Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is an author's version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/41593

Please be advised that this information was generated on 2021-10-19 and may be subject to change.

CAPTURING FINE-PHONETIC VARIATION IN SPEECH THROUGH AUTOMATIC CLASSIFICATION OF ARTICULATORY FEATURES

Odette Scharenborg, Vincent Wan, Roger K. Moore

Speech and Hearing Research Group
Department of Computer Science, University of Sheffield, Sheffield, UK
{O.Scharenborg, V.Wan, R.K.Moore}@dcs.shef.ac.uk

ABSTRACT

The ultimate goal of our research is to develop a computational model of human speech recognition that is able to capture the effects of fine-grained acoustic variation on speech recognition behaviour. As part of this work we are investigating automatic feature classifiers that are able to create reliable and accurate transcriptions of the *articulatory* behaviour encoded in the acoustic speech signal. In the experiments reported here, we compared support vector machines (SVMs) with multilayer perceptrons (MLPs). MLPs have been widely (and rather successfully) used for the task of multi-value articulatory feature classification, while (to the best of our knowledge) SVMs have not. This paper compares the performances of the two classifiers and analyses the results in order to better understand the articulatory representations. It was found that the MLPs outperformed the SVMs, but it is concluded that both classifiers exhibit similar behaviour in terms of patterns of errors.

1. INTRODUCTION

In everyday speech it is quite common for there to be no pauses between lexical items; words flow smoothly one in to another with adjacent sounds often coarticulated. This means that, if words are assumed to be constructed from a limited set of abstract 'phonemes', virtually every phoneme string is compatible with many alternative word sequence interpretations. Human listeners, however, appear to be able to recognise intended word sequences almost effortlessly. Even in the case of *fully* embedded words such as "ham" in "hamster", listeners can make the distinction between the two interpretations even *before* the end of the first syllable "ham". There is now considerable evidence that sub-segmental (i.e. acoustic-phonetic) and supra-segmental (i.e. prosodic) cues in the speech signal modulate human speech recognition (HSR), and help the listener segment a speech signal into syllables and words (e.g. [5],[10],[19]). However, currently no computational models of HSR exist that are able to model this '*fine-phonetic variation*' [8].

Our ultimate goal is to refine an existing computational model of HSR (SpeM [20]) such that it is able to capture and use fine-grained acoustic-phonetic variation during speech recognition. In this study, we investigate the use of '*articulatory features*' (AFs) that describe properties of speech *production* and that can be used to represent the acoustic signal in a compact manner. Furthermore, AFs offer the possibility of representing coarticulation and assimilation effects as feature value changes.

In automatic speech recognition (ASR), there has also been a growing interest in the use of articulatory features for improving the performance of current ASR systems. AFs are often put forward as the solution (e.g. [12],[23],[24]) to the problem of modelling the variation in speech using the standard 'beads-on-a-string' paradigm [17], in which the acoustic signal is described in terms of phones, and words as phone sequences.

Over the years, many different approaches have been investigated for incorporating AFs into ASR systems. Artificial neural networks (ANNs), for instance, have shown high accuracies for classifying AFs ([11],[12],[23]). Frankel *et al* [7] provide a short overview of other modelling schemes, such as hidden Markov models (HMMs) [12], linear dynamic models (LDMs) [6] and dynamic Bayesian networks (DBNs) [15]. However, for smaller (and simpler) tasks, support vector machines (SVMs) offer two favourable properties: (i) good generalisation and (ii) the ability to learn from a relatively small amount of high-dimensional data. For these reasons, SVMs have also been applied to the classification of articulatory features [9],[16]. For instance, Juneja [9] developed SVM-based landmark detectors for classifying binary place and voicing features in TIMIT [14] where he reported accuracies ranging from 79% to 95%. Also, Niyogi and Sondhi [16] used SVMs to detect stop consonants in TIMIT.

However, the research reported so far using SVMs to classify articulatory features have been mainly concerned with *binary* decision tasks, or with a limited domain. In the area of *visual* automatic speech recognition, however, SVMs have been used successfully for the automatic classification of *multi*-level articulatory features [18]. This leads us to suspect that SVMs could also be used for the classification of *multi*-level *acoustic* articulatory features.

