

Using Temporal Information for Improving Articulatory-Acoustic Feature Classification

Barbara Schuppler, Joost van Doremalen, Odette Scharenborg, Bert Cranen, Lou Boves

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

{B.Schuppler, J.vanDoremalen, O.Scharenborg, B.Cranen, L.Boves}@let.ru.nl

Abstract—This paper introduces new acoustic features that obtain both a high temporal and a high frequency resolution such that is able to detect and reliably classify articulatory events of short duration, such as bursts in plosives. SVM classification experiments on TIMIT and a subset of SVitchboard showed that acoustic features based on MFCCs derived from a long window of 25ms and from a short window of 5ms that are both shifted with 2.5ms steps outperform standard MFCCs. Finally, comparison of the TIMIT and SVitchboard results showed that classifiers trained on data that allows for asynchronously changing AFs (SVitchboard) outperform classifiers trained on data where these changes do not occur (TIMIT).

I. INTRODUCTION

Most automatic speech recognition (ASR) systems are based on the principle that words are composed of a sequence of phones, also referred to as the ‘beads on a string’ model of speech [1]. This model works reasonably well for carefully produced speech. However, ASR performance drops tremendously for spontaneous speech, mainly due to the high pronunciation variability [2]. Phone-based modeling of pronunciation variation has its limitations, because it is not able to capture the overlapping, asynchronous gestures of the articulators, e.g., [3]. Therefore, there has been an increased interest in articulatory-acoustic features (AFs), which are the acoustic correlates of articulatory events. With this type of features speech can be represented in a way that does not impose a sequence of discrete segments. An estimate of the degree of asynchrony in AF changes in speech is given in [4] in terms of AF combinations. AF representations derived from the canonical phonemic transcriptions resulted in 62 AF combinations. When the AFs were allowed to change asynchronously, this number increased to 351.

AF classifiers have been used to improve speech recognition performance in adverse conditions [5], [6], to build language independent phone recognizers [7], and to improve computational models of human word recognition [8]. Furthermore, AF-based descriptions of the speech material are now being used to investigate pronunciation errors by learners of a second language [9] and for the automatic analysis of fine-phonetic detail [10]. For these latter two applications, an accurate and reliable classification is crucial.

However, the evaluation of the performance of AF classifiers suffers from the absence of large corpora that provide reliable labeling of AF values. As a consequence, training and testing of AF classifiers is generally done on the basis of data that is labeled on the phoneme level after which all phonemes are replaced by their (canonical) AF values.

Thus, the AF values change synchronously at Phone boundaries. Moreover, it is unclear to what extent the classifiers trained with AFs obtained in this fashion can be assumed to yield classification results that truly reflect articulatory gestures.

The aim of the present study is two-fold. The first aim is to build an AF classifier that can be used for reliable and accurate detection of slight pronunciation errors and the automatic analysis of fine-phonetic detail. One error second language learners often make is the confusion of fricatives and plosives [11]. Moreover, plosives show tremendous articulatory variation in casual speech [10]. Therefore, in this paper, we focus on improving the automatic classification of the manner of articulation (see Section I.A.). Secondly, we investigate the effect of the AF labeling of the training and test material on performance estimates (see Section I.B).

A. Improving manner classification through capturing temporal information in the acoustic features

Previous research has attempted to improve AF classifiers from basically two directions. First, different statistical classifiers have been tested and their (frame-based) classification accuracies have been compared [5], [12], [13]. Second, different methods to parameterize the acoustic waveforms have been evaluated for the task of AF classification [5], [14]. Most AF classification systems use MFCCs as the acoustic features. However, independent of what features may be used, hardly anyone has moved away from using a window length of 25ms and a window shift of 10ms¹, probably because these settings are known to be close to optimal for ASR purposes. With this window length and shift reasonably good results can be obtained for more or less stationary sounds, for whose correct classification a high resolution in the frequency domain is more important than a high temporal resolution. For instance, around 90% of the vowels were classified correctly in a classification task [12] on TIMIT [17]. However, many articulatory events are not stationary or are much shorter than 25ms and can therefore not be properly captured. This is especially the case for plosives. For comparison, the same study [12] showed that only 80% of the plosives were correctly classified. In the TIMIT database, which is often used for investigation of AFs, 10% of the segments are shorter than 25ms.

