Available online at www.sciencedirect.com
ScienceDirect
Speech Communication xxx (2007) xxx–xxx
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Towards capturing fine phonetic variation in speech using articulatory features ☆

Odette Scharenborg *, Vincent Wan, Roger K. Moore

*Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK*

Received 31 March 2006; received in revised form 15 December 2006; accepted 11 January 2007

**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 analysed the classification results from support vector machines (SVMs) and multilayer perceptrons (MLPs). MLPs have been widely and 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 performance of the two classifiers and analyses the results in order to better understand the articulatory representations. It was found that the SVMs outperformed the MLPs for five out of the seven articulatory feature classes we investigated while using only 8.8–44.2% of the training material used for training the MLPs. The structure in the misclassifications of the SVMs and MLPs suggested that there might be a mismatch between the characteristics of the classification systems and the characteristics of the description of the AF values themselves. The analyses showed that some of the misclassified features are inherently confusable given the acoustic space. We concluded that in order to come to a feature set that can be used for a reliable and accurate automatic description of the speech signal; it could be beneficial to move away from quantised representations.


*Keywords:* Human speech recognition; Automatic speech recognition; Articulatory feature classification; Fine phonetic variation

## 1. Introduction

In everyday speech it is quite common for there to be no pauses between lexical items; words flow smoothly one into another with adjacent sounds coarticulated. This means that, if words are assumed to be constructed from a limited set of abstract phonemes, then virtually every contiguous phoneme string is compatible with many alternative word sequence interpretations. Human listeners, however, appear to be able to recognise intended word sequences without much difficulty. 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 from psycholinguistic and phonetic research that sub-segmental (i.e. subtle, fine-grained, acoustic–phonetic) and supra-segmental (i.e. prosodic) detail in the speech signal modulates human speech recognition (HSR), and helps the listener segment a speech signal into syllables and words (e.g., Davis et al., 2002; Kemps et al., 2005; Salverda et al., 2003). It is this kind of information that appears to help the human

☆ Earlier results of the research presented in this article have been published in the proceedings of the ISCA Tutorial and Research Workshop on Speech Recognition and Intrinsic Variation, Toulouse, France, May 2006.

* Corresponding author. Tel.: +44 114 222 1907; fax: +44 114 222 1810.
*E-mail addresses:* O.Scharenborg@dcs.shef.ac.uk (O. Scharenborg), V.Wan@dcs.shef.ac.uk (V. Wan), R.K.Moore@dcs.shef.ac.uk (R.K. Moore).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.01.005

perceptual system distinguish short words (like *ham*) from the longer words in which they are embedded (like *hamster*). Salverda et al. (2003), for instance, showed that the lexical interpretation of an embedded sequence is related to its duration; a longer sequence tends to be interpreted as a monosyllabic word more often than a shorter one. Kemps et al. (2005) found that, in addition to duration, intonation seems to help the perceptual system in distinguishing singular forms from the stems of plural forms. However, currently no computational models of HSR exist that are able to model this *fine phonetic variation* (Hawkins, 2003). Our ultimate goal is to refine an existing computational model of HSR 'SpeM' (Scharenborg et al., 2005) such that it is able to capture and use fine-grained acoustic–phonetic variation during speech recognition. SpeM is a computational model of human word recognition built using techniques from the field of automatic speech recognition (ASR) that is able to recognise speech.

*Articulatory features* (AFs) describe properties of speech production and can be used to represent the acoustic signal in a compact manner. AFs are abstract classes which characterise the most essential aspects of articulatory properties of speech sounds (e.g., voice, nasality, roundedness, etc.) in a quantised form, leading to an intermediate representation between the signal and the lexical units (Kirchhoff, 1999). In this work, we are in search of automatic classifiers able to create reliable and accurate transcriptions of the acoustic signal in terms of these articulatory features for the development of a computational model of HSR that is able to model the effect of fine-grained acoustic variation on HSR.

In the field of ASR, AFs are often put forward as a more flexible and parsimonious alternative (Kirchhoff, 1999; Wester, 2003; Wester et al., 2001) to modelling the variation in speech using the standard 'beads-on-a-string' paradigm (Ostendorf, 1999), in which the acoustic signal is described in terms of phones, and words as phone sequences. It is known that speech recognition in adverse conditions poses severe problems for current phone-based ASR systems. However, Kirchhoff (1999) showed that an ASR system based on AFs outperformed HMM-based ASR systems in certain adverse conditions. Furthermore, the modelling of spontaneous speech is a difficult issue for phone-based ASR systems. Many techniques and approaches have been tried to model spontaneous speech phenomena such as coarticulation, but only to limited successes (for an overview, see Strik and Cucchiarini, 1999). AFs offer the possibility of representing coarticulation and assimilation effects as simple feature spreading. For these reasons, we investigate the use of AFs to capture fine phonetic (subphonemic) variation.

Over the years, many different approaches have been investigated for incorporating AFs into ASR systems. For instance, artificial neural networks (ANNs) have shown high accuracies for classifying AFs (King and Taylor, 2000; Kirchhoff, 1999; Wester, 2003).

Frankel et al. (2004) provide a short overview of other modelling schemes, such as hidden Markov models (Kirchhoff, 1999), linear dynamic models (Frankel, 2003) and dynamic Bayesian networks (Livescu et al., 2003). For smaller tasks, support vector machines (SVMs) offer favourable properties: good generalisation given a small amount of high-dimensional data. SVMs have also been applied to the classification of articulatory features (Juneja, 2004; Niyogi and Sondhi, 2002). For instance, Juneja (2004) developed SVM-based landmark detectors for classifying binary place and voicing features in TIMIT (Garofolo, 1988) and reported accuracies ranging from 79% to 95%. Also, Niyogi and Sondhi (2002) 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 (Saenko et al., 2005). This leads us to hypothesise that SVMs could also offer a performance advantage in the classification of multi-level acoustic articulatory features.

In the work reported here, we investigated the possibility of classifying multi-level acoustic articulatory features using SVMs. Given the existing high performance of ANNs on the task of AF classification, the classification performance of the SVMs has been compared with that of multilayer perceptrons (MLPs). Simultaneously, we use the SVMs as a tool for analysis in order to come to a better understanding of the AFs and their respective values (see Section 2.2). In our experiments, we started with a set of articulatory features that has been widely used (e.g., Kirchhoff, 1999; Wester, 2003; Wester et al., 2001) in the front-end of automatic speech recognition systems. An analysis of the AF value classification results is carried out to determine whether those AFs can also be used reliably to describe the speech signal as needed by a computational model able to capture and use fine phonetic detail. The expectation is that an analysis of why specific AF values are more difficult to classify than others, for instance because they are more difficult to derive reliably from the acoustic signal, will lead to ideas for defining an improved set of articulatory features and values that better capture the fine phonetic detail in the speech signal. The results of the experiments and analyses will thus be used to infer a modified set of articulatory features.

In order to allow a direct comparison between the SVM and the MLP, both types of systems have been trained on the same material (see Section 2.1) using the same AF set (Section 2.2). The remainder of Section 2 presents an overview of the experiments presented in this paper and their evaluation. Section 3 outlines details of the two classification systems that were used. Section 4 presents and analyses the results obtained using SVMs. Section 5 presents the results obtained using MLPs and compares these with those of the SVMs. Section 6 discusses the most notable findings. Lastly, conclusions as well as promising directions for future research are presented in Section 7.

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

## 2. Experimental set-up

### 2.1. Material

The training and testing material used in this study are taken from the TIMIT corpus (Garofolo, 1988). 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 3696 utterances. The test set (excluding the sa sentences) consisted of 1344 utterances. The speech was parameterised with 12th order MFCCs and log energy, augmented with 1st and 2nd order derivatives, resulting in 39-dimensional acoustic feature vectors. The features were computed on 25 ms windows shifted by 10 ms per frame.

### 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 (Frankel et al., 2004).

The chosen set is based on the six AFs proposed in Wester (2003). An initial experiment showed that the accuracies for the AF values in the 'place' AF class improved if the vowel-related AF values (*high*, *mid*, *low*) were removed from 'place' and were put in a separate (new) 'high–low' AF class. 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 Ladefoged (1982); note that, following Wester (2003), the silence part of a plosive is mapped onto *stop* and not onto *silence* in our experiments (we return to this in Section 4.4). Table 2 presents an overview of the feature value specification of each of the phone labels in the TIMIT set.

### 2.3. Experiments and evaluation

In the first experiment (Section 4), we trained two types of SVM classification systems for the seven AFs (Table 1).

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' | *+voice*, *−voice* |
| 'high–low' | *high*, *mid*, *low*, *nil*, *silence* |
| 'fr–back' | *front*, *central*, *back*, *nil* |
| 'round' | *+round*, *−round*, *nil* |
| 'static' | *static*, *dynamic* |

For the '−WIN' SVM system, the input of the SVM was presented with single MFCC frames; no context window was used. For the '+WIN' SVM system, the input of the SVM was presented with a context window that included the three preceding and three following frames. This distinction allowed us to discern the potential benefit of using a context window to take into consideration the dynamic nature of speech. In the second experiment (Section 5), we trained a multilayer perceptron (MLP) system also using the +/− three frames context window.

