Int J Speech Technol (2017) 20:699–713
DOI 10.1007/s10772-017-9436-y

# Rethinking classification results based on read speech, or: why improvements do not always transfer to other speaking styles

Barbara Schuppler[1]

Received: 16 January 2017 / Accepted: 27 June 2017 / Published online: 18 July 2017

**Abstract** With the growing interest among speech scientists in working with natural conversations also the popularity for using articulatory–acoustic features as basic unit increased. They showed to be more suitable than purely phone-based approaches. Even though the motivation for AF classification is driven by the properties of conversational speech, most of the new methods continue to be developed on read speech corpora (e.g., TIMIT). In this paper, we show in two studies that the improvements obtained on read speech do not always transfer to conversational speech. The first study compares four different variants of acoustic parameters for AF classification of both read and conversational speech using support vector machines. Our experiments show that the proposed set of acoustic parameters substantially improves AF classification for read speech, but only marginally for conversational speech. The second study investigates whether labeling inaccuracies can be compensated for by a data selection approach. Again, although an substantial improvement was found with the data selection approach for read speech, this was not the case for conversational speech. Overall, these results suggest that we cannot continue to develop methods for one speech style and expect that improvements transfer to other styles. Instead, the nature of the application data (here: read vs. conversational) should be taken into account already when defining the basic assumptions of a method (here: segmentation in phones), and not only when applying the method to the application data

**Keywords** Articulatory–acoustic features · Conversational speech · Pronunciation variability · Segments

## 1 Introduction

Speech science and technology used to rely on the assumption that speech utterances can be described as a sequence of words and that words are composed of a sequence of phones, also known as the 'beads on a string' model of speech (Ostendorf 1999). This model works satisfactorily for carefully produced speech, but it runs into problems with conversational speech, mainly due to the high pronunciation variability (Saraçlar et al. 2000). To some extend, the incorporation of pronunciation variants into the lexicon has shown to improve ASR systems e.g., Baum (2003); Lehtinen and Safra (1998). However, by adding variants also the internal confusability increases. For instance, Kessens et al. (2003) found that adding variants in principle decreases the WER compared to a lexicon with only canonical pronunciations as long as the average number of variants is ≤2.5 (specific for their system architecture). Nonetheless, there is evidence that the variability in conversational speech is much higher than what is possible to model with such a low number of variants. For instance, Greenberg (1999) reports an average of 22.2 pronunciation

Earlier results and parts of the research presented in this article are published as workshop paper (Schuppler et al. 2013). SPASR Workshop on Speech Production in Automatic Speech Recognition, Lyon, France) In comparison to the Workshop paper, here we present a more profound review of the literature, a more detailed analysis and discussion of the results as well as additional experiments to provide better support for the conclusions drawn.

✉ Barbara Schuppler
b.schuppler@tugraz.at

[1] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c, 8010 Graz, Austria

variants for the 100 most frequent words of the Switchboard corpus. Phone-based modeling of pronunciation variation is simply not able to capture the overlapping, asynchronous gestures of the articulators (e.g., Kirchhoff 1998; Fosler-Lussier et al. 1999). Therefore, there is an interest in articulatory-acoustic features (AFs), i.e., the acoustic correlates of articulatory gestures, as basic unit of representation. AFs that can change asynchronously seem to offer a natural way for representing (semi-) continuous articulatory gestures and the ensuing acoustic characteristics of speech signals (e.g., Frankel et al. 2007b).

In the last decades, AFs have received increasing interest in the field of speech technology (e.g., Bitar and Espy-Wilson 1996; Frankel et al. 2007b; Hasegawa-Johnson et al. 2005; Juneja 2004; Juneja and Espy-Wilson 2008; King and Taylor 2000; King et al. 2007; Kirchhoff et al. 2000; Manjunath and Sreenivasa Rao 2016; Naess et al. 2011). Instead of building acoustic models for phones, separate classifiers are trained for articulatory-acoustic features such as manner of articulation, place of articulation, and voicing. AF classifiers have been successfully used for speech recognition in adverse conditions (e.g., Kirchhoff 1999; Kirchhoff et al. 2002; Schutte and Glass 2005), to build language-independent phone recognizers (e.g., Stüker et al. 2003; Lin et al. 2009; Siniscalchi et al. 2008; Siniscalchi and Lee 2014), in the area of visual automatic speech recognition (Saenko et al. 2005), and in computational models of human spoken-word recognition (Scharenborg 2010). Furthermore, the combination of phone-based acoustic modeling with AFs have shown to reduce the word error rate in task-oriented spontaneous speech (Kirchhoff et al. 2000; Metze 2005, 2007).