¹ To our knowledge, only two studies used a different window length and shift: [15] used a 16ms window and a 8ms shift, while [16] used a 25ms window shifted with 2.5ms.

In order to obtain both a high time and frequency resolution, we propose an acoustic feature vector with MFCCs derived from a window of 25ms and from a short window of 5ms that are both shifted with 2.5ms steps. We will investigate the usefulness of such extended acoustic features by training an AF classifier based on a support vector machine (SVM) and subsequently test it on TIMIT. The reason for using an SVM is that this type of classifiers provides good generalization given a small amount of high-dimensional data, and that SVMs have shown good AF classification results (e.g., [12]). The classifier trained with the new acoustic feature vectors based on a combination of short and long windows will be compared to a baseline classifier with ‘standard’ MFCCs in order to investigate whether these acoustic features achieve a better frame classification accuracy.

B. Dealing with inaccurate material

Ideally, AF classifiers should be trained and tested on observed articulatory trajectories or on speech corpora that are manually transcribed on the articulatory level. However, these data are not available in sufficient quantity, and the creation of these data is extremely time-consuming. For instance, [19] reported that transcriptions of utterances on the AF level take 1000 times real-time for SVitchboard [20] (a selection of the Switchboard corpus that can be considered a small vocabulary data set). The conventional ‘solution’ to obtain enough training and testing material is to use a canonical mapping from phonetic alignments to AF values. Thus, the phoneme /t/ would map to the AF values ‘voiceless’, ‘alveolar’, and a sequence of ‘closure’ and ‘burst+release’. However, for Dutch, it has been shown that only 11.5% of word-final /t/s are realized according to this canonical feature representation [10]. Similar pronunciation variation can be expected for other sounds and in other languages. Thus, the mapping from phonemic labels to AF labels results in classifiers that are trained and tested on stretches of speech that may not contain the assigned feature value at all. This raises doubts about the validity of the data any classifier is trained and evaluated on.

In the absence of a sufficient amount of AF labeled data, we followed the standard procedure: the reference frame labels are derived by replacing the frame-level phonemic TIMIT labels by the canonical AF values. This thus implies that if one wants to evaluate the degree of success with which a classifier is able to correctly classify the acoustic correlates of underlying articulatory gestures, while in actual practice the phone-based canonical labels are used as the reference, asynchronously changing AFs may be erroneously marked as errors. The impact on *apparent* frame accuracy due to the lack of a transcription that accounts for asynchronously changing AFs is illustrated by [21]. They showed that if a feature is allowed to change within a range of $-/+ 2$ frames from the phone boundary, the measure “all frames correct” increases significantly by 9% to 63%. Therefore, the number of such virtual errors occurring at phone boundaries creates a substantial (and for diagnostic purposes, misleading) decrease in the frame accuracy when not allowing asynchronously changing AFs. Note that 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 unknown. In order to investigate the effect of the synchronously changing AF values around Phone boundaries on the errors the AF classification systems make, we carry out an in-depth analysis of the classification results around the phone boundaries.

Finally, the fact that the AF classifier are trained on stretches of speech that may not contain the assigned feature value can result in an AF classifier that suggests a good performance in detecting plosives even if part of the plosives are not produced as a closure and a release. Since our goal is to build classifiers that are able to detect slight pronunciation errors and fine-phonetic detail, it is important that the classifier can distinguish different plosive realizations. Therefore, in addition to evaluating the classifier on the TIMIT data, we also test the new acoustic features on the SVitchboard utterances that have been transcribed at the AF value level [19]. Comparing the new classifier with the baseline system on both data sets will show whether the new classifier is better able to describe the speech signal at the articulatory feature level.

TABLE I
MAPPING OF TIMIT PHONE SYMBOLS TO THE MANNER AF VALUES.