The results for all AF classification experiments are reported in terms of the percentage frames correctly classified, and they are presented at two different levels: per AF (the overall AF classification score) and per AF *value*. This was done because 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. The significance of the difference in performance between two sets of results is calculated using a significance test to compare continuous speech recognisers (Harborg, 1990) and is based on the standard *t*-test.

One of the benefits of using AFs is that they are able to change *a*synchronously, which makes them suitable to describe the variation occurring in natural speech arising from effects such as coarticulation and assimilation. An estimate of the degree of the asynchrony in feature changes in speech is given in Wester et al. (2004) in terms of AF combinations. Feature representations derived from the canonical phonemic transcription resulted in 62 AF combinations. When the features were allowed to change asynchronously, the number of AF combinations increased to 351. A transcription of the speech signal, however, that accounts for asynchronously changing features does not exist. In our experiments, the reference frame labels have therefore been derived by replacing the frame-level phonemic TIMIT labels by the canonical AF values, which causes the features to change *synchronously*. During classification, asynchronously changing AFs will thus be erroneously marked as errors. The impact on frame accuracy the lack of a transcription that accounts for asynchronously changing AFs has is illustrated by King and Taylor (2000). They showed that if the feature is allowed to change within a range of −/+ 2 frames from the phone boundary, the measure "all frames correct" increases significantly by 9% absolute to 63%. The number of errors occurring at canonical phoneme boundaries, thus, when not allowing asynchronously changing features, creates a substantial decrease in the frame accuracy.

The lack of a transcription of the speech signal that accounts for asynchronously changing AFs also means that it is impossible to achieve 100% correct classification on the given task and that the 'upper-bound' of the classification accuracy is also unknown. Nevertheless, we present the

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

Table 2
Feature value specification of each phone label in the TIMIT set

| Phoneme | 'manner' | 'place' | 'voice' | 'high–low' | 'fr–back' | 'round' | 'static' |
|---|---|---|---|---|---|---|---|
| *ae* | vowel | nil | +voice | low | front | −round | static |
| *ax* | vowel | nil | +voice | mid | central | −round | static |
| *ao* | vowel | nil | +voice | low | back | +round | static |
| *aw* | vowel | nil | +voice | low | front | −round | dynamic |
| *ay* | vowel | nil | +voice | low | front | −round | dynamic |
| *b* | stop | bilabial | +voice | nil | nil | nil | dynamic |
| *ch* | fricative | alveolar | −voice | nil | nil | nil | dynamic |
| *d* | stop | alveolar | +voice | nil | nil | nil | dynamic |
| *dh* | fricative | dental | +voice | nil | nil | nil | dynamic |
| *dx* | stop | alveolar | +voice | nil | nil | nil | dynamic |
| *eh* | vowel | nil | +voice | mid | front | −round | static |
| *er* | retroflex | nil | +voice | nil | nil | nil | dynamic |
| *ey* | vowel | nil | +voice | mid | front | −round | dynamic |
| *f* | fricative | labiodental | −voice | nil | nil | nil | static |
| *g* | stop | velar | +voice | nil | nil | nil | dynamic |
| *hh* | fricative | velar | −voice | nil | nil | nil | static |
| *ix* | vowel | nil | +voice | high | front | −round | static |
| *iy* | vowel | nil | +voice | high | front | −round | dynamic |
| *jh* | fricative | alveolar | +voice | nil | nil | nil | dynamic |
| *k* | stop | velar | −voice | nil | nil | nil | dynamic |
| *l* | approximant | alveolar | +voice | nil | nil | nil | dynamic |
| *m* | nasal | bilabial | +voice | nil | nil | nil | static |
| *n* | nasal | alveolar | +voice | nil | nil | nil | static |
| *ng* | nasal | velar | +voice | nil | nil | nil | static |
| *ow* | vowel | nil | +voice | mid | back | +round | dynamic |
| *oy* | vowel | nil | +voice | low | back | +round | dynamic |
| *p* | stop | bilabial | −voice | nil | nil | nil | dynamic |
| *r* | retroflex | alveolar | +voice | nil | nil | nil | dynamic |
| *s* | fricative | alveolar | −voice | nil | nil | nil | static |
| *sh* | fricative | alveolar | −voice | nil | nil | nil | static |
| *t* | stop | alveolar | −voice | nil | nil | nil | dynamic |
| *th* | sricative | dental | −voice | nil | nil | nil | static |
| *uh* | vowel | nil | +voice | high | back | +round | static |
| *uw* | vowel | nil | +voice | high | back | +round | dynamic |
| *v* | fricative | labiodental | +voice | nil | nil | nil | static |
| *w* | approximant | velar | +voice | nil | nil | nil | dynamic |
| *y* | approximant | velar | +voice | nil | nil | nil | dynamic |
| *z* | fricative | alveolar | +voice | nil | nil | nil | static |
| *zh* | fricative | alveolar | +voice | nil | nil | nil | static |
| *em* | nasal | bilabial | +voice | nil | nil | nil | dynamic |
| *en* | nasal | alveolar | +voice | nil | nil | nil | dynamic |
| *eng* | nasal | velar | +voice | nil | nil | nil | dynamic |
| *nx* | nasal | alveolar | +voice | nil | nil | nil | static |
| *axr* | retroflex | alveolar | +voice | nil | nil | nil | dynamic |
| *aa* | vowel | nil | +voice | low | back | +round | static |
| *ah* | vowel | nil | +voice | mid | central | −round | static |
| *ih* | vowel | nil | +voice | high | front | −round | static |
| *hv* | fricative | velar | −voice | nil | nil | nil | static |
| *el* | approximant | alveolar | +voice | nil | nil | nil | dynamic |
| *ux* | vowel | nil | +voice | high | back | +round | dynamic |

results in terms of the percentage of correctly classified frames, for which the output of each of the systems (in the form of an AF value for each frame) is aligned with the reference frame labels. In addition, the most often occurring AF value confusions for each system are presented. We want to compare the performance of SVMs with that of MLPs on the same task; therefore, both systems will 'suffer' the same consequences of being compared to the same reference transcription. The absolute levels of performance of both systems will likely be a bit lower but the differences in the absolute levels of performance will be the same irrespective of the reference transcription used.

Section 4 presents the most remarkable findings and differences between the −WIN and +WIN systems, while Section 5 presents those for the +WIN and MLP systems. Both Sections 4 and 5 end with an in-depth analysis and discussion. The −WIN/+WIN systems' comparison provides insights into: (1) the effect of having knowledge about the spectral change (in the +WIN condition) on the classification accuracy; (2) which AF values are still

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

being classified badly even though knowledge about the context is known. The +WIN SVM/MLP systems' comparison is a cross-check that investigates whether there is a mismatch between the characteristics of the classification systems and the characteristics of the description of the AF values.

## 3. The AF classification systems

### 3.1. Multilayer perceptron AF classification

Seven MLPs (one for each AF) were trained using the NICO Toolkit (Ström, 1997). All MLPs consisted of three layers. Each MLPs' input layer, with 273 nodes, was presented with 39-dimensional MFCC frames with a context window of plus and minus three frames. The hidden layer had tanh transfer functions and a different number of nodes depending upon the AF (see Table 8). In an initial experiment to determine the optimum network size, networks with various numbers of hidden units were trained. The network configurations that gave the best performance in the initial tests are used in the experiments and results presented in Section 5. The output layer was configured to estimate the posterior probabilities of the AF values given the input. The number of output nodes for each MLP is also listed in Table 8.

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.

### 3.2. Support vector machine AF classification

SVMs are binary maximum margin classifiers (for a full introductory text, the reader is directed to Burges, 1998). For this paper we present a brief introduction to provide an insight into these classifiers.

One of the benefits of SVMs over MLPs is that their training may be formulated as a quadratic programming optimisation problem that guarantees a globally optimal solution. However, unlike MLPs, SVMs are not statistical classifiers and do not estimate posterior probabilities directly. The *maximum margin principle* underlying an SVM is illustrated in Fig. 1. Given two separable classes the decision boundary is found by maximising the (margin or) distance between the two dotted parallel lines such that no data occupy the space in-between. The decision boundary is chosen to be the solid line midway between the dotted lines. In this case, the solution for the decision boundary is expressed entirely in terms of the points that lie on the dotted lines, which are known as the support vectors (SVs), and all other points may be discarded.

When the data is non-separable then a soft margin is used that allows some points to enter the margin or be misclassified entirely. Incursions into the margin are penalised so a search for the best solution maximises the margin and minimises the penalties simultaneously. The trade-off between the two is controlled by a single regularisation parameter, $c$, applied as a multiplying factor on the penalties. Smaller values of $c$ will result in solutions that weight margin maximisation more importantly while larger values will move the focus towards fitting the training data which may lead to poor generalisation. Thus $c$ controls how well an SVM generalises to test data. In this case, the SVs are those points that lie within the margin (including those on the 'dotted-line' boundaries) or are misclassified.

SVMs are easily extended to non-linear problems by mapping the data non-linearly onto a manifold embedded in a higher dimensional space and constructing a linear boundary 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 efficiently within the SVM framework by the use of kernel functions (for more detail see Burges, 1998).