Even though the main reason for using AFs instead of phones is that AFs have more potential for capturing pronunciation variation, most investigations on AF classification have been carried out on read speech, while conversational speech is (far) more prone to pronunciation variation than read speech (Schuppler et al. 2014). Even though it is well known that TIMIT (Garofolo 1988), a read speech corpus of American English, is non-generic, it continues to be the most popular corpus for research into AF classification American English (e.g., Pernkopf et al. 2009; Pruthi and Espy-Wilson 2004; Siniscalchi et al. 2007). The main reasons for using TIMIT are the high number of speakers, the quality of the manually created phonetic transcriptions and the large set of phonetic labels. To our knowledge, only the work done during the 2004 Johns Hopkins Summer Workshop (Frankel et al. 2007a; Hasegawa-Johnson et al. 2005; Livescu et al. 2007), and the work by Greenberg and Chang (2000), Pruthi and Espy-Wilson (2007) and Naess et al. (2011) used Switchboard (Godfrey et al. 1992), a corpus of American English spontaneous conversations. Other studies which go beyond AF classification for read speech are those based on German task-oriented spontaneous speech by Kirchhoff et al. (2000) and Metze (2005, 2007). Recently, Manjunath and Sreenivasa Rao (2016) compared the performance of their AF classifiers for read, spontaneous and conversational speech of a Corpus of the Indian language Bengali.

Research on AF classification has focused on finding the ideal set of acoustic parameters for building multi-value or binary classifiers (e.g., Salomon et al. 2004; King and Taylor 2000; Niyogi et al. 1999; Scharenborg et al. 2007) or the ideal statistical classification method (e.g., a comparison of ANNs with SVMs by Chang et al. 2005). For an overview of different methods used for AF classification see King et al. (2007). These experiments have shown improvements of the newly proposed acoustic parameters and classification methods over baseline conditions, for read speech. In general it is assumed that classification performance in spontaneous speech is simply lower than in read speech but that new methods also yield an improvement there. It remains to be actually shown whether improved classification generalizes from read speech to conversational speech.

This paper presents two studies that investigate this question from two different viewpoints: acoustic parameter selection and classification method development. These two experiments will be outlined in the two following subsections. The first study aims to develop acoustic parameters for accurate classification of *manner* of classification. The requirement for the acoustic parameters, thus, is to have both a high frequency resolution (for stationary sounds like nasals and fricatives) and to have a high temporal resolution (for short acoustic events like bursts in plosives). The second study aims to improve the training material for AF classifiers by using a data selection approach. Both studies compare the performance achieved on read speech to the performance achieved on conversational speech and present analyses of observed discrepancies. Section 2 describes the acoustic parameters, the classification method, and the speech material used in both studies. In Sects. 3 and 4 we present and discuss our results from Studies 1 and 2, respectively. The paper closes with a general discussion and conclusions.

### 1.1 Acoustic parameters for manner classification

Research aimed at finding acoustic parameters for accurate AF classification can be distinguished into two research lines. In one line of research, exemplified by Frankel et al. (2007b), King and Taylor (2000), Manjunath and Sreenivasa Rao (2016) and Scharenborg et al. (2007), attempts are made to cover a full set of features with a single multi-value classifier (with seven classes). In the second line, exemplified by Niyogi et al. (1999), Pruthi and Espy-Wilson (2007), and Schutte and Glass (2005), research concentrates on finding an optimal set of acoustic parameters for building a detector for one specific *manner* feature for, e.g., vowel nasalization (e.g., Pruthi and Espy-Wilson 2007), nasal manner (e.g., Chen 2000; Pruthi and Espy-Wilson 2004), or stops (e.g., Abdelatti Ali et al. 2001; Niyogi et al. 1999). These type of binary classifiers tend to use highly specific parameters and perhaps also highly specific decision mechanisms.

