were used as inputs for the SVM classifiers. From each frame we extracted 12 Mel-frequency Cepstral Coefficients together with first and second order differences, frequencies of formants (F1-F5) with first order differences, bandwidths of detected formants, and fundamental frequency. The length of the speech signal frames was set to 25 ms and step between two adjacent frames to 10 ms. The original speech signal was sampled at 16 kHz. The distributions of classes vary significantly for different types of features. While the distribution of classes is almost equal (the case of AF *vocalic*) for half of the articulatory features, in the rest of the cases the positive classes are rare in the data. This has a strong influence on the recall of the positive classes while the overall accuracy remains high.

In Tables 3, 4 and 5 we present results for SVMs with linear kernel and with polynomial kernel of second and third order, respectively, for the recognition of manner features based on FER on TIMIT core test set. The values of recall, precision, and F1-measure are presented for positive and negative classes of a articualtory feature.

The comparison of individual kernels in the SVM classification leaves us with observation that the performance improves for all articulatory features with increasing order of used kernels. From comparison of the performances with bagging we see that all SVMs perform better in terms of F1-measure for all features except the feature *fricative*.

Interestingly, drop in the ratio of cases with positive class in the data need not necessarily lead to drop in performance if it is accompanied by increase of 'compactness' of the class. This can be seen from the better performance for the feature *closure* which is on the frame level a subset of the feature *stop*.

5. Conclusion

We presented support vector machines with three types of kernels as approaches to recognition of articulatory manner features that we use as a building block of a continuous speech recognizer. The comparison was made between a linear and two polynomial kernels of second and third order for isolated frame recognition approach. Our results show high dependence of the performance on positive/negative class balance in the data whereby with increasing unbalance of the class distributions the performance of recognizers degrades.

According to the frame based values of F1-measure the SVM with polynomial kernel of third order gave superior performance over SVMs with the remaining two types of kernel for all articulatory manner features. These superiority of the third order polynomial kernel is underlined by monotone increase of the F1-measure for all classes. The comparison of SVM with third order polynomial kernel with bagging gives very similar results except for the articulatory feature *fricative* where the performance is better for bagging.

Performance of the SVMs was dependent on the frequency of occurrence of classes in the data. It achieved better performance in terms of recall, accuracy and F1-measure in cases where the distribution of positive and negative classes was not too unbalanced. An especially interesting case of this influence is the feature *flap* for which the positive class is contained in less the one percent of the speech frames. Although this feature was virtually undetected with our methods, from the point of view of speech recognition its practical importance is obviously smaller than that of more frequent articulatory features.

6. Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. 02/IN1/I100. We would also wish to express our gratitude to Dr. Anja Geumann for support of our work.

7. References

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

Christopher J. C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Julie Carson-Berndsen. 1998. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Kluwer, Dodrecht, Netherlands.

Shuangyu Chang, Mirjam Wester, and Steven Greenberg. 2005. An Elitist Approach to Automatic Articulatory-acoustic Feature Classification for Phonetic Characterization of Spoken Language. *Speech Communication*, 47(3):290–311, November.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST, USA.

Hynek Heřmanský. 1999. Mel cepstrum, deltas, double-deltas,.. - What else is new? In *Proc. of Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland.

Roman Jakobson, Gunnar Fant, and Morris Halle. 1952. *Preliminaries to Speech Analysis: The distinctive features and their correlates*. MIT Press.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods—Support Vector Learning*, pp. 169–184. MIT Press.

Supphanat Kanokphara, Jan Macek, and Julie Carson-Berndsen. 2006. Comparative Study: HMM & SVM for Automatic Articulatory Feature Extraction. In *Proc. of the 19th Int'l. Conference IEA/AIE*, Annecy, France.

Katrin Kirchhoff. 1999. *Robust Speech Recognition using Articulatory Information*. Ph.D. thesis, Bielefeld.

Benoît Launay, Olivier Siohan, Arun Surendran, and Chin-Hui Lee. 2002. Towards Knowledge-Based Features for HMM Based Large Vocabulary Automatic Speech Recognition. In *Proc. of IEEE ICASSP*, vol. 1, pp. 817–820, Orlando, FL, USA, May.

Jan Macek, Supphanat Kanokphara, and Anja Geumann. 2005. Articulatory-acoustic Feature Recognition: Comparison of Machine Learning and HMM methods. In *Proceedings of SPECOM 2005*, vol. 1, pp. 99–103.

John R. Quinlan. 1987. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234.

Bernhard Schölkopf and Alexander J. Smola. 2002. *Learning with Kernels*. MIT Press, Cambridge, MA, USA.

Kenneth N. Stevens. 1980. Acoustic correlates of some phonetic categories. *JASA*, 68(3):836–842.

Kenneth N. Stevens. 2000. *Acoustic Phonetics*. MIT Press, Cambridge, MA, USA.

Sebastian Stüker, Tanja Schulz, Florian Metze, and Alex Waibel. 2003. Multilingual Articulatory Features. In *Proc. of IEEE ICASSP*, vol. 1, pp. 144–147. IEEE.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer Verlag, New York.