Classification of more than two classes is achieved by combining the decisions of several binary SVMs by error correcting codes (Wu et al., 2004). The number of SVMs required depends on the number of classes. For each data point the adopted approach takes the hard decisions of each SVM (encoded, for example, as a 1 if the data point lies on one side of the boundary and 0 if it lies on the other) and combines them into a binary number string. Each unique string is then mapped to a class label.

In our experiments, we used LIBSVM (Chang and Lin, 2001). For the −WIN condition, the input of the SVMs consisted of single MFCC frames. For the +WIN condition a context window of plus and minus three frames was also presented resulting in 273-dimensional MFCC vectors. Two common kernels are the polynomial kernel and the radial basis function (RBF) kernel. In an initial experiment, we tested both the polynomial and the RBF

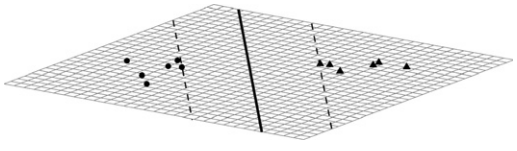

Fig. 1. Finding the decision boundary using SVMs.

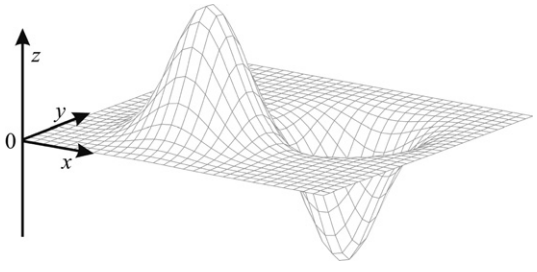


Fig. 2. Visualising the output of an RBF kernel SVM.

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

kernels and found the RBF performed better. Let us examine RBF kernel SVMs in more detail.

Fig. 2 is an illustration of the output of an RBF kernel SVM (it is important to note that the figure is not an illustration of the non-linear mapping to the higher dimensional space). The $x$- and $y$-directions correspond to the inputs to the SVM while the $z$-direction represents the (unthresholded) output score of the SVM at each $\{x, y\}$ coordinate. Fig. 2 shows the simplest possible SVM solution with two SVs, one for each class. The SVM places a spherical Gaussian (with a standard deviation $\sigma$) centred on each SV. The sign of the Gaussian is dependent upon the SV's class label. One may try to infer that RBF kernel SVMs have a loose analogy to density estimation. However, it is not a true density estimate and is actually closer to nearest neighbour clustering. In a classifier involving many SVs, the output score is the weighted sum over a set of basis Gaussians, one centred on each of the SVs and all with the same $\sigma$.

## 4. SVM AF classification results

### 4.1. Classification results per AF

Table 3 (without a context window; '−WIN') and Table 4 (with a context window of ±3 frames; '+WIN') show the classification results of the SVM systems for varying amounts of training data; from 2K training frames (or 0.18% of the total amount of training data) to 500K training frames (or 44.2% of the total amount of training data). These smaller training sets are created by randomly selecting frames from the full training set while maintaining the same prior distribution of the AF value classes as in the full training set. In the case of 'voice', 'fr–back', 'manner', 'static', and 'high–low' there are no results for the 500K training set, because the optimisation did not finish after two weeks. The results are reported in terms of the percentage frames correctly classified for each AF classifier separately. Also, the number of training frames and the percentage of support vectors for each AF classifier 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 $1/\sigma^2$ and $c$ (see Section 3.2) for each SVM in both the −WIN and +WIN condition are listed in Table 5. The results show increasing accuracies (and percentage of support vectors) for increasing number of training utterances. For both −WIN and +WIN conditions, the ranking of the best performing AF classifiers is identical; the best performance is obtained for 'voice', followed by 'round'.

Comparing the results in Tables 3 and 4 shows that, unsurprisingly, using a context window increases the AF accuracies for all AFs. (Although the −WIN condition also uses context knowledge via the first and second order derivatives, such knowledge is more reliable and is encoded differently when using a context window.) The size of this

Table 3
SVM AF classification accuracies (Acc; decreasing from left to right) for each AF and the percentage of support vectors (SV) in each SVM when using no context window

| #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 | 89.5 | 61.6 | 83.2 | 41.8 | 80.3 | 60.7 | 73.8 | 92.4 | 73.3 | 58.5 | 73.1 | 76.4 | 69.7 |
| 10K | 26.8 | 90.3 | 48.7 | 84.8 | 36.9 | 82.3 | 51.2 | 77.0 | 85.4 | 76.0 | 53.1 | 75.9 | 66.7 | 73.5 |
| 50K | 25.1 | 90.8 | 40.4 | 86.1 | 34.0 | 83.4 | 46.8 | 78.9 | 76.4 | 78.0 | 48.7 | 77.6 | 57.5 | 76.4 |
| 100K | 24.2 | 91.0 | 37.3 | 86.6 | 33.3 | 83.7 | 44.6 | 79.6 | 72.0 | 78.6 | 47.8 | 78.0 | 53.8 | 77.5 |
| 500K | 22.9 | 91.3 | 32.3 | 87.3 | 32.0 | 84.3 | 41.5 | 80.8 | 60.9 | 79.8 | 45.8 | 79.0 | 47.3 | 79.4 |

Table 4
SVM AF classification accuracies (Acc; decreasing from left to right) for each AF and the percentage of support vectors (SV) in each SVM when using a 7-frame context window

| #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 | 36.4 | 89.6 | 48.1 | 83.9 | 75.9 | 81.4 | 77.4 | 76.5 | 90.2 | 75.0 | 85.2 | 74.6 | 69.0 | 71.2 |
| 10K | 29.4 | 90.4 | 39.7 | 85.6 | 60.8 | 83.5 | 59.6 | 78.9 | 80.4 | 79.5 | 76.4 | 78.3 | 59.6 | 75.9 |
| 50K | 25.8 | 91.1 | 33.7 | 87.0 | 51.6 | 85.4 | 53.4 | 83.1 | 68.3 | 82.3 | 64.2 | 81.0 | 49.9 | 79.6 |
| 100K | 24.6 | 91.4 | 31.6 | 87.6 | 47.9 | 86.0 | 48.7 | 84.0 | 62.7 | 83.2 | 60.7 | 81.9 | 46.7 | 80.6 |
| 500K | – | – | 27.9 | 88.6 | – | – | – | – | – | – | – | – | 40.4 | 83.1 |

Table 5
Values of the $1/\sigma^2$ and $c$ parameters for each SVM without ('−WIN') and with a 7-frame context window ('+WIN')

| AF | −WIN | | +WIN | |
|---|---|---|---|---|
| | $1/\sigma^2$ | $c$ | $1/\sigma^2$ | $c$ |
| voice | 0.5 | 5 | 0.5 | 1 |
| round | 1.5 | 1 | 0.1 | 1 |
| fr–back | 0.01 | 300 | 0.1 | 5 |
| manner | 0.01 | 15 | 0.01 | 5 |
| static | 10 | 10 | 1 | 1 |
| high–low | 0.01 | 100 | 0.05 | 5 |
| place | 0.1 | 3 | 0.005 | 5 |

increase, however, is not the same for all AFs: comparing the accuracies after training on 100K training frames shows that the increase in accuracy for 'voice' is only 0.4%, while the increase for 'static' is the biggest at 4.6%. The difference in accuracies between the −WIN and +WIN condition is, however, significant at the 99% confidence level for each AF. Adding a ± three frame context window is, thus, beneficial for all AFs but not to the same extent. We return to this issue in Section 4.4.

A second difference between the AF classifiers for 'voice', 'manner', 'fr–back', and 'high–low' is that the percentage of support vectors *in*creased when a context window was used, while there was a *de*crease for 'round', 'static', and 'place'. The explanation of the increase in support vectors is rather straightforward. In the case where no context window is used, the dimensionality of the MFCCs is 39, while the dimensionality increases to 273 when a context window is used. Because of this high dimensionality the data points are more dissimilar, resulting in more support vectors needed to cluster the data. The reduction in support vectors for 'static', 'round', and 'place' is thus surprising, we return to this issue in Section 4.4.

### 4.2. Classification results per AF value

Table 6 lists the classification accuracies in terms of frames correctly classified for each AF value for the SVM classification systems (trained on 100K training frames) as well as the difference in accuracy (all differences are significant at the 99% confidence level). A quick glance at the results shows that the +WIN condition also outperforms the −WIN condition on an AF value level, with the exception of *vowel* and *front*. The differences in accuracies can be as high as 14.9% (e.g., for *central*). The higher AF accuracies for the +WIN condition – reported in Tables 3 and 4 – are thus not simply a result of a better classification of *silence* and *nil*.