In the work reported here, we investigate the possibility of classifying multi-level acoustic articulatory features using SVMs in the context of the larger goal of developing a computational model of HSR that is able to model the effect of fine-grained acoustic variation on HSR. For this computational system, we are in search of feature classifiers that are able to create reliable and accurate feature transcriptions of the acoustic signal. Given the existing high performance of ANNs on the task of AF classification, this paper reports on an in-depth comparison between the performances of the ANNs and the SVMs and analyse the results to better understand the articulatory features.

In order to allow a direct comparison between the ANN and the SVM systems, both systems have been trained on the same material (see Section 2.1) using the same AF set (see Section 2.2). Section 2 outlines details of the two classification systems that were used, Section 3 presents and analyses the results for

both systems, then Section 4 discusses the results. Conclusions are presented in the final section.

2. EXPERIMENTAL SET-UP

2.1. Material

The training and testing material used in this study are taken from the TIMIT corpus [14]. TIMIT consists of reliably hand labelled and segmented data of quasi-phonetically balanced sentences read by native speakers of eight major dialect regions of American English. Of the 630 speakers in the corpus, 438 (70%) were male. We followed TIMIT's standard training and testing division, in which no sentence or speaker appeared in both the training and test set. The training set consisted of 3,696 utterances. The test set (excluding the `sa` sentences) consisted of 1,344 utterances.

2.2 Articulatory features

In this research, we used the set of seven articulatory features shown in Table 1. The names of the AFs are self-explanatory, except maybe for *static*, which gives an indication of the rate of acoustic change, e.g., during diphthongs.

The set is based on the six AFs proposed in [23]. After initial experiments, we added a seventh AF to replace the corresponding values in *place*: *high-low*. This improved the accuracies for the AF values in both AF classes compared to when the *high-low* AF values were in *place*. For the training and testing data, the frame-level phonemic TIMIT labels were replaced by the canonical AF values using a table look-up procedure. The mappings between the phonemes and the AF values are based on [13].

2.3. Multilayer perceptron AF classification

For the first experiment, seven multilayer perceptrons or MLPs (one for each AF) were trained using the NICO Toolkit [21]. All MLPs consisted of three layers. Each MLP's input layer, with 273 nodes, was presented with 12 MFCC coefficients plus log energy (for 25 ms frames, with a 10 ms frame shift), their first and second derivatives and a context window of plus and minus 3 frames. The hidden layers had hyperbolic tan transfer functions and a different number of nodes depending upon the AF. To determine the optimum network size, networks with various numbers of hidden units were trained in an initial experiment. The network configurations that gave the best performance in the initial tests are used in the experiments and results presented below. The output layer was configured to estimate the posterior probability of the AF value given the input. The number of output nodes for each MLP is listed in Table 3.

When training each MLP the performance on a validation set (consisting of 100 utterances randomly selected and taken from the training material) was monitored and training was terminated when the validation set's error rate began to increase. During classification, the class with the highest associated posterior probability is chosen.

2.4. Support Vector Machine AF classification

SVMs (for an introductory text, the reader is referred to [1]) are binary classifiers which make their decisions by constructing a hyper-plane that separates the two classes such that the boundary

Table 1. Specification of the AFs and their respective quantised values.

AF	Values
manner	approximant, retroflex, fricative, nasal, stop, vowel, silence
place	bilabial, labiodental, dental, alveolar, velar, nil, silence
voice	voiced, unvoiced
high-low	high, mid, low, nil, silence
fr-back	front, central, back, nil
round	rounded, unrounded, nil
static	static, dynamic