In our work we are interested in creating a feature set that covers all aspects of manner of articulation for the automatic transcription of speech. Hence, this study follows the approach of building a single multi-value classifier. Here, the goal is to develop acoustic parameters for accurate classification of both stationary sounds (e.g., nasals) and short acoustic events (e.g., bursts in plosives). A big challenge when using a single classifier for all manner features is that the accurate classification of some manner features requires high frequency resolution (e.g., glides), whereas others require high time resolution (e.g., the detection of bursts). In Study 1, we investigate whether the use of a combination of MFCCs calculated for both short and long windows serves this purpose and whether this set of acoustic parameters shows not only similar overall performance but also similar improvements in TIMIT and Switchboard (see Sect. 3).

### 1.2 Impact of inaccurate labeling

The development and evaluation of AF classifiers suffer from the absence of large corpora that provide manually labeled AF values. As mentioned earlier, only a small set of 78 Switchboard utterances was manually transcribed at the AF level (Livescu et al. 2007). As a consequence, training and testing of AF classifiers is generally done on the basis of data that is labeled with broad phonetic transcriptions (such as those that come with the TIMIT corpus), after which all phones are automatically replaced with their corresponding (canonical) AF values using a lookup table. As a result, these AF values change synchronously at phone boundaries, which obviously violates the observation that, at least some, AFs tend to change independently and asynchronously. For example, /n/ would map to the AF values [+voiced, alveolar, nasal] but vowels surrounding the /n/ would not be automatically mapped onto [nasal]. However, it is well known that vowels following and preceding nasal consonants tend to become nasalized in English (Ogden 2009). Thus, AF transcriptions created from phone transcriptions do not reflect the overlapping nature of articulation. What is more, the effect of this automatic mapping on AF classification performance may well be much larger in conversational speech than in read speech. Even though the reason for using AFs in the first place is to get away from phones as basic unit, phones are still used as basic unit to create training material. In the current paper, we will show that this automatic mapping causes substantial performance drops and that this effect is stronger for conversational speech.

For read speech, the impact of using synchronously changing AF labels on classification accuracy was illustrated by King and Taylor (2000). When evaluating the performance of their classifier, they showed that if features are allowed to change within ±2 frames (with a duration of 10 ms) from the phone boundary, the measure ‘all frames correct’ increases significantly by 9% absolute to 63%. Therefore, applying very strict criteria with respect to phone boundaries may introduce virtual errors that have a substantial impact on the apparent frame accuracy. Potentially incorrect labels at phone boundaries do not only have an impact when testing the classifiers but also when training them. It is unclear to what extent classifiers trained and tested on error-prone data will yield classification results that truly reflect the underlying acoustic phonetic events one is interested in. Given that in conversational speech the number of phones realized canonically is much lower than in read speech (c.f. Schuppler et al. 2014), the number of incorrect labels in training and test data is higher. For this same reason, the creation of broad phonetic transcriptions is a much more difficult task for conversational speech than for read speech, which is for instance also reflected by substantially higher inter-labeler disagreement [5.6% for read speech vs. 21.2% for conversational speech (Kipp et al. 1996, 1997)].

In order to deal with these labeling issues, Chang et al. (2005) have developed the so-called *elitist approach*, a form of data selection with which a reliably labeled subset of training material is extracted with the help of the classifier. Afterwards, the same (training) set is used for testing and only those frames that have a classification output larger than a certain threshold (0.7 in Chang et al. 2005) are selected for subsequent training and testing. Using this approach, one hopes to eliminate frames that cannot be unambiguously assigned to a particular AF value. For TIMIT, Chang et al. (2005) yielded an 8% absolute improvement in classification accuracy. Whereas Chang et al. (2005) did not test the *elitist approach* on other speaking styles than read speech, Study 2 of this paper will investigate whether such a data selection method can be used to improve AF classification for conversational speech.

## 2 Method and material

### 2.1 Support vector machines (SVMs)

For building AF classifiers, different statistical classifiers have been used, such as artificial neural networks