For both conditions, the three easiest AF values to classify are +*voice*, −*voice*, and *vowel*, while the three most difficult are *dental*, *central*, and +*round* for +WIN, and *dental*, *central*, and *approximant* for −WIN. The variation in the AF value classification accuracies is rather wide: ranging from 12.5% (*central*) to 91.9% (*vowel*) for the −WIN condition and from 27.3% (*dental*) to 91.7%

Table 6
AF value classification accuracies and differences ('Diff') for the −WIN and +WIN SVM systems and for the MLP system and the difference with the +WIN SVM system, and the percentage of training frames for each AF ('%Frames')

| AF value | Accuracy (%) | | | | | %Frames |
|---|---|---|---|---|---|---|
| | −WIN | +WIN | Diff | MLP | Diff | |
| manner | | | | | | |
| *approximant* | 43.2 | 54.8 | 11.6 | 54.7 | 0.1 | 4.8 |
| *retroflex* | 65.1 | 72.4 | 7.3 | 71.1 | 1.3 | 5.8 |
| *fricative* | 81.7 | 85.6 | 3.9 | 87.2 | −1.6 | 17.1 |
| *nasal* | 73.3 | 77.3 | 4.0 | 79.1 | −1.8 | 6.3 |
| *stop* | 70.9 | 80.3 | 9.4 | 86.1 | −5.8 | 16.1 |
| *vowel* | 91.9 | 91.4 | -0.5 | 91.0 | 0.4 | 34.1 |
| place | | | | | | |
| *bilabial* | 55.1 | 63.2 | 8.1 | 68.4 | −5.2 | 6.1 |
| *labiodental* | 57.8 | 65.4 | 7.6 | 70.9 | −5.5 | 3.1 |
| *dental* | 21.8 | 27.3 | 5.5 | 22.3 | 5.0 | 1.4 |
| *alveolar* | 75.2 | 77.4 | 2.2 | 78.3 | −0.9 | 29.5 |
| *velar* | 50.8 | 55.8 | 5.0 | 64.4 | −8.6 | 8.2 |
| high–low | | | | | | |
| *high* | 70.4 | 71.3 | 0.9 | 71.0 | 0.3 | 12.3 |
| *mid* | 45.3 | 53.4 | 8.1 | 54.9 | −1.5 | 10.5 |
| *low* | 71.3 | 73.9 | 2.6 | 75.7 | −1.8 | 11.4 |
| voice | | | | | | |
| *+voice* | 91.3 | 91.7 | 0.4 | 93.8 | 2.1 | 61.3 |
| *–voice* | 90.4 | 90.8 | 0.4 | 90.3 | 0.5 | 38.7 |
| fr–back | | | | | | |
| *front* | 82.0 | 81.6 | -0.4 | 81.5 | 0.1 | 21.8 |
| *central* | 12.5 | 27.4 | 14.9 | 33.2 | −5.8 | 3.4 |
| *back* | 48.2 | 57.3 | 9.1 | 54.0 | 3.3 | 8.8 |
| round | | | | | | |
| *+round* | 49.2 | 51.5 | 2.3 | 56.8 | −5.3 | 8.9 |
| *−round* | 81.8 | 84.2 | 2.4 | 82.3 | 1.9 | 25.3 |
| static | | | | | | |
| *static* | 81.0 | 85.6 | 4.6 | 84.4 | 1.2 | 56.7 |
| *dynamic* | 75.6 | 80.2 | 4.6 | 81.2 | −1.0 | 43.3 |

(*+voice*) for +WIN. There seems to be, however, no straightforward relationship between the percentage of training frames available for a certain AF value (see column '%Frames' in Table 6) and the obtained AF value's classification accuracy. For instance, the percentage of training frames for *high*, *mid*, and *low* is more or less balanced, but the classification accuracy for *mid* is about 25% lower than the classification accuracies for *high* and *low*. Furthermore, the classification accuracy for *labiodental* is 7% higher than the classification accuracy for *velar*, while there are 2.6 times more *velar* frames in the training data than there are *labiodental* frames. So, there has to be other reasons as to why some of the AF values are so difficult to classify. We will return to this issue in Section 6.

### 4.3. AF value confusions

Table 7 shows an overview of the six most frequently occurring AF value confusions for both the −WIN and the +WIN SVM systems. What is immediately evident is that four of the five AF values that scored the lowest accuracies as listed in Table 6 appear within the top 4 most frequent confusions in Table 7. Furthermore, the six most frequent −WIN confusions are also the six most frequent +WIN confusions, but with a slightly different ranking; both systems, thus, make the same confusions most frequently. The misclassifications made by the two AF classification systems are thus not 'random', but contain some structure.

### 4.4. Discussion and analysis

A context window is usually added to take into consideration the dynamic nature of speech, which usually spans more than the size of one frame (i.e. 25 ms). There are three AFs that are critically dependent on the availability of information on spectral change (as will be explained below), these are 'static', 'manner', and 'place'. These three AFs are, indeed, among the four AFs that have the biggest improvement when using a context window. Unlike the other AF values, 'voice' does not benefit much from using a context window. This is because all the information that is needed for a proper classification can be found in a single frame. The fundamental frequency (F0) range for a male speaker is 80–200 Hz, and for females 150–350 Hz. This means that even in case of the lowest F0 (80 Hz), two full periods are present in a frame of 25 ms.

In our classification scheme (based on Ladefoged, 1982), [w], [j], and [l] are marked as *approximant*, but the formant structure of *approximant* is very similar to the formant structure of high vowels, with the difference being that *approximant* has relatively slow, but clear, formant changes shortly before and after a (incomplete) constriction. Furthermore, the formant structure of *retroflex* is also very similar to the formant structure of *vowel*, with the difference being here that *retroflex* has clear formant changes at the transitions with adjacent vowels and consonants. To be able to distinguish between *vowel*, *approximant*, and *retroflex* information about the spectral changes is fundamental. An analysis of the most occurring confusions showed that the dramatic improvements for *approximant* and *retroflex* (11.6% and 7.3%, respectively) are indeed mainly due to a dramatic decrease in *approximant–vowel* (10.6% absolute improvement) and *retroflex–vowel* (6.9% absolute improvement) confusions. The slight decrease in *vowel* classification is caused by an increase of 0.5% of *vowel-approximant* confusions. Nevertheless, it is clear that the availability of the spectral changes is vital for the classification of *approximant* and *retroflex*. However, the accuracy for *approximant* is still low. Also in the case of *nasal*, the 4.0% improvement in accuracy is due entirely to a reduction in *nasal–vowel* confusions.

A context window can also help to distinguish between a *fricative* and a *stop*. It can help in determining whether a silent period preceded the frication, and thus whether the frication comes from a *fricative* or a *stop*. This is exactly what happened. An analysis of the most frequent confusions revealed that the 9.4% increase in classification accuracy

for +WIN is mainly due to a 3.6% reduction in *stop–fricative* confusions and a 3.5% reduction in *stop–silence* confusions. The 3.9% improvement for *fricative* is largely due to a decrease in *fricative–stop* confusions by 2.4%. These reductions are mainly caused by a reduction in the number of (erroneous or unwanted) 'frame changes': the number of times two consecutive frames have a different label. The number of times two consecutive frames were marked as *silence* and *stop* is 4053 for the −WIN condition while this only occurs 2591 times in the +WIN condition. Likewise, the number of frame changes from *stop* to *fricative* reduces from 3906 to 3321. Adding a context window thus not only 'smoothes' the output in time, but it also ensures that both the silence and the frication parts are marked as *stop*. This suggests that the model of *stop* does not model the different stages of a plosive very well. It might therefore be better not to model the different stages using one AF value, but to relabel the silence part of a *stop* as *stop-closure*.

The AF 'place' is also helped by using a context window. The second formant (F2) values and the amount and direction of F2 changes in vowels adjacent to a plosive or nasal consonant give a clear indication of *where* in the vocal tract the constriction for the plosive or nasal occurred. To execute a gesture and reach the articulatory target, a speaker's articulators need between 30 ms and 100 ms (Rietveld and van Heuven, 1997). Therefore, in this study, not every frame will contain information about articulator movement, since the MFCCs used in this study were created on the basis of a 25 ms windows with a 10 ms shift. Hence, a context window provides this information to the classifiers thereby making it easier to classify the correct place of articulation. This explains the relatively large increase in accuracy (3.1% absolute) for 'place'.

In general, many of the 'place' AF values are often confused with *alveolar*. The relatively large reduction in percentage confusions for *bilabial* is due to a reduction of 3.3% absolute of the *bilabial–silence* confusions and a reduction of 2.9% absolute of the *bilabial–alveolar* confusions. The biggest factor for the improvement of *labiodental* is a reduction of 3.5% in the confusions with *alveolar*. Furthermore, for *dental* the 5.5% improvement is due to a 2.6% absolute reduction in *dental–alveolar* confusions and a 1.6% absolute reduction in *dental–bilabial* confusions. Finally, *velar* had a reduction in *velar–alveolar* confusions (1.4% absolute), and an absolute reduction of 2.2% in *velar–alveolar* confusions. Overall, the classification accuracies for the AF values are somewhat low. We return to this point in Section 6.