is geometrically furthest away from both classes. Unlike MLPs, SVMs are not statistical classifiers and do not estimate posterior probabilities. The decision hyper-plane is defined by $\boldsymbol{x}\cdot\boldsymbol{w} + b = 0$ where $\boldsymbol{w}$ is the normal to the hyper-plane and $\boldsymbol{x}$ is a d-dimensional vector. For linearly separable data labelled by y, $\{\boldsymbol{x}_n, y_n\}$ for $\boldsymbol{x}_n \in \Re^d$, $y_n \in \{-1,+1\}$, $n = 1\ldots N$, the optimal decision hyper-plane is found by maximising the *margin* between the two classes, which in turn is achieved by minimising $\|\boldsymbol{w}\|^2$ subject to the inequalities $(\boldsymbol{x}_n\cdot\boldsymbol{w} + \mathrm{b})\, y_n \geq 1$ for all n. The solution for the optimal hyper-plane $\boldsymbol{w}_0$ is a linear combination of a small subset of the training data $\boldsymbol{x}_s$, $s \subset \{1\ldots N\}$, which are known as the support vectors. These support vectors also satisfy the equality $(\boldsymbol{x}_s\cdot\boldsymbol{w}_0 + b)\, y_s = 1$.

When the data is non-separable then no hyper-plane exists for which all points satisfy the inequality above. To overcome this problem slack variables ξ_n are introduced into the inequalities relaxing them so that some points are allowed to lie within the margin or be misclassified completely. The resulting problem is then to minimise

$$\|\boldsymbol{w}\|^2 + c\Sigma_n\xi_n \quad \text{subject to } (\boldsymbol{x}_n\cdot\boldsymbol{w} + \mathrm{b})\, y_n \geq 1 - \xi_n \quad \forall\, n \qquad (1)$$

The solution for $\boldsymbol{w}_0$ is then a linear combination of all points that have $\xi > 0$ as well as those that lie on the margin.

SVMs are easily extended to construct non-linear boundaries. This is achieved by mapping the data non-linearly onto a manifold embedded in a higher dimensional space and constructing the decision hyper-plane there. A practical way to demonstrate this is to fold a flat sheet of paper (a 2D space) into a 3D shape, cut it linearly and unfold to reveal the non-linear cuts. Such transformations are implemented easily by the use of kernel functions. The optimisation procedure for $\boldsymbol{w}_0$ is expressed entirely in terms of the inner product between pairs of vectors. Kernel functions compute the corresponding inner product in the higher dimensional space as a function of the original vectors without explicitly applying any mapping. The most basic and common kernel functions are the polynomial kernel,

$$K_{\mathrm{poly}}(\boldsymbol{x}_1\cdot\boldsymbol{x}_2) = (A\,\boldsymbol{x}_1\cdot\boldsymbol{x}_2 + B)^p \qquad (2)$$

and the radial basis function (RBF) kernel,

$$K_{\mathrm{RBF}}(\boldsymbol{x}_1\cdot\boldsymbol{x}_2) = \exp -0.5\cdot\gamma\,(\boldsymbol{x}_1 - \boldsymbol{x}_2)^2 \qquad (3)$$

In our experiments, we used LIBSVM, which achieves multi-class classification by error correcting codes [2]. In an initial experiment, we tested both the polynomial and the RBF kernel on the same task. Since the RBF kernel showed a better result

Table 2. SVM AF classification accuracies (Acc; decreasing from left to right) for each AF and the percentage of support vectors (SV) used for each SVM.

#utts	voice		round		fr-back		manner		static		high-low		place	
	SV(%)	*Acc(%)*	*SV(%)*	*Acc(%)*	*SV(%)*	*Acc(%)*	*SV(%)*	*Acc(%)*	*SV(%)*	*Acc(%)*	*SV(%)*	*Acc(%)*	*SV(%)*	*Acc(%)*
2K	30.2	89.54	61.6	83.16	41.8	80.25	60.7	73.78	92.4	73.34	58.5	73.14	76.4	69.68
10K	26.8	90.32	48.7	84.81	36.9	82.26	51.2	77.02	85.4	76.05	53.1	75.93	66.7	73.51
50K	25.1	90.80	40.4	86.08	34.0	83.43	46.8	78.91	76.4	77.98	48.7	77.63	57.5	76.39
100K	24.2	90.95	37.3	86.57	33.3	83.72	44.6	79.64	72.0	78.63	47.8	77.98	53.8	77.55

Table 3. MLP AF classification accuracies (Acc; in decreasing order), the number of hidden nodes, and the number of output nodes used for each MLP.