(ANNs) (e.g., King and Taylor 2000; Kirchhoff 1998; Scharenborg et al. 2007; Siniscalchi et al. 2008) and FFNNs (Manjunath and Sreenivasa Rao 2016), HMMs (e.g., Kirchhoff 1999; Manjunath and Sreenivasa Rao 2016), linear dynamic models (Frankel 2003), *k* nearest neighbor (*k*-NNs) (Naess et al. 2011) and dynamic Bayesian Networks (e.g., Pernkopf et al. 2009; Frankel et al. 2007b; Jyothi 2013). Niyogi et al. (1999), Pruthi and Espy-Wilson (2007), Yoon et al. (2010), Scharenborg et al. (2007), and Schutte and Glass (2005) used support vector machines (SVMs) for the classification of articulatory-acoustic features, among other reasons because these show good generalization from a small amount of high-dimensional training data. Juneja (2004) developed SVM-based landmark detectors for classifying binary place and voicing features in TIMIT, while Niyogi and Sondhi (2001) used SVMs to detect stop consonants in TIMIT. Also King et al. (2007) mentions SVMs as a powerful classification technique for binary tasks. Then, Scharenborg et al. (2007) showed that SVMs even compare favorably to ANNs on the task of multi-level articulatory–acoustic feature classification, which is also the task at hand in the current work. We therefore use SVMs for our investigations.

SVMs learn the optimal hyperplane from labeled training material by separating two classes using the *maximum margin principle* (Cortes and Vapnik 1995). The margin is defined as the distance between the hyperplane and the data points in both classes closest to the plane. SVM classifiers are solely based on the data points at the margin, so called *support vectors*. When data cannot be separated by a hyperplane, SVMs can be made to use *soft margins*, allowing for some data points to be on the wrong side of the hyperplane (Schölkopf et al. 2000). The soft-margin approach introduces a parameter $C$, which controls the trade-off between the size of the margin and the number of misclassified data points. For problems that are not linearly separable directly, SVMs can construct models by mapping the input space into a higher dimensional space in which the optimal separating hyperplane is calculated. For this purpose, we use the Radial Basis Function (RBF) kernel which has one parameter $\gamma$ that controls the width of the kernel functions.

Originally, SVMs were designed for two-class problems. However, multi-class problems can also be handled by reducing them to a set of binary problems. In *one-versus-rest*, $N$ binary classifiers are trained to separate one class from the $N-1$ other ones; in *one-versus-one*, $(N-1)N/2$ classifiers are trained, all separating one class from one other class. In our experiments, we view the AF manner classification task as a multiclass problem with seven classes and adopt the one-versus-one classification method. For training and testing the models, we use the LibSVM package (Chang and Lin 2001).The parameters C and $\gamma$ are optimized for each experiment separately using a grid search (see Sects. 3, 4).

## 2.2 Evaluation of frame-level classification

For all experiments, we will present frame-level classification accuracies in terms of percentage correctly classified frames of the test material along with the 95% confidence intervals. The 95% confidence interval reflects a (conservative) significance level of 0.05, i.e., if two confidence intervals do not overlap, the difference between two values is significant (Field 2013). Additionally, we will present the statistical measure *F-score* for each class (Powers 2011). Since the F-score considers both the precision $p$ (number of true positives divided by the sum of true positives and false positives) and the recall $r$ (number of true positives divided by the sum of true positives and false negatives), it gives a fairer picture of the actual performance of a classifier than the classification accuracy alone. The F-score is calculated as the harmonic mean of precision and recall:

$$F = 2 \times \frac{p \times r}{p + r} \tag{1}$$

## 2.3 Articulatory–acoustic feature values

Speech scientists do not agree on a unique mapping between articulatory gestures and AF values. Here we would like to go into more detail for the case of plosives. The canonical realization of a plosive consists of three stages: a closure, a burst and a subsequent release-friction. In not-canonical realizations, which are especially frequent in conversational speech (e.g., 88.5% of word-final /t/ in Dutch Schuppler et al. 2009a), one or more of these stages can be absent. The stages of plosives have been mapped in different ways onto AF values. Plosives can be mapped as a whole on one AF value ‘plosive’ (King and Taylor 2000; Scharenborg et al. 2007; Siniscalchi et al. 2008), or on a sequence of ‘closure’ and ‘friction’, due to which the burst and the release friction are modeled together with friction coming from fricative consonants, e.g. the phones /f/ and /z/ (Frankel et al. 2007a; Kirchhoff et al. 2002). This mapping is quite plausible, since the difference between a release realized with friction and a release realized as burst and friction is only salient within the first milliseconds of the release, i.e., the steepness of its amplitude rise. In order to distinguish friction from fricatives and releases from plosives that might be realized with a burst, plosives are mapped on to the sequence ‘closure’ and ‘burst+release’ (Frankel et al. 2007b; Pernkopf et al. 2009). We also opt for the latter sequence of values, for two reasons. First, modeling plosives as one unit violates the assumption of SVMs that the