Using a context window for the AF 'static' improves both AF values to the same extent. However, it is interesting that the overall classification performance of this binary classification task performs 11.5% worse than the other binary task 'voice' in the −WIN condition (see Table 3, 500K training frames) and 5.8% in the +WIN condition (see Table 4, 500K training frames). The percentage of support vectors for 'static' was relatively high in the −WIN condition (almost three times as many as for 'voice'). Following Frankel et al. (2004), the value *dynamic* in the 'static' class is assigned to frames that come from various diverse (groups of) phonemes, which have spectral change during production in common. These include, e.g., diphthongs, laterals, trills, and plosives. Classifying 'static' is thus a difficult task for SVMs that do not use a context window. A deeper analysis of the SVMs in the −WIN condition showed that the support vectors had Lagrange multipliers that did not reach $c$, which means that they are able to separate the training data completely. However, the width of the RBFs is also small (indicated by the large value for $1/\sigma^2$). 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. This can be explained by the great diversity of the (groups of) phonemes assigned with the *dynamic* label. This problem is somewhat alleviated by using a context window. Using a context window reduced the percentage of support vectors and increased the width of the RBFs (indicated by the much smaller value for $1/\sigma^2$). This indicates that the clusters representing *static* and *dynamic* are less irregularly distributed and less localised. Using a context window thus simplifies the classification task for 'static'; it is easier to group together the input samples, resulting in the biggest increase of the classification accuracy for all AFs. However, compared to 'voice' the percentage support vectors is still much higher and the value for $1/\sigma^2$ is larger. This indicates that the clusters representing *+voice* and *−voice* are much more coherent and regularly distributed than the clusters representing *dynamic* and *static*. This is not surprising since the acoustic phenomena associated with *static* and *dynamic* are much more diverse than those for *+voice* and *−voice*.

The effects of using a context window on the AF values of 'fr–back' are quite different. First of all, the classification accuracy of *front* deteriorates somewhat. This is due to an increase in the *front–central* and *front–back* confusions of 1.0% and 0.4% absolute, respectively. The classification accuracy for *central* in the −WIN condition is quite bad, it is thus not surprising that the accuracy is improved for the +WIN condition. The 14.9% absolute improvement for *central* is due to a 9.9% absolute reduction in *central–front* confusions and a 4.5% absolute reduction in *central–nil* reductions. For *back*, the 9.1% absolute increase is caused by a big reduction in *back–front* confusions (7.9% absolute) and a small increase in *back–central* confusions (1.5% absolute). Although a context window improves the classification accuracy for both *central* and *back*, the classification accuracies remain low, this will be further discussed in Section 6.

Of the 'high–low' AF values, *mid* is least well classified and therefore has the most room for improvement. When a context window is used, indeed *mid* benefits most, increasing with 8.1% absolute in classification accuracy. This increase is mostly due to a reduction in *mid–high* confusions (3.0% absolute) and a reduction in *mid–low* confusions (3.1% absolute). In Section 6, this is further discussed.

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

Table 7
The six most frequent occurring AF value confusions for the SVM and MLP systems where a 'from' AF value is labelled as the 'to' AF value

| −WIN | | | +WIN | | | MLP | | |
|---|---|---|---|---|---|---|---|---|
| *from* | to | % | from | to | % | from | to | % |
| *approx* | *vowel* | 46.6 | *dental* | *alveolar* | 38.6 | *dental* | *alveolar* | 36.9 |
| *dental* | *alveolar* | 41.2 | *approx* | *vowel* | 36.0 | *approx* | *vowel* | 35.5 |
| *central* | *front* | 40.4 | *central* | *front* | 30.5 | *+round* | *−round* | 26.9 |
| *+round* | *−round* | 30.4 | *+round* | *−round* | 30.0 | *central* | *front* | 28.5 |
| *central* | *nil* | 29.5 | *central* | *nil* | 25.0 | *retroflex* | *vowel* | 23.5 |
| *retroflex* | *vowel* | 29.0 | *retroflex* | *vowel* | 22.1 | *central* | *nil* | 22.5 |

A side-effect of using a context window is that the number of times a change in AF value occurs between two consecutive frames is greatly reduced. For instance, for 'place' adding a context window reduced the number of 'frame changes' with 7.4% absolute, while for 'static' the number of frame changes was reduced with 45.0% absolute. Adding a context window thus 'smoothes' the output and removes quick alterations in AF values.

## 5. MLP AF classification results

### 5.1. Classification results per AF

Table 8 shows, in decreasing order, 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 'voice', followed by 'round'. The results in Table 8 are not quite as good as the results presented in Wester (2003), with the exception of the performances for 'place' and 'high–low', which are better.

From Table 8, 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 (see also the discussion on this difference in Section 4.4). On the other hand, 'manner' has seven output nodes, but gets a relatively high accuracy compared to, for instance, 'place' and 'high–low' which have an equal and a lower number of output nodes, respectively, but a lower accuracy.

### 5.2. Comparing MLP and SVM AF classification results

For convenience, the best AF classification accuracies for the SVM classifiers (i.e. those +WIN classifiers trained on 100K training frames) and for the MLP classifiers are listed side-by-side in Table 9. Table 9 also shows the AF classification accuracy at chance level; i.e. that accuracy that would be obtained by a classifier that labelled all frames with the most frequent AF value – since it can safely be assumed that speech/silence detection for TIMIT is easy the chance levels are calculated on the non-silence frames only. It is clear that both SVM and MLP systems perform far above chance level. Comparing the results of the SVM classifiers and the MLP classifiers in Table 9 shows that the two systems have similar performance; the overall rankings for the best performing classifiers are very much alike, with only 'place' and 'high–low' swapping places. The SVMs outperform the MLPs significantly (at the 99% confidence level) for 'fr–back', 'static', and 'high–low', while the MLPs significantly (again at the 99% confidence level) outperform the SVMs for 'voice', 'manner', and 'place'. The slightly better performance of the SVMs for 'round' is not significant. Thus, the SVM systems outperform the MLP systems while only using 8.8% of training frames (100K training frames) that was used to train the MLPs. When increasing the training set for the SVMs to 500K (44.2% of the full training set), the SVMs outperformed the MLPs for another two AF values, 'round' and 'place'.

The SVMs outperformed the MLPs for five out of seven AFs despite using less training material. The training algorithm for SVMs guarantees that a global optimum will be reached, while the back-propagation training algorithm for MLPs only converges to a local optimum.

Table 8
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' | 92.5 | 100 | 2 |
| 'round' | 87.5 | 100 | 3 |
| 'fr–back' | 85.6 | 200 | 4 |
| 'manner' | 84.8 | 300 | 7 |
| 'static' | 82.9 | 100 | 2 |
| 'place' | 81.6 | 200 | 7 |
| 'high–low' | 80.8 | 100 | 5 |

Table 9
Overview of the AF classification accuracies at chance level, for the +WIN SVM systems trained on 100K training frames, and the MLP systems

| AF | Chance level (%) | SVM Acc (%) | MLP Acc (%) |
|---|---|---|---|
| 'voice' | 74.0 | 91.4 | 92.5 |
| 'round' | 59.7 | 87.6 | 87.5 |
| 'fr–back' | 59.7 | 86.0 | 85.6 |
| 'manner' | 40.3 | 84.0 | 84.8 |
| 'static' | 51.8 | 83.2 | 82.9 |
| 'high–low' | 59.7 | 81.9 | 80.8 |
| 'place' | 42.6 | 80.6 | 81.6 |

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

### 5.3. Comparing MLP and SVM AF value classification results

A comparison of the AF value accuracies of the MLP classifiers and the +WIN SVM classifiers is shown in Table 6. Except for *approximant*, *high*, and *front* all differences are significant at the 99% confidence level. A quick glance at the results shows that there is little difference between the two systems. Some of the AF values are better classified by the SVMs while others are better classified by the MLPs; the deteriorations and improvements per AF more or less 'balance' each other. For both types of system, the three easiest AF values to classify are +*voice*, *vowel*, and −*voice*, while the three most difficult are *dental*, *central*, and +*round* for the SVM system, and *dental*, *central*, and *back* for the MLP system. These latter observations are discussed in Section 6.

### 5.4. Comparing MLP and SVM AF value confusions

Similar to the −WIN and +WIN SVM systems, the six most frequently occurring AF value confusions are listed in Table 7. Again similar to the −WIN and +WIN SVM systems, for the MLP system, four of the five AF values that have the lowest accuracies as listed in Table 6 appear within the top 4 most frequent confusions in Table 7. Comparing the SVM and MLP systems shows that the most frequent confusions for both types of systems are the same, but again with a slightly different ranking.

### 5.5. Discussion and analysis

The above analyses indicate that the SVM and MLP systems give similar classification results. But what is striking is that a more detailed analysis of all occurring confusions for both the SVM and the MLP systems revealed that both types of systems also tend to make the same relative number of confusions. The analysis revealed that, in general, if the AF value accuracy for the SVM system is higher than that of the MLP system then all possible confusions for that specific AF value for the SVM system are proportionally lower; and vice versa. There are, however, a couple of differences. Firstly, for all 'manner' and 'place' AF values, remarkably, more *AF value–silence* confusions occur for the SVM system. Secondly, with respect to 'place', *bilabial*, *labiodental*, *dental*, and *velar* are more often classified as *alveolar* by the SVM system. Looking more closely at the AF value classification accuracies for both systems reveals that for all 'real' consonants (during articulation a closure or stricture of the vocal tract occurs that is sufficient to cause audible turbulence; this thus excludes more vowel-like consonants like *approximant* and *retroflex*), the MLP system outperforms the SVM system, with the exception of *dental* but this can be explained by the fact that there are only few training frames for *dental* (see also Section 6.1) and that SVMs are better able in dealing with sparse data. Like the 'real' consonants, the AF value *dynamic* is critically dependent on the availability of spectral change.