AF	Acc. (%)	#hidden nodes	#output nodes
voice	92.4	100	2
round	87.2	100	3
manner	84.6	300	7
fr-back	84.3	200	4
static	82.9	100	2
place	81.9	200	7
high-low	80.2	100	5

than the polynomial kernel, we used the RBF kernel for the experiments reported in this paper.

The input of the SVMs consisted of 12 MFCC coefficients plus log energy, and their first and second derivatives, resulting in 39-dimensional acoustic feature vectors for 25 ms frames, with a 10 ms frame shift. Unlike the MLP experiments, no context window was used.

3. RESULTS

3.1. Classification results per AF

Table 3 shows the MLP classification results in terms of percentage frames correctly classified for each AF separately. Furthermore, the sizes of the hidden and output layers of each MLP are listed. The best results are obtained for the *voice* AF, followed by the *round* AF. The results in Table 3 are similar (though slightly worse, with the exception of the performances for *place* and *high-low*) than the results presented in [23].

From Table 3, it is not possible to deduce a clear relationship between the number of output nodes (or the difficulty of the classification task) and the accuracy of the AF classifier. For instance, *static* has two output nodes, like *voice*, but the performance of *static* is almost 10% lower. On the other hand, *manner* has seven output nodes, but gets a relatively high accuracy.

Table 2 shows the classification results of the SVM system for varying amounts of training utterances. The results are reported in terms of the percentage frames correctly classified, for each AF separately. Also, the number of training frames and the percentage of support vectors for each are listed. The percentage of support vectors can give an indication of the relative difficulty of the task and/or separability of the AF values: A larger percentage suggests either more complex decision boundaries or highly overlapping data. The values for γ (see Eq. 3) and c (see Eq. 1) for each SVM are listed in Table 4. A large γ implies narrower RBFs and c sets the amount of regularisation (simpler decision boundaries vs. fitting the training data): If c is large then the SVM constructs more complex decision boundaries to better fit the training data but may result in poor generalisation.

Table 4. Values of the γ and c parameters for each SVM.

AF	γ	c
manner	0.01	15
place	0.1	3
high-low	0.01	100
voice	0.5	5
fr-back	0.01	300
round	1.5	1
static	10	10

Table 5. AF value classification accuracies and differences for the MLP and the SVM systems.

AF value	Accuracy (%)		
	MLP	**SVM**	**Diff**
manner			
approximant	54.9	43.2	11.7
retroflex	70.2	65.1	5.1
fricative	86.7	81.7	5.0
nasal	79.0	73.3	5.7
stop	86.3	70.9	15.4
vowel	91.0	91.9	-0.9
place			
bilabial	68.3	55.1	13.2
labiodental	67.4	57.8	22.8
dental	19.7	21.8	-2.1
alveolar	78.3	75.2	3.1
velar	63.1	50.8	12.3
high-low			
high	70.2	70.4	0.2
mid	55.4	45.3	10.1
low	76.2	71.3	4.9
voice			
+voice	93.8	91.3	2.5
-voice	89.8	90.4	-0.6
fr-back			
front	76.8	82.0	-5.2
central	35.5	12.5	23.0
back	58.3	48.2	10.1
round			
+round	54.0	49.2	4.8
-round	85.0	81.8	3.2
static			
static	84.5	81.0	3.5
dynamic	81.0	75.6	5.4

The results in Table 2 show increasing accuracies (and percentage of support vectors) for increasing number of training utterances. The best performance is (similar to the MLPs) obtained for *voice*, followed by *round*.

Comparing the results of the MLP classifiers in Table 3 with the results of the SVM classifiers in Table 2 shows that the two systems have similar performance; the overall rankings for the best performing classifiers are very much alike, with *manner* and *fr-back* swapping places, just like *place* and *high-low*. Nevertheless, MLPs outperform SVMs for each AF. It must be noted that the SVMs were trained on much less training material than the full set of training frames used for training the MLPs: 100K frames are only 8.8% of the full training set. Furthermore, the SVMs did not use a context window.