Phone	Manner AF value
sil, pau	silence
l, el, r	liquid
w, y	glide
em, en, eng, m, n, ng, nx	nasal
dh, f, hh, s sh, th, v, z, zh, hv	fricative
b, d, g, p, t, k, q	burst+release
bcl, dcl, gcl, pcl, tcl kcl	closure
dx, epi, all vowels	NIL

II. METHOD

A. Articulatory feature values

In the past, different methods have been proposed to characterize m manner of articulation. For example, plosives can either be mapped as a whole on one AF value ‘plosive’ [12], [21], [22] or split up into two parts ‘closure’ and ‘release’, where the release can be modeled together with ‘fricative’ [5], [23] or separate from the fricatives as a ‘burst+release’ [13], [15]. The latter is also the way we deal with plosives, for two reasons. First, modeling plosives as a unit violates the assumption of SVMs that the sequence of frames assigned to a sound can be considered as drawn from one population, which is definitely not the case for plosives that consist of a sequence of ‘closure’ and ‘release’. Secondly, we aim at using AF classifiers to analyze how the sounds were actually produced, for instance whether a plosive was realized as a closure followed by friction or as a closure plus burst and friction. Therefore we train separate classifiers for ‘fricative’ and ‘burst+release’.

We excluded all ‘vowel’ material, because it is possible that certain AF values overlap in time, for instance in the case of nasalized vowels. However, when including both ‘vowel’ and ‘nasal’ as an AF value these AF values cannot co-occur in

a frame, as the classifier has to make a choice between the two. A full overview of the manner AF values is given in Table 1.

B. Speech material

1) TIMIT

The speech material used in this study is taken from TIMIT, which contains hand-labeled and hand-segmented quasi-phonetically balanced sentences read by 630 speakers (of which 70% were male) speakers from eight major dialects of American English. We followed TIMIT’s training and testing division, in which no sentence or speaker appears in both the training and test set. The training set consists of 3,696 utterances.

To train the SVM classifiers, a smaller training set of 25,210 10ms frames was created by randomly selecting frames from the full training set with the same prior distribution of the AF value classes as in the full training set. Note: since the window shift is 2.5ms in the case of the new acoustic features, the original 25,210 training frames were split into 100,842 2.5ms frames to train the new classifiers.

In the first experiment (Section III.A.), the classifiers were tested on the TIMIT test set consisting of 1,344 utterances (excluding the sa sentences); i.e., 236,984 10ms frames and 943,604 2.5ms frames.

The TIMIT database is labeled using 59 Arpabet symbols, which have been relabeled in terms of AF values according to Table 1. We should note that it is possible that segments that have been annotated as ‘burst+release’ in the TIMIT material indeed contain a burst; however, it is also possible that the burst is actually missing, just leaving frication.

2) SVitchboard

For the second experiment, we used a data set that consisted of 78 utterances (a total of 119s of speech, excluding initial and final silences) drawn from SVitchboard, a small-vocabulary subset of Switchboard, which contains spontaneous telephone speech [18]. This subset is converted into a set of 13,295 10ms frames, and a set of 53,115 2.5ms frames. The subset was transcribed manually on the AF level by ‘XC’ for the 2006 JHU Summer Workshop [19]. The original set of AFs did, however, not match our set of AFs. Therefore, manual adaptations were made by one of the authors (BS) in the following way. Starting point was the transcriptions from the tier on which the feature set ‘Dg1’ (Degree of forward constriction) was annotated. These transcriptions were modified to our feature set. Table 2 shows the mapping between these two feature sets. When modifying the annotations,

- the original boundaries were maintained, but the labels were changed according to Table 2; e.g., the original label ‘fricative’ was changed to ‘burst+release’ in those plosives where a burst was present. Releases without burst maintained the label ‘fricative’.
- new boundaries were placed when one feature in the ‘Dg1’ set is transcribed as two features in our set; e.g., an ‘approximant’ in the original set occasionally was replaced by a ‘liquid’ followed by a ‘glide’ in our set.

- new boundaries were placed to separate background speakers from silence and given the label ‘BG’.
- the boundaries of the ‘Nasality’ tier of the original transcription were used to annotate the nasal consonants.
- obvious labeling mistakes that occurred in two utterances were corrected.