## 6. General discussion and further analyses

### 6.1. Place of articulation of consonants ('place')

Miller and Nicely (1955) already pointed out that place of articulation is the easiest to see on the speaker's lips, but the hardest to hear out of all features they investigated (voicing, nasality, affrication, duration, and place of articulation). This study shows that, just as it is for human listeners, place of articulation is the most difficult feature for automatic systems to classify: for all three systems (−WIN SVM, +WIN SVM, and MLP) 'place' and 'high–low' were the two AFs that were classified worst.

As pointed out before, many of the AF 'place' values are most often confused with *alveolar*. We analysed the MFCCs to investigate why *bilabial*, *labiodental*, *velar*, and especially *dental* are confusable with *alveolar*. For the 10K training frames set, we calculated the Bhattacharyya distances for each of the 39 Mel-frequency cepstral coefficients for the *bilabial*, *labiodental*, *velar*, and *dental* frames compared with each of the MFCCs of the *alveolar* frames. The Bhattacharyya distance is a separability measure between two Gaussian distributions and is explained in many texts on statistical pattern recognition (e.g., Fukunaga, 1990). The results are plotted in Fig. 3. The $x$-axis shows the coefficient number, while the $y$-axis shows the Bhattacharyya distance. What immediately stands out is that overall the differences between the distributions are rather small, with the exception of only one coefficient: coefficient 13, which represents the energy (and in case of *bilabial* and *velar* also the first coefficient, representing the overall spectrum shape).

As pointed out in Section 4.4, the change of F2 gives a clear indication of where in the vocal tract the constriction for plosives and nasals occurs. Given that (1) the higher order Mel-frequency cepstral coefficients represent more detailed spectral structure between the formants, and (2) all *bilabial* consonants in English are either plosive or nasal, one would expect that the differences between the means for the *bilabial–alveolar* pair in Fig. 3 for coefficients 6 to 12 are relatively large. There are indeed minor peaks for *bilabial–alveolar* at coefficients 8 and 9. For the other AF values, there are peaks at coefficients 6 and 8 for *labiodental–alveolar*, and peaks at coefficient 9 for *dental–alveolar* and *velar–alveolar*. This suggests that indeed, some information about F2 is represented there, but these peaks are not as pronounced as one might have expected. The largest Bhattacharyya distances are to be found for the lower order MFCCs, which represent the overall spectrum shape and the general formant structure, especially for *labiodental–alveolar* and to a lesser extent *dental–alveolar*.

Of the *bilabial–alveolar* confusions, most confusions are caused by the *nasal* consonants. The two *bilabial* nasals were 29.0% and 36.4% confused with *alveolar*; while for

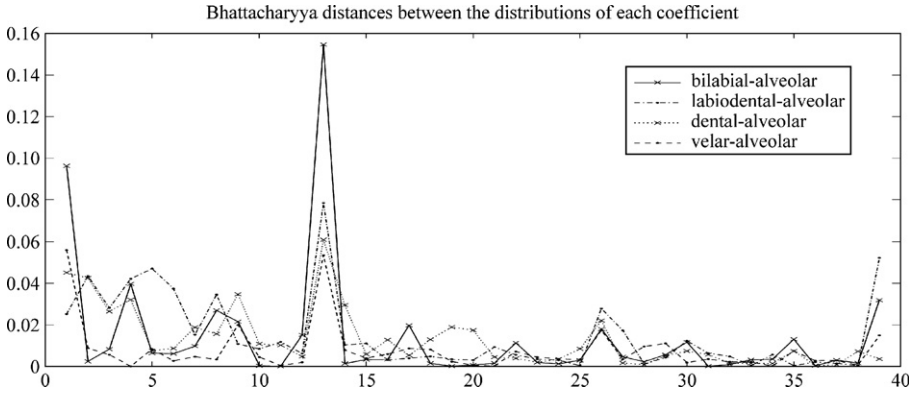


Fig. 3. The Bhattacharyya distances between the 'bilabial', 'labiodental', 'dental', 'velar' frames and the 'alveolar' frames for each of the 39 MFCCs, calculated on the 10K training frames set.

the two *bilabial* plosives this was only 12.4% and 14.4%. Also for *velar* consonants, the nasals were much more often confused with *alveolar* than the other types (fricatives, approximants, and plosives). The confusion percentages for *velar* nasals with *alveolar* ranged from 45.0% to 46.9%; while for the other consonants the confusion percentages ranged from 4.5% to 26.9%. This suggests that the peaks for *bilabial–alveolar* and *velar–alveolar* that are around coefficients 8 and 9 are mostly 'caused' by the *bilabial* and *velar* plosives. It looks like the MFCCs are able to represent the acoustics for the place of articulation for plosives better than for nasal sounds.

As Table 6 shows, *dental* is classified worst of all 'place' AF values and is most often confused with *alveolar* (see Table 7). In English, all *dental* consonants are fricatives. Unlike non-fricative sounds, there are no clear formant changes for *dental* and *alveolar* fricatives that give an indication of where in the vocal tract the constriction occurs; they are distinguished by the overall energy and the duration of the noise. The peaks in the Bhattacharyya distances, especially for the lower order MFCCs, are most likely due to the fact that there are no non-fricative *dentals* but there are non-fricative *alveolars*, which thus have clear formant changes. Since the places of articulation of *dental* and *alveolar* are very close to one another, there is only 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. Fig. 3 confirms that the most important coefficient for distinguishing *dental* and *alveolar* indeed turns out to be the energy (coefficient 13), with coefficient 1 representing the overall spectrum shape being of a slightly lesser importance. Secondly, during training, significantly fewer examples of *dental* were encountered than for the other 'place' AF values – just over 15K frames in the full training set (1.4%, see Table 6). 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 1356 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. Furthermore, the percentage of *alveolar* frames in the training material is the highest (29.5%, see Table 6) in the training material, thus it is to be expected that *alveolar* has a better estimated posterior probability distribution or decision boundary than the other AF values.

Like *dental* consonants, *labiodental* consonants in English are fricatives (with the exception of the nasal [ɱ] which is an allophone of [m], but this phoneme is not transcribed in TIMIT). This means that *labiodental* fricatives can only be distinguished from *alveolar* fricatives using the intensity and the duration of the energy. Again, this is shown in Fig. 3: coefficient 13, the energy, contributes most to the difference between *alveolar* and *labiodental*. The comparatively low percentage of confusions of *labiodental* with *alveolar* (16.8%; the lowest of all AF values) is most likely due to the rather large differences in the lower order MFCCs; they are more pronounced than the differences in the lower order MFCCs for *dental-alveolar*.

### 6.2. Place of articulation of vowels ('front–back' and 'high–low')

From Tables 6 and 7 it can be deduced that the poor classification of *central* (27.4% for +WIN) results from the high number of confusions with *front* and, surprisingly, *nil*. Table 6 shows that 8.8% of the training frames are labelled as *back*, 3.4% as *central*, 21.8% as *front*, and thus 66.0% as *nil*. As explained above, this will result in a good posterior probability for *nil* – and good classification accuracy for *nil* – but poorer ones for the other three AF values. This might explain the rather low accuracies for both *central* and *back* (57.3%). Furthermore, this might also explain the *central–nil* and *back–nil* confusions. However, there are also high numbers of *central–front* (30.5%), *back–front* (19.4%), and *central–back* (17.1%) confusions. An explanation might be that the confusability of these AFs results from the fact that *back*, *central*, and *front* are positions along a continuum of tongue positions from the *back* of the mouth to the *front* of the mouth. Thus, the continuous positions have to be quantised. There are, however,

problems with using quantised values. People have different lengths and shapes of the vocal tract; articulation is thus speaker dependent. Furthermore, articulation positions are not absolute. These two factors combined can result in a (broad) range of possible values of MFCCs associated with the same quantised AF value. The high number of confusions of *central* with *front* combined with the high number of confusions of *back* with *front* also suggests that the distribution of *central* frames is rather broad. We examined this by plotting the distribution of the SVM scores calculated by the SVMs for the test material.

We trained three separate classifiers, one each for *front*, *central*, and *back*, to discriminate that specific AF value from the other two. We then scored the test material using the three classifiers and examined the distribution of the SVM scores. Fig. 4 shows scatter plots of the SVM scores for the *central* (darker crosses) and the *back* (lighter dots) frames. The *x*-axis denotes the SVM score obtained with the classifier trained on the *back* frames; the *y*-axis denotes the SVM score of the corresponding frame obtained with the classifier trained on the *central* frames. In the left panel of Fig. 4, the SVM scores of all *central* and *back* frames are plotted; in the right panel, the SVM scores of only the correctly classified *central* and *back* frames are plotted. Note that misclassified points that were removed include points that were classified as *front*. It can be seen from the right panel that there are many points close to the diagonal line indicating that the SVMs scored those points similarly. The left panel shows how much the SVM score distributions overlap. Similar figures are obtained for all combinations of fr–back. Fig. 5 shows the distribution of the SVM scores of the test material as scored by the classifier trained to classify *front* (left panel), *central* (middle panel), and *back* (right panel). The solid line represents the distribution of the SVM scores of the *front* frames, the dashed line represents the *central* frames, and the dotted line represents the *back* frames. The further apart the distribution of the AF value on which the classifier is trained and the distributions of the other two AF values, the easier that AF value can be separated from the two other. So, in the left panel, one wants to see the distribution of *front* as far away as possible from the distributions of *central* and *back*.