3.2. Classification results per AF value

As pointed out above, our ultimate goal is to build a computational model of HSR that is able to recognise fine-grained acoustic-phonetic variation, and to use it during speech recognition. Therefore, we are not only interested in overall classification scores, since these also include the classification of 'nil' or 'silence' (except for *static* and *voice*), but also in the classification of each AF value separately.

Table 5 lists the classification accuracies in terms of frames correctly classified for each AF value for the two classification systems as well as the difference in accuracy. A first quick glance at the results shows that the MLPs also outperform the SVMs on an AF value level, with the exception for the AF values 'vowel', 'dental', '-voice', and 'front'. The differences in accuracies can be as high as 23.0% (for 'central'). The higher AF accuracies for the MLPs – as reported in Tables 2 and 3 – are thus not simply a result of a better classification of 'silence' and 'nil'. For both types of system, the three easiest AF values to classify are '+voice', '-voice', and 'vowel', while the three most difficult are 'dental', 'central', and '+round' for the MLP system, and 'dental', 'central', and 'approximant' for the SVM system. These latter observations are discussed in Section 4.

The great diversity in AF value classification accuracy is remarkable: ranging from 19.7% ('dental') to 93.8% ('+voice') for the MLPs and from 12.5% ('central') to 91.9% ('vowel') for the SVMs. Furthermore, even though the overall classification results for *round* are second highest for both systems, the classifications for '+round' and '-round' differ by more than 30%.

3.3. AF value confusions

Figure 1 shows the graphical representations of the confusion matrices for each of the classifiers in order of decreasing accuracy listed in Table 3. The left hand side of the figure shows the confusion matrices for the MLPs and the right hand side for the SVMs. The vertical axis of each confusion matrix denotes the label of the frame being classified in the reference transcription, while the horizontal axis denotes the label given by the classifier for that frame. The shade of each cell in the matrix refers to the percentage of the reference labels classified as each of the labels on the horizontal axis: white is 100%, black is 0%. Ideally, all cells on the diagonal should be white, and black elsewhere.

The confusion matrices show that *voice*, *manner*, *static*, *place*, and *high-low* overall have few confusions (with the exception of *manner* where a lot of frames have been misclassified as *vowel*). For these confusion matrices, the diagonal has indeed the lightest colour. The confusion matrices for *round* and *fr-back*, however, are a bit 'messier' (see also Section 4).