Like in Experiment 1, vowel segments have been excluded. Furthermore, frames labeled as flap or containing background speech have been removed.

TABLE 2
PHONE-TO-AF MAPPING FOR THE ORIGINAL ‘Dg1’ (DEGREE OF FORWARD CONstriction) AND OUR AF SET.

Phone	Dg1	Our AF set
l, el	closure	liquid
er, r	approximant	liquid
w, y	approximant	glide
em, en, eng, m, n, ng, nx	closure	nasal
dh, f, hh, s sh, th, v, z, zh, hv	fricative	fricative
b, d, g, p, t, k, q	fricative	burst+release or fricative
bcl, dcl, gcl, pcl, tcl kcl	closure	closure
	silence	silence

TABLE 3
THE %SV AND VALUES FOR c AND γ FOR DIFFERENT ACOUSTIC FEATURES.

	%SV	c	γ
Baseline	55.0	0.4	2
Short	46.1	0.8	1
Long	31.1	0.5	4
Both	41.8	0.3	2

C. Support Vector Machines

The AF classifiers built in this study are support vector machines [24]. In our experiments, we used the LIBSVM package [25], which achieves multi-class classification by error correcting codes. The RBF kernel was used for the experiments reported in this paper.

For the AF classifiers, trained using four sets of acoustic features (cf. Section III.A), the number of support vectors (SVs) is listed in Table 3 as a percentage of the amount of training data (%SV). The percentage of SVs indicates the task complexity: more SVs suggest either more complex decision boundaries or more overlapping data. Table 3 also lists the γ and c parameters in the SVMs, estimated on an independent development set of 2,000 frames. A large γ implies narrower RBFs. If c is large, the more complex decision boundaries are constructed to fit the training data, but this may result in poor generalization.

Table 3 shows that the %SV is lowest for *Long*, while the value for γ is highest. This indicates that the width of the RBFs is reasonably small, which suggests that the clusters in the *Long* model are more localized, with little overlap between the AF values, resulting in a fair generalization. The c values of the four models do not differ much. The %SV is highest for *Baseline*, while the value for γ is reasonably low; this indicates that the width of the RBFs is reasonably large resulting in less localized AF clusters and less generalization

than *Long* and *Both*. *Short* has the lowest generalization; while *Both* is most likely in between *Baseline* and *Long*.

III. EXPERIMENTS

A. Experiment 1

The first experiment investigated the effect of different sizes of the window over which the MFCCs are calculated. This experiment gives insights into the effect of improving the temporal resolution in the MFCCs. In total four different MFCC representations were compared:

- *Baseline*: window size: 25ms; window shift: 10 ms.
- *Short*: window size: 5ms; window shift: 2.5ms.
- *Long*: window size: 25ms; window shift: 2.5ms.
- *Both*: the *Short* and *Long* MFCCs are concatenated.

For all acoustic features, the input speech is first divided into overlapping Hamming windows of 25ms or 5ms with a 10ms or 2.5ms shift and a pre-emphasis factor of 0.97. For the 25ms windows 12 MFCCs plus C0, and their first and second order derivatives were calculated (39 features). For the 5ms windows, 6 MFCCs plus C0 and first and second order derivatives were calculated (21 features). Afterwards, cepstral mean subtraction (CMS) was applied.

Adding context information has shown to improve classification performance (e.g., [12]). We therefore carried out different tests to determine the optimal amount of temporal context, which was 30ms at both sides. For *Baseline*, 3 frames (30 ms) left and right of a frame were taken into account resulting in MFCC vectors of length $7 \cdot 39 = 273$. Since the window shift is different for the baseline system and the three other systems, the context was incorporated slightly differently in the *Short*, *Long* and *Both* classifiers: for these three classifiers +/- 3 frames were taken but taking only every fourth frame, in order to cover the same temporal context left and right of the frame as *Baseline*. This resulted in feature vectors of length 273 for *Long* and 147 for *Short*. To combine the window lengths (25ms and 5ms) for *Both*, feature vectors of both windows with the same midpoint were concatenated, resulting in feature vectors of length 273 (from the 25ms window) + 147 (from the 5ms window) = 420.