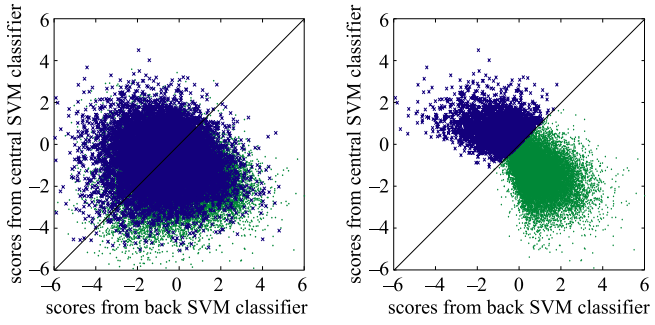


Fig. 4. Scatter plots of the SVM scores of the *central* and *back* test material as scored by the classifiers trained to discriminate *back* from *front* and *central* (*x*-axis) and *central* from *front* and *back* (*y*-axis). In the left panel, the SVM scores of all *central* (darker crosses) and *back* (lighter dots) frames are plotted; in the right panel, the SVM scores of only the correctly classified *central* and *back* frames are plotted.

The most important observation to be made from Fig. 5 is that the distribution for *front* overlaps the distribution for *central* entirely for each classifier. This means that it is impossible for these classifiers to separate *central* from *front* given the MFCCs. This finding explains the low classification accuracy of *central* and its high confusion rate with *front* (see Table 7). Fig. 5 also presents an explanation for the rather low classification accuracy for *back*: there is only a small number of *back* frames that lie outside of the distribution for *front*. These frames will most likely be classified correctly, while the others are being classified as *front*. Our assumption that the high number of confusions of *central* and *back* with *front* was due to the quantisation of the front-back continuum is thus correct: the MFCCs do not allow reliable estimation of the front–back continuum.

Similar to the 'fr–back' AF value *central*, the classification accuracy for the 'high–low' AF value *mid* is relatively poor (53.4% for +WIN), with high confusion rates with *low* (14.7%) and *high* (14.1%). Unlike 'fr–back', however, the distribution of the training frames for the 'high–low' AF values is more or less balanced (*high*: 12.3%; *mid*: 10.5%; *low*: 11.4% of the training frames). This suggests that the placing of the decision boundaries or the estimation of the posterior probabilities is not necessarily leading to a bias towards one of the AF value classes. A further

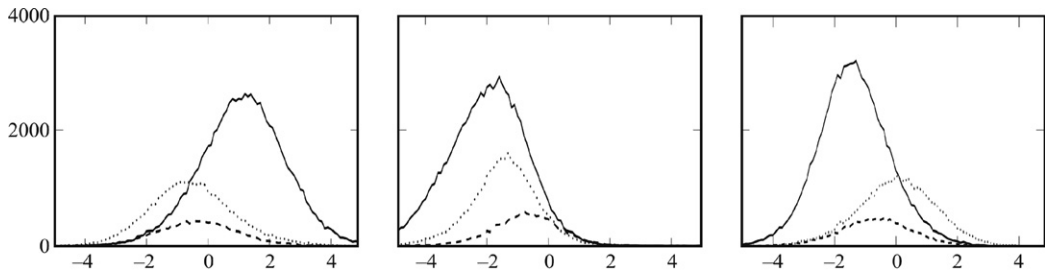

Fig. 5. Distributions of the SVM scores of the 'fr–back' test material as scored by the classifier trained to discriminate *front* from *central* and *back* (solid line, left panel), *central* from *back* and *front* (dashed line, middle panel), and *back* from *front* and *central* (dotted line, right panel). The solid lines represent the distribution of scores for the *front* AF value, the dashed lines represent *central*, and the dotted lines represent *back*.

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

analysis of the confusions for *low* and *high* revealed that they were hardly ever confused with one another, but relatively often with *mid* (11.1% for *low* and 11.3% for *high*). The same explanation as for 'fr–back' might be applicable for 'high–low'; the AF values are positions along a continuum of tongue positions from the roof of the mouth to the bottom, which has to be quantised, introducing some error. Again, we also investigated the distribution of the SVM scores calculated by the SVMs for the test material in order to investigate the separability of the three classes.

Again, we trained three separate classifiers, one for *low*, *mid*, and *high*, tested each classifier on the test material, and examined the distribution of the SVM scores. We believe that the distributions of the SVM scores as presented in Fig. 5 are easier and clearer to interpret than the scatter plots in Fig. 4; therefore, all subsequent results are presented in the form of the distributions of the SVM scores. Fig. 6 shows the distributions of the SVM scores of the test material as scored by the classifier trained to discriminate *low* (left panel), *mid* (middle panel), and *high* (right panel) from their respective other AF values. In each figure, the solid line represents the distribution of the SVM scores of the *low* frames, the dashed line represents the *mid* frames, and the dotted line represents the *high* frames.

Fig. 6 shows that the distributions for the three AF values generated by the three classifiers are overlapping in a fashion consistent with the rather low classification accuracies as listed in Table 6. More importantly, however, the distribution for *mid* as scored by the *mid* classifier (middle panel) is overlapping with *low* and *high* even more. This can be explained by taking into account the actual pronunciation of *mid* vowels. For the articulation of *mid* vowels, the tongue will be in-between the positions for the articulation of *low* and *high* vowels, the acoustics associated with *mid* will thus be in-between the acoustics of *low* and *high*. The explanation for the somewhat disappointing results for *mid* are thus caused by the – not so surprising – fact that the frames for *mid* are very similar to *low* and *high* frames, making it very hard for the classification systems to tell *mid* apart from the other two AF values. The finding that *low* and *high* are more often confused with *mid* than each other is also to be explained by looking at these distributions. In the left panel, the *mid* distribution is much closer to the *low* distribution than the *high* distribution; and vice versa for the right panel.

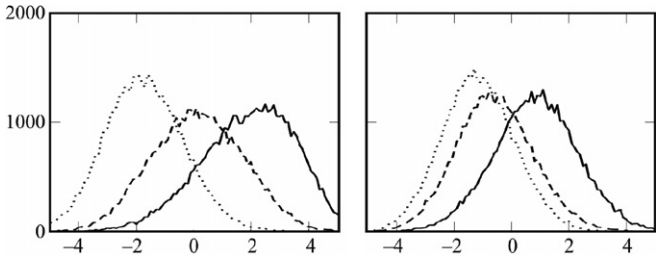


Fig. 7. Distributions of the SVM scores of the 'high–low' test material as scored by the classifier trained to discriminate *low* from *high* only (left panel) and *low* from *mid* only (right panel). The solid lines represent the distribution of scores for the *low* AF value, the dashed lines represent mid and the dotted lines represent *high*.

In addition to these classifiers (see Fig. 6), we also trained classifiers to discriminate *low* from *high* only (Fig. 7, left panel) and *low* from *mid* only (Fig. 7, right panel). The distributions of Fig. 7 show how the SVMs behave when they have been trained on only two out of the three AF values. In the left hand panel, *mid* was excluded from training but during testing it is clearly placed in-between the distributions of *high* and *low*. When *high* was omitted from training (in the right hand panel), the SVM still places the *high* distribution to the left of *mid*. These results suggest that there is indeed a continuum from *low* to *high* via *mid* and that the mapping from phonemic labels to a quantised set of AF values is inaccurate with respect to the acoustic phenomena associated with the *low*, *mid*, and *high* frames. Indeed quantisation is inconsistent with the physical system that we wish to model. Furthermore, Fig. 7 suggests that it is better to train SVMs on the extremes of 'high–low' and allow them to infer the continuum than to train separate classifiers to identify artificially quantised AF values. The accuracy of the classifiers can be inferred from the areas under the curves; in the left panel of Fig. 6, setting the threshold at the crossover point between the *high* and *low* distributions (ignoring the *mid* AF value in the test data) the accuracy of discriminating *high* from *low* is 84.9% for *high* and 82.4% for *low*; in the right panel of Fig. 6, setting the threshold at the crossover point between the *high* and *low* distributions, the accuracy

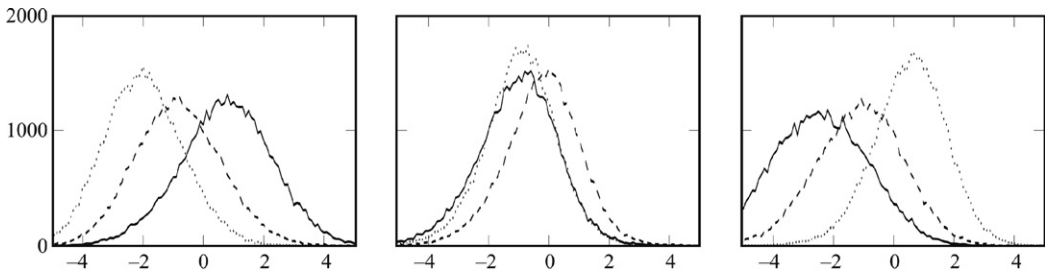


Fig. 6. Distributions of the SVM scores of the 'high–low' test material as scored by the classifier trained to discriminate *low* from *mid* and *high* (solid line, left panel), *mid* from *low* and *high* (dashed line, middle panel), and *high* from *mid* and *low* (dotted line, right panel). The solid lines represent the distribution of scores for the *low* AF value, the dashed lines represent *mid*, and the dotted lines represent *high*.