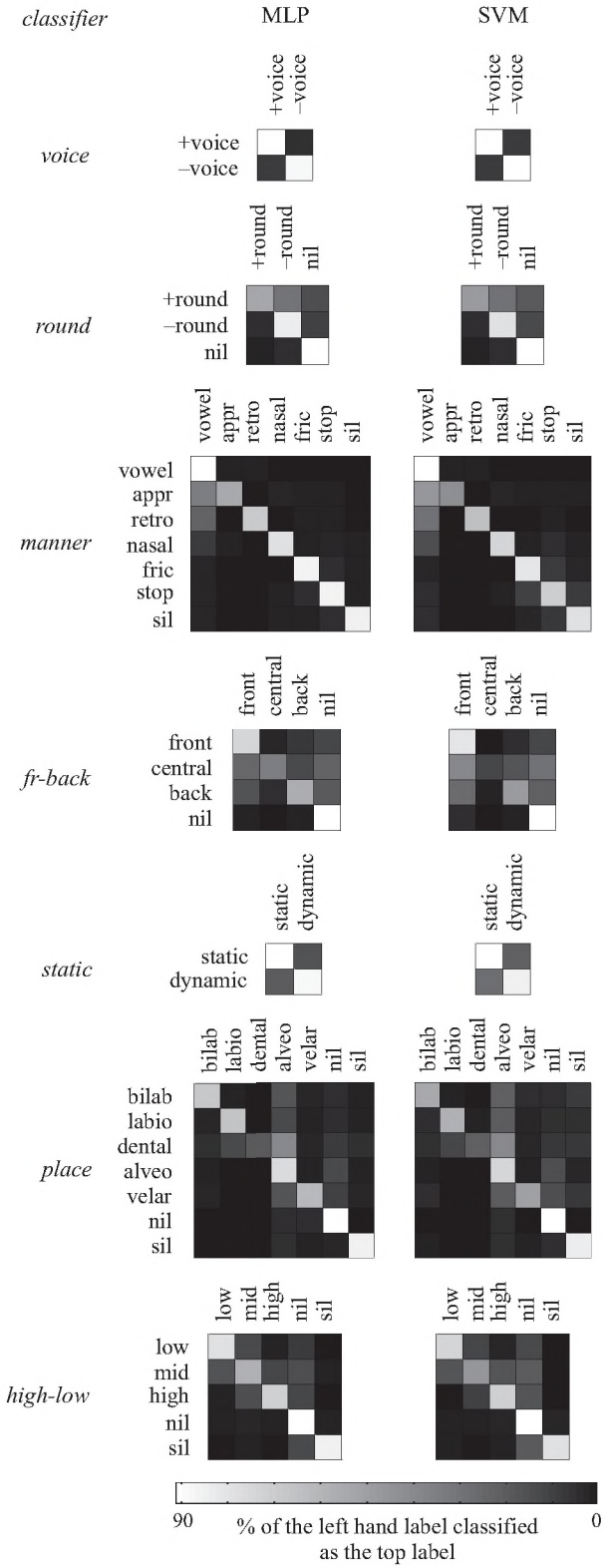

Figure 1. AF value confusion matrices for the MLP (left hand) and SVM (right hand) systems in order of decreasing accuracy (following Table 3).

Table 6. AF value confusions for the MLP and SVM systems, where more than 21% of the frames for a 'from' AF value were labelled as the 'to' AF value.

MLP			SVM		
from	*to*	*%*	*from*	*to*	*%*
dental	alveolar	41.1	approx	vowel	46.6
approx	vowel	35.5	dental	alveolar	41.2
+round	-round	30.9	central	front	40.4
central	front	26.2	+round	-round	30.4
central	nil	24.2	central	nil	29.5
retroflex	vowel	23.5	retroflex	vowel	29.0
			back	front	27.3
			dynamic	static	24.4
			back	central	22.9
			bilabial	alveolar	21.5

What is striking in Figure 1 is that the MLP and SVM confusion matrices are so alike. Both systems tend to make the same ((relative) number of) confusions: The shading in the cells differs only slightly between the two systems.

Table 6 provides more detail on Figure 1. It shows an overview of the AF value confusions for both the MLP and the SVM systems, where more than 21% of the frames for a 'from' AF value were labelled as the 'to' AF value. The most important thing which is immediately evident is that the AF values that scored the lowest accuracies as listed in Table 5 appear within the top 4 highest confusions in Table 6. Furthermore, the six most often occurring MLP confusions are also the six most often occurring SVM confusions, but with a slightly different ranking. Table 6, thus, backs up what could already be seen in Figure 1; both systems make the same confusions the most often. The misclassifications made by the two AF classification systems are thus not 'random', but contain some structure, as was already suggested by Figure 1. This suggests that there might be something wrong with the description of the AF value itself, instead of something inherently wrong with the classification systems we chose, which are very different: MLPs estimate posterior probabilities, whereas SVMs estimate the optimum decision boundary by maximising the margin between AF values.

4. DISCUSSION

During training, significantly fewer examples of 'dental' were encountered than for the other *place* AF values – just over 15,000 frames in the full training set (1.4%). The poor classification results for 'dental' are thus likely caused by a poor estimation of the posterior probability for 'dental', which leads to a bias towards the other AF value classes. Note that, although the SVM for *place* only received 1,356 frames for 'dental' (in the 100K training frames set), it detects 'dental' better than the MLP, which is expected as SVMs tend to generalise better to sparse data. As shown in Table 6, for both the MLP and the SVM systems, 'dental' is most often classified as 'alveolar'. The places of articulation of 'dental' and 'alveolar' are very close to one another, leading to a small articulatory difference between the two. Both 'dental' and 'alveolar' consonants have a concentration of energy in the higher frequency regions of the spectrum, so the acoustics seem to be similar. Furthermore, the percentage of 'alveolar' frames in the training material is the highest (27.7%) in the training material, thus it is to be expected that 'alveolar' has a better estimated posterior probability distribution or decision boundary than 'dental'.