1) Results

Table 4 shows the performance of the four AF classifiers in terms of percentage correctly classified frames on the TIMIT test material in confusion matrices. For clarity, the percentages along the diagonal are in bold. Comparing the three new acoustic features with the baseline system shows that the *Short* classifier performs best for ‘burst’ (Bur) as is to be expected, because the burst of a plosive is an event with a short duration. The *Long* and *Both* classifiers perform best for ‘frication’ (Fri). The AF value that profits most from adding temporal information is ‘liquid’ (Liq). In summary, the best performing classifier is *Both*, which outperforms the other classifiers, except for silence (Sil) and closure (Clo).

TABLE 4
CONFUSION MATRICES FOR THE FOUR BEST PERFORMING AF CLASSIFIERS ON TIMIT; BL=THE BASELINE SYSTEM.

BL	Sil	Liq	Gli	Nas	Fri	Bur	Clo
Sil	0.83	0.00	0.01	0.00	0.02	0.03	0.10
Liq	0.00	0.86	0.04	0.04	0.03	0.02	0.01
Gli	0.00	0.13	0.81	0.03	0.01	0.01	0.01
Nas	0.01	0.03	0.02	0.85	0.03	0.01	0.05
Fri	0.01	0.02	0.01	0.02	0.87	0.05	0.03
Bur	0.03	0.02	0.01	0.01	0.12	0.77	0.05
Clo	0.05	0.01	0.00	0.03	0.04	0.04	0.83

Short	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	0.72	0.00	0.02	0.01	0.05	0.06	0.14
Liq	0.00	0.84	0.05	0.05	0.03	0.01	0.01
Gli	0.00	0.16	0.75	0.04	0.02	0.01	0.01
Nas	0.01	0.07	0.03	0.76	0.05	0.01	0.06
Fri	0.01	0.01	0.01	0.02	0.88	0.05	0.02
Bur	0.02	0.02	0.01	0.01	0.12	0.78	0.04
Clo	0.10	0.01	0.00	0.02	0.04	0.04	0.79

Long	Sil	Liq	Gli	Nas	Friq	Bur	Clo
Sil	0.78	0.00	0.01	0.00	0.03	0.05	0.13
Liq	0.00	0.89	0.04	0.04	0.02	0.01	0.01
Gli	0.01	0.12	0.81	0.03	0.01	0.01	0.01
Nas	0.01	0.03	0.01	0.87	0.03	0.00	0.05
Fri	0.02	0.01	0.01	0.01	0.89	0.04	0.03
Bur	0.03	0.02	0.01	0.01	0.12	0.77	0.05
Clo	0.06	0.01	0.00	0.03	0.04	0.04	0.81

Both	Sil	Liq	Gli	Nas	Fri	Bur	Clo
Sil	0.76	0.00	0.01	0.00	0.03	0.05	0.15
Liq	0.00	0.89	0.04	0.03	0.02	0.01	0.01
Gli	0.00	0.12	0.81	0.03	0.01	0.01	0.01
Nas	0.01	0.03	0.01	0.86	0.03	0.00	0.05
Fri	0.02	0.01	0.01	0.01	0.89	0.03	0.03
Bur	0.03	0.02	0.01	0.01	0.01	0.78	0.04
Clo	0.05	0.01	0.00	0.03	0.04	0.04	0.82

2) Analysis

As explained above, the synchronously changing AF values at the phone boundaries in the training data are likely to incur errors around those phone boundaries in the test material. In order to further analyze these errors and to investigate the influence of improving the temporal resolution of the MFCCs on these errors, the classification output of the best performing new classifier *Both* was further analyzed and compared to the *Baseline* system. To that end, the percentage correctly classified frames for the 20ms following a phone boundary or leading towards a phone boundary are calculated for each of the AF values separately. As *Baseline* consists of a 10ms window shift, only two frames are analyzed; the *Both* system uses 2.5ms window shifts, resulting in eight frames that are analyzed. With this method, only phonemes of a minimum duration of 40ms could be analyzed, and the presented figures therefore only represent those 75.4% of segments in TIMIT that have this minimum length.