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

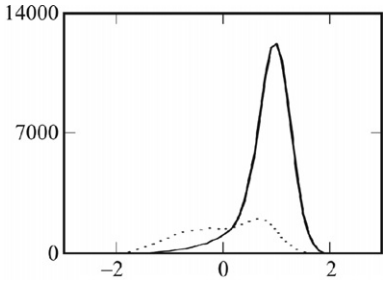


Fig. 8. Distributions of the SVM scores of the 'round' test material for +*round* (dotted line) and −*round* (solid line).

of discriminating *high* from *low* is 78.9% for *high* and 86.9% for *low*; in Fig. 7 (left panel), again setting the threshold at the crossover of the *high* and *low* distributions, the corresponding accuracies are 88.5% and 87.6%, i.e. including *mid* for training reduces the accuracy of *high* and *low*.

### 6.3. Round

Table 6 shows an approximately 30% difference in AF value accuracies for +*round* and −*round*. As Table 7 shows, this is almost totally caused by the classification of +*round* frames as −*round*. Although the percentage of frames in the training data labelled as +*round* (8.9%, see Table 6) is about one third the percentage of training frames labelled as −*round* (25.3%, see Table 6), the amount of training data is not as unbalanced as, for instance, was the case for *dental* in relation to *alveolar*.

In a subsequent analysis, we examined the distribution of the SVM scores calculated by the SVMs for the test material like we did for 'fr–back' and 'high–low'. Fig. 8 shows the distributions of the SVM scores of the +*round* (dotted line) frames and −*round* (solid line) frames of the test material. The further apart the two distributions are, the easier the +*round* and −*round* frames can be separated. As is clear from Fig. 8, the distribution of –*round* is Gaussian shaped and the majority of the –*round* frames are above '0'. The distribution of +*round*, however, seems bimodal, with the frames belonging to the right-most 'Gaussian' shape having an SVM score above '0' as well. These +*round* frames will incorrectly be classified as −*round*. Based on these results, we suspect there to be a mismatch between the articulatory description as derived from TIMIT and the actual way the speaker has produced the sound, in other words, the behavioural reality. This needs further investigation.

## 7. Concluding remarks and future work

To develop a computational model of HSR that is able to simulate the effect of fine-grained acoustic variation on human speech perception, we are in search of AF classifiers that are able to create reliable and accurate AF transcriptions of the acoustic signal. To this end, we analysed the classification results from SVMs and MLPs. MLPs have been widely used for the task of articulatory feature classification and have a reasonable level of performance, while SVM classifiers had until 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 maximising a margin. Despite this difference, both systems show similar classification behaviour as is shown by our analyses of the performances of the two systems. However, the SVMs significantly outperformed the MLPs for five out of the seven AFs while only using 8.8–44.2% of the training material used to train the MLPs.

The classification accuracies obtained for the AF values varied widely; by more than 70% absolute. This behaviour and the very low classification scores for, for instance, *dental*, *central*, *mid*, and *approximant* (see also Table 6 and Section 4.2) cannot simply be explained by the fact that an AF reference transcription of the speech signal was used in which the AF values did not change asynchronously, since the errors introduced by using the canonical AF transcription will have affected all AF values to more or less the same extent (see also Section 2.3). An in-depth analysis of the classification performance of the SVMs was carried out to get a better understanding of the articulatory features and to explain why some of the investigated AFs were so difficult to classify. The expectation was that this analysis would give an indication of the way to proceed towards the definition of a feature set that can be used for a reliable and accurate automatic description of the speech signal. The structure in the misclassifications of the SVMs and MLPs suggested that there might be a mismatch between the characteristics of the classification systems and the characteristics of the description of the AF values themselves. The analyses presented in Section 6.2 showed that some of the misclassified features are inherently confusable given the acoustic space. Furthermore, the preliminary results presented in Section 6.2 suggested that it is better to train SVMs on samples of the extremes of an AF class distribution and allow them to infer the intermediate points of the continuum than to train separate classifiers to identify artificially quantised AF values. Thus, in order to come to a feature set that can be used for a reliable and accurate automatic description of the speech signal, it is concluded that it could be beneficial to move away from quantised representations. This will be confirmed in follow-up research.

In our experiments, we used MFCCs as input for the two classification systems. It is, however, questionable whether MFCCs are the most appropriate type of acoustic features for the task of articulatory feature classification (for instance, as was already pointed out in Section 6.2, MFCCs do not allow reliable estimation of the 'front–back' continuum), and whether other types of acoustic feature will yield improved performance because they better capture the AF value information. Additionally,

Please cite this article in press as: Scharenborg, O. et al., Towards capturing fine phonetic variation in speech using ..., Speech Commun. (2007), doi:10.1016/j.specom.2007.01.005

other types of acoustic feature might show differences between the MLP and SVM classifiers that can further improve our understanding of the AFs and provide insights into an improved definition of a feature set that can be used for a reliable and accurate automatic description of the speech signal. Future research will investigate whether acoustic features based on the human auditory system (Cooke, 1993) will improve the classification performance of the SVMs and MLPs and our understanding of the AFs.

In the current study, we tried to get the best possible performance for the SVM and MLP classification systems. However, for a computational model of HSR it is important to model human recognition behaviour. For example, such a computational model should correctly model the pattern of confusion of AF values. Comparisons of human and computer recognition behaviour have shown that, e.g., *voicing* information is recognised much more poorly by machines than by human listeners (Cooke, 2006; Meyer et al., 2006). In follow-up research, we intend to extend the comparisons of the articulatory feature recognition/classification confusion patterns of human listeners and computers to include all AF values. Capturing a better understanding of fine phonetic detail will be achieved by examining such confusion patterns at the AF value level.

## 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 for his help with the MLP experiments and Stuart Wrigley for help with the analysis. Furthermore, the authors would like to thank three anonymous reviewers for their useful comments on an earlier version of this manuscript.

## References

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Disc. 2 (2), 1–47.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available from: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Cooke, M.P., 1993. Modelling Auditory Processing and Organization. Cambridge University Press, Cambridge, UK.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Amer. 199 (3), 1562–1573.

Davis, M.H., Marslen-Wilson, W.D., Gaskell, M.G., 2002. Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. J. Exp. Psychol.: Human Percept. Perform. 28, 218–244.

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

Frankel, J., Wester, M., King, S., 2004. Articulatory feature recognition using dynamic Bayesian networks. In: Proc. Interspeech, Jeju Island, Korea.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed. Academic Press Inc.

Garofolo, J.S., 1988. Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database, National Institute of Standards and Technology (NIS), Gaithersburgh, MD.

Harborg, E., 1990. Hidden Markov Models applied to automatic speech recognition. Dr.ing-thesis, NTH, Trondheim.

Hawkins, S., 2003. Roles and representations of systematic fine phonetic detail in speech understanding. J. Phonetics 31, 373–405.

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

Kemps, R.J.J.K., Ernestus, M., Schreuder, R., Baayen, R.H., 2005. Prosodic cues for morphological complexity: the case of Dutch plural nouns. Memory Cognition 33, 430–446.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Comput. Speech Lang. 14, 333–353.

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

Ladefoged, P., 1982. A Course in Phonetics, second ed. Harcourt Brace Jovanovich.

Livescu, K., Glass, J., Bilmes, J., 2003. Hidden feature models for speech recognition using dynamic Bayesian networks. In: Proc. Eurospeech, Geneva, Switzerland, pp. 2529–2532.

Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B., 2006. A human–machine comparison in speech recognition based on a logatome corpus. In: Proc. Workshop on Speech Recognition and Intrinsic Variation, Toulouse, France.

Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27 (2), 338–352.

Niyogi, P., Sondhi, M.M., 2002. Detecting stop consonants in continuous speech. J. Acoust. Soc. Amer. 111, 1063–1076.

Ostendorf, M., 1999. Moving beyond the 'beads-on-a-string' model of speech. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, CO, pp. 79–84.

Rietveld, A.C.M., van Heuven, V.J., 1997. Algemene Fonetiek. Coutinho: Bussum, The Netherlands.

Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J., Darrell, T., 2005. Visual speech recognition with loosely synchronized feature streams. In: Proc. ICCV, Beijing, China.

Salverda, A.P., Dahan, D., McQueen, J.M., 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition 90, 51–89.

Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M., 2005. How should a speech recognizer work? Cognitive Sci. 29 (6), 867–918.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Comm. 29 (2–4), 225–246.

Ström, N., 1997. Phoneme probability estimation with dynamic sparsely connected artificial neural networks. Free Speech J. 5.

Wester, M., 2003. Syllable classification using articulatory-acoustic features. In: Proc. Eurospeech, Geneva, Switzerland, pp. 233–236.

Wester, M., Greenberg, S., Chang, S., 2001. A Dutch treatment of an Elitist approach to articulatory-acoustic feature classification. In: Proc. Eurospeech, Aalborg, Denmark, pp. 1729–1732.

Wester, M., Frankel, J., King, S., 2004. Asynchronous articulatory feature recognition using dynamic Bayesian networks. In: Proc. IEICI Beyond HMM Workshop, Kyoto, Japan.

Wu, T.-F., Lin, C.-J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. J. Machine Learning Res. 5, 975–1005.