# Comparative Study: HMM&SVM for Automatic Articulatory Feature Extraction

Supphanat Kanokphara, Jan Macek and Julie Carson-Berndsen

University College Dublin
School of Computer Science and Informatics
Belfield, Dublin 4, IRELAND
{supphanat.kanokphara, jan.macek, julie.berndsen}@ucd.ie

**Abstract.** Generally speech recognition systems make use of acoustic features as a representation of speech for further processing. These acoustic features are usually based on human auditory perception or signal processing. More recently, Articulatory Feature (AF) based speech representations have been investigated by a number of speech technology researchers. Articulatory features are motivated by linguistic knowledge and hence may better represent speech characteristics. In this paper, we introduce two popular classification models, Hidden Markov Model (HMM) and Support Vector Machine (SVM), for automatic articulatory feature extraction. HMM-based systems are found to be best when there is good balance in the numbers of positive and negative examples in the data while SVM is better in the unbalanced data condition.

## 1 Introduction

HMMs are predominantly used as acoustic models in current speech recognition systems. The reason for this is that HMMs can normalize the time-variation of the speech signal and characterize the speech signal statistically and optimally. HMM-based speech recognition systems do not deal with the speech signal directly but rather deal with a specific kind of speech representation. Such speech representations have two roles in speech recognition systems. First, they reduce data for further speech recognition processing. Second, they ideally present only useful information from speech signal to the systems and ignore non-acoustic and/or non-linguistic information such as speaking style, noise, emotion, etc.

From the reasons mentioned above, it is needless to say that the choice of speech representation can also greatly affect the accuracy of systems. Speech representations today are usually based on the acoustic information of the signal [1]. 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 ways to solve this problem is to integrate linguistic knowledge as suggested by [2, 3].

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 cross-lingual recognizer [4]. The AF representations have also proven to be good in noisy environment [5].

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 [6, 7]. The first detailed description of distinctive features [8] assumed that they had identifiable counterparts.

In this paper, two systems are introduced to extract articulatory features from the speech signal. The first system is based on an HMM. As mentioned previously, an HMM is designed to map a speech signal into a sequence of units. These units can be words, syllables, demi-syllables, phones, etc. This approach is therefore suitable for articulatory feature extraction.

As a second approach to articulatory acoustic manner feature recognition we used Support Vector Machine (SVM) classifiers with a linear kernel in the SVMLight implementation [8]. In contrast to the HMM approach, we run the SVMs in a static description setting, i.e. only parameters describing the recognized frame and four adjacent frames are used as a source of information for prediction. Although such a setting makes no use of contextual information we show that the achieved performance is better than HMMs in the highly skewed data condition.

Systematically, this paper is organized as follows. Section 2 explains the experimental details of the experimental paradigm used in this paper, i.e. the corpus, the evaluation method and the feature table. Section 3 describes the HMM-based AF extraction system while section 4 presents the SVM-based AF extraction system. Finally, discussion and conclusion work are presented in section 5.

## 2 Experimental Setup

### 2.1 The Corpus

The experiments use the standard TIMIT corpus [10] 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 without SA sentences is used as the training set while only the core test set is used as the test set.

### 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 and precision. FER is widely used for articulatory feature extraction evaluation [11]. This is because, in current speech technology, articulatory features are commonly used as an alternative or additional speech representation. The speech representation is a sequence of numeric vectors where each numeric vector represents speech in each time frame. Therefore, AF extraction systems are usually evaluated on the frame level.

To understand the differences between achieved accuracies better under highly skewed data condition we present *precision* and *recall* rates in Table 3. The precision is defined as the ratio

$$\frac{\#\text{ of correctly classified instances of class } c}{\#\text{ of instances classified as class } c} \tag{1}$$

and the recall is defined as the ratio

$$\frac{\#\text{ of correctly classified instances of class } c}{\#\text{ of instances of class } c} \tag{2}$$

Both measures analyse the correct classification for each feature class individually. It is important to mention the trade-off between precision and recall.

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 AF exists. In this paper, we directly convert reference phone transcriptions into AF transcriptions. These transcriptions lack the co-articulation properties which would be found in annotated. However, this is the only resource we have and it is widely accepted as the reference transcriptions for AF evaluation.

For our experiments we have been using the TIMIT corpus that is annotated at the phone level, thus for our task it was necessary to map from original phone annotations

**Table 1.** Assignment of articulatory-acustic manner feature classes to phonemes and their frequency in the TIMIT corpus

| 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 approximant | **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 |

to articulatory acoustic features. The assignment of articulatory-acoustic manner feature classes to phonemes is shown in Table 1.

## 3 The HMM-Based AF Extraction System

The HMM-based AF extraction system is actually a normal HMM-based phone recognizer. First the phone recognizer hypothesizes a phone sequence from the input speech. The sequence is then mapped to AF sequences according to each individual tier from the feature table. The advantage of this system is that the information on different tiers can be used to help hypothesize the correct articulatory features which results in higher accuracy.