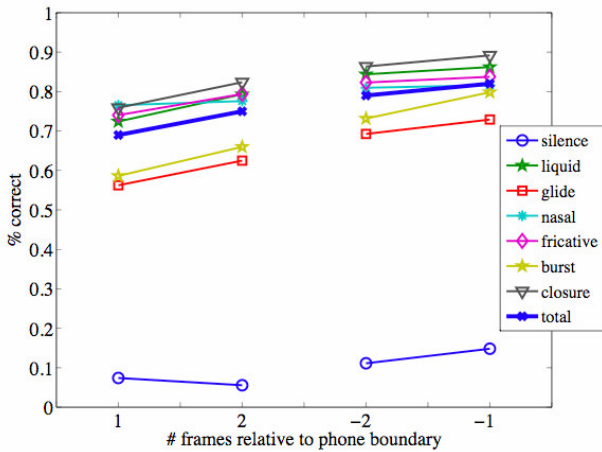


Fig. 1. The frame accuracy over time for each of the AF values for *Baseline*. The bold line shows the overall performance.

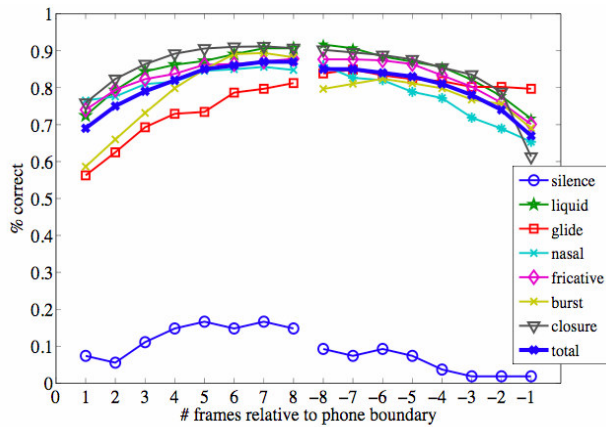


Fig. 2. The frame accuracy over time for each of the AF values for *Both*. The solid line shows the overall performance.

Figs. 1 (*Baseline*) and 2 (*Both*) show the percentage correctly classified frames for each of the AF values separately and the overall classification score over time. The positive numbers indicate the frame numbers counting from the start of the phone, the negative numbers indicate the frames numbers counted from the end of the phone. As expected, the performance of the classifiers is lower towards the boundaries as there the co-articulation effects are strongest. These results suggest that indeed the synchronously changing AFs in the test material results in a reduced frame accuracy. However, the central frames of the phone do not reach 100% correct classification accuracy; this might be due to the fact that the models are corrupted due to the synchronously changing AFs in the training material. While the frame accuracies over time for most AF values show the same pattern, there are a few exceptions. Silence shows extremely low accuracies close to the phone boundaries. The nasals have the most constant frame accuracy throughout the phone. Glides and bursts show opposite patterns, namely the accuracy

for glides is very low at the beginning of the phone while this is the case for bursts at the end of the phone. Note that those AFs that do not follow the overall pattern are those that profit the least from the *Both* system compared to *Baseline*.

Finally, comparing *Baseline* and *Both* shows that although the first frame after a boundary and the last before a boundary have relatively low accuracies, the accuracies rise faster for *Both* than those for *Baseline*. These results seem to suggest that improving the temporal resolution results in better classifiers, which are also more robust against the errors introduced by the synchronously changing AFs.

B. Experiment 2

In the second experiment, the best performing new acoustic features are compared with the baseline system on the SVitchboard data. To that end, new SVMs are trained using the acoustic features of the *Both* system and the *Baseline* system on the SVitchboard data and subsequently tested in a 5-fold cross validation scheme. Folds were generated by randomly dividing the whole dataset into five parts. Note that, as opposed to the TIMIT experiment, there was not a stringent separation between utterances from training and test speakers. *Both* classifiers were trained five times on 80% of the material and tested on the remaining 20%. Each time, γ and c parameters were optimized on the training material by training and testing on two random subsets (from the training set) with different parameters.