From Tables 5 and 6 it can be deduced, that the poor classification of 'central' is contributable to the high number of confusions with 'front' and, surprisingly, 'nil'. Looking at the training material distribution shows that 62.3% of the training frames are labelled as 'nil', 5.4% as 'back', 20.4% as 'central', and 11.9% as 'front'. As explained above, this will result in a good classifier (distribution) for 'nil', but poorer ones for the other three AF values. This might explain the 'central'-'nil' confusions. Within this same AF class, there is, however, also a high number of 'back'-'front' and 'back'-'central' confusions. An explanation might be that the confusability of these AFs is contributable to the fact that 'back', 'central', and 'front' are positions along a continuum. Thus, the continuous positions had to be quantised. On top of that, people have different lengths and shapes of the vocal tract. The high number of confusions of 'central' with 'front' combined with the high number of confusions of 'back' with 'central' also suggests that the distribution of 'central' frames is rather broad.

Table 5 shows a more than 30% difference in AF value accuracies for '+round' (8.9% of the training frames) and '-round' (28.7% of the training frames). As Table 6 and Figure 1 show, this is almost totally contributable to the labelling of '+round' frames as '-round'. We suspect there to be a mismatch between the articulatory description as derived from TIMIT and the behavioural reality. This needs further investigation.

Finally, the poor classification results for 'approximant' (Table 5) are contributable to the fact that approximants are in many ways comparable to vowels, making it difficult to distinguish between approximants and vowels (see also Table 6). In our classification scheme (based on [13]), [w], [j], and [l] are marked as 'approximant', but the acoustic properties of [w] and [j] differ little from high vowels. Approximants, however, are usually classified phonetically as consonants due to their functional role in syllabic structure rather than because of their acoustic properties [3].

Although the performances obtained with the *static* AF classifiers are not too bad, what is conspicuous – at least from the SVM system – is the relatively high percentage of support vectors used. This might give an indication as to why this simple two-class classification task performs so much worse than the other two-class task: *voice*. Following [7], the value 'dynamic' in the *static* class is assigned to frames that come from various diverse (groups of) phonemes, which have spectral change occurring during production in common. These include, e.g., diphthongs, laterals, trills, fricatives (e.g., [ð]), and plosives. The SVMs might be at a disadvantage here, since rate of spectral change is better determined with a context window. Our SVMs get only one frame presented at a time, unlike the MLPs which have a context window of -3 and +3 frames. Classifying *static* is thus a difficult task for SVMs. A deeper analysis of the SVMs showed that the support vectors had Lagrange multipliers that did not reach c, which means that they are able to completely separate the training data. However, the width of the RBFs is also small (indicated by the large value for γ). This, coupled with the large number of support vectors, suggests that the clusters representing 'static' and 'dynamic' are irregularly distributed and highly localised, resulting in poor generalisation compared to the MLP system. This can be explained by the great diversity of the (groups of) phonemes assigned with the 'dynamic' label.

5. CONCLUSION AND FUTURE WORK

In our search for automatic AF classifiers that are able to create reliable and accurate AF transcriptions of the acoustic signal, we compared SVMs with MLPs. MLPs have been widely used for this task and have a reasonable level of performance, while SVM classifiers had up till now (to the best of our knowledge) not been used for the task of *multi*-value *acoustic* AF classification.