**Table 1** Mapping of TIMIT phone symbols to the manner AF values

| Phone | Manner AF value |
|---|---|
| sil, pau, h# | 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 |
| ch, jh, dx, epi, t all vowels | NIL |

**Table 2** Phone-to-AF mapping in the SV-AF data

| 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 | Ficative |
| b, d, g, p, t, k, q | Fricative | Burst+release or fricative |
| bcl, dcl, gcl, pcl, tcl, kcl | Closure | Closure |
| Silence | Silence | Silence |

sequence of frames assigned to a sound can be considered as drawn from one population, which is definitely not true for plosives that consist of a sequence of ‘burst+release’. Second, we aim at using AF classifiers to analyze how the speech was actually produced, for instance whether a plosive was realized as a closure followed by friction, in which case the release friction of the plosive would be expected to be classified as ‘fricative’, or as a closure followed by a clear burst, in which case the frames would be expected to be classified as ‘burst+release’ (cf. Schuppler et al. 2009a). Therefore, we train separate classifiers for ‘fricative’ and ‘burst+release’.

In the literature on AF classification there is no consensus on whether to add the value ’vowel’ to the *manner* classifier (e.g., Scharenborg et al. 2007; Chang et al. 2005), or whether to exclude vocalic stretches of speech from *manner* classification. We opt for the latter, because from a phonetic point of view, manner (and also place) of articulation is only defined for consonants. A full overview of the *manner* features used in this research is given in Table 1. Note that affricates were excluded from our experiments. Affricates consist of a sequence of a plosive and a fricative; however, the boundary between these two acoustic events was not provided in the TIMIT transcriptions.

## 2.4 Read speech corpus: TIMIT

TIMIT contains phonetically balanced sentences read by 630 speakers of American English (Garofolo 1988). We followed TIMIT’s training (3696 utterances) and test division (1344). The TIMIT database comes with manual phone level transcriptions, which we have automatically relabeled in terms of AF values according to Table 1.

## 2.5 Conversational speech corpus: switchboard

Switchboard is a corpus of telephone bandwidth speech from spontaneous conversations from 500 speakers of American English (Godfrey et al. 1992).

### *2.5.1 SVitchboard-AF (SV-AF)*

SVitchboard-AF (SV-AF) consists of 78 utterances (a total of 119s of speech, excluding silences) (Livescu et al. 2007) taken from the Switchboard corpus. SV-AF comes with manually created transcriptions at the phone and AF level. Since we use SV-AF as a gold standard, it is important to note that the inter-transcriber agreement was very high for the AF labels (kappa-values between 0.82 and 0.95), which is approximately 0.1 higher than for the phone-label transcriptions of Switchboard.

We manually adapted the original set of AF labels from Livescu et al. (2007) to our set of AF labels using the mapping shown in Table 2, starting from the tier ‘Dg1’ (Degree of forward constriction). In the original transcriptions, both fricatives and plosive—release sequences were transcribed as ‘fricative’. We changed the label of releases of plosives that started with a burst to ‘burst+release’. The boundaries of the ‘nasality’ tier of the original transcription were used to annotate the nasal consonants. Apparent labeling mistakes that were observed in two utterances were corrected. When modifying the annotations, none of the original boundaries was moved or deleted, although new boundaries were placed when one value in the ‘Dg1’ set is transcribed as a sequence of two values in our set. For example, an ‘approximant’ in the original set occasionally was replaced by a ‘liquid’ followed by a ‘glide’ in our set. Additionally, since background noise was labeled as silence, new boundaries were placed to separate background speakers from silence, because often their speech was of similar amplitude as the foreground speaker. The resulting transcriptions are referred to as manual-SV-AF in the remainder of the paper.

In order to be able to make direct comparisons between the Switchboard and TIMIT results, the AF labels (and boundaries) from the 78 manual-SV-AF utterances were also generated automatically, by automatically changing the phone labels into their AF values using a table lookup procedure. Note that this is the ‘standard‘ procedure in AF research. These transcriptions are referred to as automatic-SV-AF.