With this experiment we intend to investigate the impact of labeling accuracy and its interaction with improved temporal resolution.

3) Results and Discussion

Table 5 shows the performance of *Baseline* and *Both* in terms of percentage correctly classified frames on the SVitchboard material in confusion matrices. As is clear from the percentages on the diagonal, *Both* outperforms *Baseline* for all AF values, except for silence (Sil) where they perform equally well. The performance of *Both* on the SVitchboard data is substantially better than on TIMIT. Despite the differences between the two experiments we firmly believe that this is mainly due to more accurate labels in SVitchboard than in TIMIT. Furthermore, the difference between *Both* and *Baseline* is much bigger than for the TIMIT data. The difference is biggest for ‘burst’ (Bur), which is according to expectation as here the temporal resolution resulting from the short windows gives additional information about the bursts, which are short and dynamic acoustic events. Thus, it seems safe to conclude that when the training material does allow for asynchronously changing (and consequently more accurate) AFs (SVitchboard), the upper bound for the classification performance is raised. Moreover, in the presence of more accurate labels in training and test the gain obtained thanks to improved temporal resolution is larger than the improvement obtained in TIMIT, where the AF labels are less accurate because of the synchronous changes imposed by the mapping from phonetic symbols to AF labels.

TABLE 5

CONFUSION MATRICES FOR THE FOUR BEST PERFORMING AF CLASSIFIERS ON SVITCHBOARD; BL=THE BASELINE SYSTEM.

BL	Sil	Liq	Gli	Nas	Fri	Bur	Clo
Sil	0.96	0.00	0.00	0.01	0.01	0.01	0.01
Liq	0.01	0.90	0.04	0.02	0.02	0.01	0.01
Gli	0.02	0.05	0.87	0.04	0.02	0.02	0.01
Nas	0.01	0.02	0.03	0.84	0.04	0.01	0.05
Fri	0.02	0.01	0.02	0.02	0.83	0.05	0.05
Bur	0.01	0.01	0.00	0.01	0.04	0.83	0.10
Clo	0.02	0.01	0.02	0.03	0.06	0.03	0.84

Both	Sil	Liq	Gli	Nas	Fri	Bur	Clo
Sil	0.96	0.00	0.00	0.00	0.02	0.00	0.01
Liq	0.01	0.95	0.01	0.01	0.01	0.00	0.00
Gli	0.00	0.02	0.95	0.01	0.01	0.00	0.00
Nas	0.00	0.02	0.01	0.93	0.02	0.00	0.02
Fri	0.02	0.01	0.00	0.01	0.89	0.03	0.04
Bur	0.00	0.00	0.00	0.00	0.04	0.92	0.03
Clo	0.01	0.00	0.01	0.01	0.05	0.01	0.90

IV. CONCLUSIONS

The aim of the present study was two-fold. The first aim was to build an AF classifier that can be used for reliable and accurate detection of slight pronunciation errors and the automatic analysis of fine-phonetic detail. In order to improve the automatic classification of the manner of articulation, we proposed new acoustic features that obtain both a high temporal and a high frequency resolution so that it becomes possible to detect and reliably classify articulatory events of short duration, such as bursts in plosives. The results showed that combining MFCCs derived from a long window of 25ms and from a short window of 5ms that are both shifted with 2.5ms steps outperforms standard MFCCs. The added value of temporal information was found when testing the SVM classifiers on TIMIT and on a subset of SVitchboard.

Secondly, we investigated the effect of the AF labeling of the training and test material on performance estimates. Comparing the results obtained on TIMIT and on SVitchboard showed that the classifiers trained on data that allows for asynchronously changing AFs (SVitchboard) outperform classifiers trained on data where these changes do not occur (TIMIT). Thus in order to train reliable and accurate AF classifiers, training and test material that allows for asynchronously changing AFs is crucial. Finally, these results also seem to suggest that classifiers trained on data that allows for asynchronously changing AFs seem to yield classification results that reflect articulatory gestures.