Both the SVMs and the MLPs are trained discriminatively, but use different optimisation criteria; MLPs estimate posterior probabilities, whereas SVMs estimate the optimum decision boundary by structural risk minimisation. Despite this difference, both systems show similar classification behaviour as is shown by our analyses of the performances of the two systems. Nevertheless, the MLPs outperformed the SVMs on most AF classifications. However, we believe that there is room for improving the SVM (and MLP) classifiers:

1. The MLP systems are trained on more training data than the SVM systems. Although SVMs are known for their ability to work with sparse data, Table 2 suggests that the SVM systems might benefit from more training data; the SVM accuracies are still rising – although they are somewhat flattening for higher number of training frames – indicating that a higher number of training frames will further improve the classification accuracies.
2. The MLP systems used a +3 and –3 frames context window, while the SVM systems did not. Using a context window or more advanced sequence kernels [22] for the SVM systems should improve the performance.
3. In our experiments, we used MFCC features as input for the two classification systems. In follow-up research, we will investigate whether acoustic features based on the human auditory system [4] will improve the performance of the SVMs and MLPs.
4. Use a different set of articulatory or acoustic-phonetic features to describe the acoustic signal.

6. ACKNOWLEDGEMENTS

The research of the first author was supported by the Netherlands Organization for Scientific Research (NWO). The first author would like to thank Louis ten Bosch and Mirjam Wester for their help with the MLP experiments.

7. REFERENCES

[1] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, *2* (*2*), 1998, 1-47.

[2] C.-C. Chang, C.-J. Lin, "LIBSVM: a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] J. Clark, C. Yallop, *An introduction to phonetics and phonology*, *2nd edition*. Oxford, UK: Blackwell Publishers Ltd, 1995.

[4] M.P. Cooke, *Modelling auditory processing and organization.* Cambridge, UK: Cambridge University Press, 1993.

[5] M.H. Davis, W.D. Marslen-Wilson, M.G. Gaskell, "Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition," *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 2002, 218-244.

[6] J. Frankel, *Linear dynamic models for automatic speech recognition*, Ph.D. thesis, The Centre for Speech Technology Research, Edinburgh University, 2003.

[7] J. Frankel, M. Wester, S. King, "Articulatory feature recognition using dynamic Bayesian networks", *Proceedings Interspeech*, Jeju Island, Korea, 2004.

[8] S. Hawkins, "Roles and representations of systematic fine phonetic detail in speech understanding," *Journal of Phonetics*, *31*, 2003, 373-405.

[9] A. Juneja, *Speech recognition based on phonetic features and acoustic landmarks*, Ph.D thesis, University of Maryland, 2004.

[10] R.J.J.K Kemps, M. Ernestus, R. Schreuder, R.H. Baayen, "Prosodic cues for morphological complexity: The case of Dutch plural nouns," *Memory & Cognition*, *33*, 2005, 430-446.

[11] S. King, P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, *14*, 2000, 333-353.

[12] K. Kirchhoff, *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefield, 1999.

[13] P. Ladefoged, *A course in Phonetics, 2nd edition*. Harcourt Brace Jovanovich, 1982.

[14] L. Lamel, R. Kassel, S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *DARPA Speech Recognition Workshop*, 1986, pp. 100-109.

[15] K. Livescu, J. Glass, J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 2529-2532.

[16] P. Niyogi, M.M. Sondhi, "Detecting stop consonants in continuous speech," *Journal of the Acoustical Society of America*, *111*, 2002, 1063-1076.

[17] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Keystone, CO, 1999, pp. 79-84.

[18] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, T. Darrell, "Visual speech recognition with loosely synchronized feature streams," *Proceedings of ICCV*, Beijing, China, 2005.

[19] A.P. Salverda, D. Dahan, J.M. McQueen, "The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension," *Cognition*, *90*, 2003, 51-89.

[20] O. Scharenborg, D. Norris, L. ten Bosch, J.M. McQueen, "How should a speech recognizer work?," *Cognitive Science*, *29* (*6*), 2005, 867-918.

[21] N. Ström, "Phoneme probability estimation with dynamic sparsely connected artificial neural networks," *The Free Speech Journal,* 5, 1997.

[22] V. Wan, S. Renals, "Speaker Verification using Sequence Discriminant Support Vector Machines," *IEEE Transactions of Speech and Audio Processing*, *13:2*, 2005, 203-210.

[23] M. Wester, "Syllable classification using articulatory-acoustic features," *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 233-236.

[24] M. Wester, S. Greenberg, S. Chang, "A Dutch treatment of an Elitist approach to articulatory-acoustic feature classification," *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, pp. 1729-1732.