*2.5.2 Switchboard Transcription Project (STP)*

The Switchboard Transcription Project (STP) (Greenberg 1997) contains 72 min of speech from the Switchboard corpus (taken from 618 conversations by 370 different speakers) that were manually transcribed at the phone level. There is no overlap between STP and SV-AF. The STP labels are related to the phone set used for TIMIT, but in STP, plosives are annotated as one segment and not as a sequence of closure and burst+release (e.g., /pcl/ and /p/ in TIMIT map to /p/ in STP). The STP plosives were therefore automatically split into closure and burst+release classes using an automatic procedure (Schuppler 2011). An independent test of this splitting procedure on the plosive segments in the SV-AF corpus showed a 63% agreement with the manually place boundaries. This agreement is in the range of what has been reported in the literature. Khasanova et al. (2009), for instance, reported that automatically and manually created burst labels coincided in 47% or the plosives in word-medial and in 97% of the plosives in word-initial position).

## 3 Study 1: improving acoustic parameters for AF classification

### 3.1 Four sets of acoustic parameters

Previous research investigated different methods to parameterize the acoustic waveforms and different window lengths, and shifts for the detection of specific acoustic events. For multi-value AF classification tasks, however, mostly MFCCs have been used (e.g., Chang et al. 2005; King and Taylor 2000, but see Scharenborg and Cooke 2008) for a comparison of different acoustic parameters for AF classification). Good results are obtained with the conventional 25 ms window shifted with 10 ms for fairly stationary features. In order to accurately detect short acoustic events such as bursts in plosives, however, shorter window lengths and shifts are needed. For instance, Salomon et al. (2004) used 5 ms windows shifted with 1ms steps. A 2.5 ms step size rather than the conventional 10 ms has also shown good results in combination with long 25 ms windows for the detection of nasals (Pruthi and Espy-Wilson 2004).

Our goal is to capture both very short (e.g., bursts) and longer acoustic events (e.g., nasality) both in read speech (TIMIT) and in conversational speech (STP). To that end, we investigate MFCCs derived using two different window lengths and shifts and their combinations:

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

For each of these types of acoustic parameters, the input speech is first divided into overlapping Hamming windows of 25 or 5 ms with a 10 or 2.5 ms shift and a pre-emphasis factor of 0.97. For the 25 ms windows, a filter bank of 22 triangular filters equally spaced on the Mel-scale was used to calculate 13 MFCCs (C0–C12) and their first and second order derivatives (39 parameters). For the 5 ms windows, a filter bank of seven triangular filters was used and seven MFCCs (C0–C6) and their first and second order derivatives were calculated (21 parameters). Cepstral mean subtraction (CMS) was applied to all parameters.

The SVM classifiers use a temporal context of 30 ms at both sides of the frame to be classified. For *Baseline*, three frames (30 ms) to the left and right of each frame were concatenated, resulting in MFCC vectors of length 7×39 = 273. For the *Short*, *Long*, and *Both* classifiers also three frames were concatenated, but taking only every fourth frame, in order to cover the same temporal context as in *Baseline*. This resulted in feature vectors of length 273 for *Long* and 147 for *Short*. For *Both*, feature vectors of long and short windows with the same midpoint were concatenated, resulting in feature vectors of length 273 + 147 = 420.

### 3.2 AF classification of TIMIT

*3.2.1 Results*

For the optimization of the C and $\gamma$ parameters, two independent subsets of 5000 feature vectors (one for training and one for testing) were randomly extracted from the original TIMIT training set. An additional 100K vectors were extracted from randomly chosen files from the TIMIT training set for training the SVM classifiers with the *Baseline* parameters. For the *Short*, *Long*, and *Both* parameters, the same audio data were used, resulting in 400K vectors (the shift is four times smaller). The resulting classifiers were tested on 294,984 10 ms frames and 1,173,665 2.5 ms frames from the TIMIT test set.

Table 3 shows the AF classification accuracy in terms of percentage correctly classified frames on the TIMIT test material. The diagonals additionally show the 95% confidence intervals. F-scores are calculated as the harmonic mean of precision and recall and shown for each class.