### 3.1 System overview

The HMM-based articulatory feature extraction systems in this paper are constructed using HTK [12] which is now widely used for HMM based speech recognition experiments. The acoustic model training system starts by converting the speech signal into a sequence of vector parameters with a fixed 25 ms frame and a frame rate of 10 ms. Each parameter is then pre-emphasized with the filter P(z) = 1-0.9*z^(-1). The discontinuities at the frame edges are attenuated by using Hamming window. A fast Fourier transform is used to convert time domain frames into frequency domain spectra. These spectra are averaged into 24 triangular bins arranged at equal mel-frequency intervals (where fmel = 2595 log10 (1+f /700)), where f denotes frequency in Hz. 12 dimensional mel-frequency cepstral coefficients (MFCCs) are then obtained from cosine transformation and lifter. The normalized log energy is also added as the 13th front-end parameter. The actual acoustic energy in each frame is calculated and the maximum selected. All log energies are then normalized with respect to maximum and log energies below a silence floor (set to -50 dB) clamped to that floor.

These 13 front-end parameters are expanded to 39 front-end parameters by appending first and second order differences of the static coefficients. The chosen parameters have been used extensively and have proven to be a good choice for HMM-based speech recognition systems [13].

Each model contains 5 states with no skip state and the covariance matrices of all states are diagonal. After the context-independent HMMs have been trained, they are expanded to context-dependent HMMs using tree-based state tying technique [12] for cross-phone network expansion. Maximum likelihood estimators are used to train HMM parameters. The number of training iterations after each change is determined automatically [14]. There is no mixture expansion. Only one mixture models are used in the experiment. For the recognition process, the Viterbi algorithm is used without any pruning factor and language model.

**Table 2.** HMM-based AF extractor results

| Manner Feature | Overall accuracy | Precision (for positive class) | Recall (for positive class) |
|---|---|---|---|
| Approximant | 94.07% | 66.00% | 70.50% |
| Closure | 95.27% | 82.93% | 84.01% |
| Flap | 97.90% | 36.22% | 62.46% |
| Fricative | 93.26% | 85.75% | 84.89% |
| Nasal | 95.90% | 77.31% | 84.17% |
| Lateral Approximant | 96.18% | 49.69% | 69.32% |
| Stop | 92.10% | 76.10% | 87.54% |
| Vocalic | 89.20% | 90.36% | 86.47% |

### 3.2 Experimental Result

According to the experimental results, HMM-based system works very well with AF extraction since all accuracies are averagely 94.24%. Considering precision and recall, HMM-based Flap and Lateral Approximant extractors do not seem to work very well. The precision for Flap is only 36.22% while it is only 49.69% for Lateral Approximant. The recall for Flap is only 62.46% while it is only 69.32% for Lateral Approximant

## 4 Articulatory manner features recognition with SVMs

Support Vector Machines learn separating hyperplanes to classify instances in the feature space that is 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 acoustic feature.

The motivation for using SVMs comes from the pattern recognition community with mathematical properties of linear classifiers and from the statistical learning theory community with the structural risk minimization properties of SVMs [15, 16].

For a binary classification task with data points $\mathbf{x}_i$ ($i$=1,…, n) and labels $y_i$ we have the decision function $f(\mathbf{x})=sign(\mathbf{w}\cdot\mathbf{x} + b)$. 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

$$\text{minimize } g(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{w}\|_2^2 \text{ subject to constraints } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \text{ for all } i$$

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

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

where $\alpha_i$ are Lagrangian multipliers.

Confidence measures for SVMs can be based on the value of $\mathbf{w}\cdot\mathbf{x} + b$. In our approach we used normalization of the SVM output into [0; 1] via normalization function $c(\mathbf{x}) = 1 - \exp(-\sigma(\mathbf{w}\cdot\mathbf{x} + b)^2)$.

### 4.1 Experiments with SVMs for articulatory feature recognition

We extracted 180 inputs for the SVM classifier. These were extracted for the current frame and two adjacent frames before and after the current frame, i.e. 5 frames in total. From each frame we extracted values of overall energy, overall entropy, 12 Mel-frequency Cepstral Coefficients with first and second order differences. The length of the frames was set to be 16 ms due to practical consideration of the speed of FFT on signals of length $2^n$. In our case we worked with 16 kHz data.

The labels for each of the instances were converted from phoneme transcriptions of the speech signal. 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 although the overall accuracy remains high.

The SVMs were trained on all training data provided in TIMIT corpus and the evaluation of training was performed on the core test set of TIMIT corpus. Frame Error Rate (FER) was used to evaluate the performance and make it comparable with HMM approach mentioned above. Table 3 shows results for recognition of manner features based on FER on TIMIT core test set. The values of recall and precision are values for the positive class, i.e. for subset of instances with class of the respective manner feature. The N/A values in the Table 3 are caused by 0% recall of the positive class actually meaning that all instances were classified as negative class and leaving the precision value undefined.