ACKNOWLEDGMENTS

Barbara Schuppler is supported by the Marie Curie Research Training Network Sound-to-Sense. Joost van Doremalen is supported by the project DISCO, funded by the Dutch-Flemish programme STEVIN. Odette Scharenborg is supported by a Veni-grant from the Netherlands Organization for Scientific Research (NWO).

REFERENCES

- [1] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," *IEEE Automatic Speech Recognition and Understanding Workshop*, 1999, pp. 79–84.
- [2] M. Saraçlar, H. Nock, S. Khudanpur, "Pronunciation modelling by sharing gaussian densities across phonetic models," *Computer Speech and Language*, 14, pp. 137–160, 2000.
- [3] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. thesis, University of Bielefeld, Germany, 1999.
- [4] M. Wester, J. Frankel, S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," *Proc. IEICI Beyond HMM Workshop*, 2004.
- [5] K. Kirchhoff, G. A. Fink, G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, 37, pp. 303–319, 2002.
- [6] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," *Proc. ICSLP*, 1998, pp. 891–894.
- [7] S. M. Siniscalchi, T. Svendsen, C.-H. Lee, "Towards a detector-based universal phone recognizer," *Proc. ICASSP*, 2008, pp. 4261–4264.
- [8] O. Scharenborg, "Using durational cues in a computational model of spoken-word recognition," *Proc. Interspeech*, 2009.
- [9] J. Tepperman, S. Narayanan, "Using Articulatory Representations to Detect Segmental Errors in Nonnative Pronunciation," *IEEE transactions on audio, speech, and language processing*, 16 (1), pp. 8–22, 2008.
- [10] B. Schuppler, W. van Dommelen, J. Koreman, M. Ernestus, "Wordfinal [t]-deletion: An analysis on the segmental and sub-segmental level," *Proc. Interspeech*, 2009.
- [11] K. Truong, A. Neri, C. Cucchiari, H. Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," *Proc. InSTIL/ICALL*, Venice, Italy, 2004, pp. 135–138.
- [12] O. Scharenborg, V. Wan, R. K. Moore, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication*, 49, 2007, 811–826.
- [13] J. Frankel, M. Wester, S. King, "Articulatory feature recognition using dynamic bayesian networks," *Computer Speech and Language*, 21 (4), pp. 620–640, 2007.
- [14] O. Scharenborg, M. P. Cooke, "Comparing human and machine recognition performance on a VCV corpus," *Proc. Workshop on Speech Analysis and Processing for Knowledge Discovery*, 2008.
- [15] F. Pernkopf, T. V. Pham, J. A. Bilmes, "Broad phonetic classification using discriminative bayesian networks," *Speech Communication*, 51, pp. 151–166, 2009.
- [16] T. Pruthi, C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, 43, pp. 225–239, 2004.
- [17] J. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIS), Gaithersburgh, MD, 1988.
- [18] S. King, C. Bartels, J. Bilmes, "SVitchboard 1: Small vocabulary tasks from switchboard 1," *Proc. Interspeech*, 2005, pp. 3385–3388.
- [19] K. Livescu, A. Bezman, N. Borges, L. Yung, Ö. Çetin, J. Frankel, S. King, M. Magimai-Doss, X. Chi, L. Lavoie, "Manual transcriptions of conversational speech at the articulatory feature level," *Proc. ICASSP*, 2007.
- [20] S. King, C. Bartels, J. Bilmes, "SVitchboard 1: Small vocabulary tasks from switchboard 1," *Proc. Interspeech*, 2005, pp. 3385–3388.
- [21] S. King, P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, 14, pp. 333–353, 2000.
- [22] S. M. Siniscalchi, T. Svendsen, C.-H. Lee, "Towards bottom-up continuous phone recognition," *Proc. ASRU*, 2007 pp. 566–569.
- [23] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, Ö. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," *Proc. Interspeech*, 2007, pp. 2485–2488.
- [24] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 2 (2), 1–47, 1998.
- [25] 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>.