**Table 3.** Frame Error Rates on TIMIT core test set for SVMs for recognition of articulatory acoustic manner features

| Manner Feature | Overall accuracy | Precision (for positive class) | Recall (for positive class) |
|---|---|---|---|
| Approximant | 91.88% | N/A | 0% |
| Closure | 94.20% | 82.54% | 56.36% |
| Flap | 99.29% | N/A | 0% |
| Fricative | 88.51% | 93.78% | 36.10% |
| Nasal | 96.26% | N/A | 0% |
| Lateral Approximant | 94.59% | N/A | 0% |
| Stop | 90.16% | 81.41% | 53.39% |
| Vocalic | 82.23% | 74.83% | 80.17% |

In our experiments we tried to modify the training set to achieve better performance on the core test set to cope with significant differences in class distributions in data as is shown in Table 1. While only three of the manner feature classes occur in more than 15% of cases, five of the used feature classes imply highly unbalanced datasets. Although in the cases of *closure*, *fricative*, *stop* and *vocalic* the performance was influenced in a negative way, we did not obtain any change in the classifiers performance for the features *approximant*, *lateral approximant*, and *nasal*. The recall was positively influenced only in the case of the manner feature *flap*, but this change was an increased recall to the value of 40.52% at precision rate of 7.75% and overall accuracy of 95.74%. These values were achieved for undersampling of the negative examples in the training set to 5% of the original count.

Although approaches of over- and undersampling of under-/overrepresented classes in unbalanced data sets are standard and were reported to give good results [17, 18]. In our case we suspect that attributes used for description of the signal frames do not contain enough information to allow the SVM to separate them even after altering the original training dataset.

## 5 Conclusion

We presented two approaches to recognition of articulatory manner features that we use as a building block of a continuous speech recognizer. The comparison was made between the sequential method of Hidden Markov Models and the isolated frame recognition approach based on Support Vector Machines.

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 FER, the HMM outperformed SVMs in most of the cases except for the manner feature *flap* and *nasal* in terms of overall accuracy. This comparison is slightly unfair as the recall rates for both of these cases is 0%.

Performance of the SVMs was highly dependent on the frequency of occurrence of positive classes in the data. It has not achieved good performance in terms of recall and accuracy in cases where the distribution of positive and negative classes was very unbalanced. Although we performed resampling of the original training data we did not observe useful positive change in the performance. We suspect this to be caused by poor discriminative power of the used parameters of the speech signal.

## References


1. H. Hermansky: Mel Cepstrum, Deltas, Double-Deltas,-What Else is New? Proc. in Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland (1999)
2. B. Launay, O. Siohan, A. C. Surendran and C.H. Lee: Towards knowledge-Based Features for HMM Based Large Vocabulary Automatic Speech Recognition. In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando (2002)
3. J. Carson-Berndsen: Time Map Phonology: Finite State Models and Event Logics in Speech Recognition. Dordrecht, Holland: Kluwer Academic Publishers (1998)
4. S. Stueker, T. Schulz, F. Metze, and A. Waibel: Multilingual Articulatory Features. In Proceedings of ICASSP, Vol. 1, (2003) 144–147
5. K. Kirchhoff: Robust Speech Recognition using Articulatory Information. Ph.D. thesis, University of Bielefeld (1999)
6. K.N. Stevens: Acoustic Correlates of some Phonetic Categories. Journal of the Acoustical Society of America (JASA), Vol. 68(3), (1980) 836-842
7. K.N. Stevens: Acoustic Phonetics. MIT Press: Cambridge (Ma), London (1998)
8. R. Jakobson, G. Fant, M. Halle. Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates. MIT Press, 9th ed. 1969 (1952)
9. T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
10. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM, NIST (1993)
11. S. Chang, S. Greenberg, and M. Wester: An Elitist Approach to Articulatory-Acoustic Feature Classification. In Proc. 7th Eurospeech, Aalborg, Denmark (2001), 1725-1728
12. S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland: The HTK Book, Microsoft Corporation and Cambridge University Engineering Department (December 2002)
13. S. B. Davis and P. Mermelstein: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Trans. Acoustic Speech and Signal Processing, Vol. 28(4), (1980), 357-366
14. P. Tarsaku and S. Kanokphara: A Study of HMM-Based Automatic Segmentations for Thai Continuous Speech Recognition System. In Proc. the Symposium on Natural Language Processing, (2002), 217-220
15. V.Vapnik (1995) *The Nature of Statistical Learning Theory*, Springer.
16. C.J.C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discovery, Vol. 2, Number 2, p. 121-167, 1998.
17. Maloof, M.A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*.
18. N. Lachiche and P. A. Flach. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In: *Proc. 20th International Conference on Machine Learning (ICML'03)*, pages 416-423. AAAI Press, January 2003.