The average frame level accuracies are: 83.4% for *Baseline*, 85.3% for *Short*, 87.0% for *Long*, and 87.7% for *Both*. Comparing our three sets of acoustic parameters with the baseline shows that the *Both* classifier performed

**Table 3** Frame-level confusion matrices for the AF classifiers trained on TIMIT and tested on TIMIT

| | Sil | Liq | Gli | Nas | Fric | Bur | Clo |
|---|---|---|---|---|---|---|---|
| BL | | | | | | | |
| Sil | **93.2** ± **0.2** | 0.3 | 0.2 | 0.8 | 3.0 | 0.6 | 2.0 |
| Liq | 0.4 | **89.0** ± **0.3** | 2.4 | 2.8 | 2.8 | 1.0 | 1.4 |
| Gli | 0.8 | 12.8 | **77.0** ± **0.7** | 2.8 | 3.3 | 1.3 | 2.0 |
| Nas | 1.6 | 2.4 | 0.6 | **86.3** ± **0.4** | 4.2 | 0.5 | 4.4 |
| Fric | 2.7 | 1.1 | 0.4 | 1.6 | **89.5** ± **0.2** | 1.6 | 3.1 |
| Bur | 2.8 | 2.9 | 0.6 | 1.9 | 12.8 | **65.2** ± **0.6** | 13.8 |
| Clo | 4.3 | 0.9 | 0.3 | 3.5 | 4.9 | 2.4 | **83.6** ± **0.3** |
| F-score | **0.93** | **0.88** | **0.81** | **0.84** | **0.88** | **0.73** | **0.83** |
| Short | | | | | | | |
| Sil | **92.5** ± **0.1** | 0.2 | 0.1 | 0.9 | 3.6 | 0.6 | 2.1 |
| Liq | 0.6 | **89.1** ± **0.2** | 2.8 | 3.7 | 2.1 | 1.0 | 0.8 |
| Gli | 0.6 | 13.8 | **78.1** ± **0.4** | 3.2 | 2.1 | 1.4 | 0.8 |
| Nas | 1.7 | 2.7 | 0.8 | **87.6** ± **0.2** | 3.2 | 0.4 | 3.5 |
| Fric | 3.0 | 0.8 | 0.4 | 2.0 | **88.5** ± **0.1** | 2.5 | 2.9 |
| Bur | 2.5 | 1.5 | 0.5 | 0.8 | 11.6 | **76.4** ± **0.3** | 2.9 |
| Clo | 4.9 | 0.6 | 0.2 | 3.2 | 4.1 | 2.1 | **84.8** ± **0.2** |
| F-score | **0.92** | **0.89** | **0.81** | **0.85** | **0.88** | **0.80** | **0.85** |
| Long | | | | | | | |
| Sil | **93.3** ± **0.1** | 0.2 | 0.1 | 0.7 | 2.9 | 0.7 | 2.1 |
| Liq | 0.5 | **90.3** ± **0.2** | 2.9 | 2.2 | 2.2 | 1.0 | 0.9 |
| Gli | 0.8 | 9.8 | **83.1** ± **0.3** | 2.0 | 2.2 | 1.2 | 1.0 |
| Nas | 1.7 | 1.8 | 0.7 | **89.0** ± **0.2** | 2.5 | 0.4 | 3.8 |
| Fric | 2.4 | 0.8 | 0.4 | 1.3 | **90.1** ± **0.1** | 2.3 | 2.7 |
| Bur | 2.4 | 1.4 | 0.5 | 0.5 | 10.3 | **78.0** ± **0.3** | 7.0 |
| Clo | 4.3 | 0.7 | 0.2 | 3.0 | 4.0 | 2.7 | **85.1** ± **0.2** |
| F-score | **0.93** | **0.90** | **0.85** | **0.88** | **0.90** | **0.81** | **0.85** |
| Both | | | | | | | |
| Sil | **93.5** ± **0.1** | 0.2 | 0.1 | 0.7 | 2.6 | 0.7 | 2.2 |
| Liq | 0.5 | **91.0** ± **0.2** | 2.5 | 2.0 | 2.0 | 1.0 | 0.9 |
| Gli | 0.6 | 9.3 | **84.2** ± **0.3** | 1.8 | 1.9 | 1.3 | 0.9 |
| Nas | 1.6 | 1.7 | 0.7 | **89.4** ± **0.2** | 2.3 | 0.3 | 3.9 |
| Fric | 2.3 | 0.7 | 0.3 | 1.2 | **90.8** ± **0.1** | 2.2 | 2.5 |
| Bur | 2.0 | 1.3 | 0.5 | 0.6 | 8.9 | **79.6** ± **0.2** | 7.0 |
| Clo | 4.1 | 0.6 | 0.2 | 2.9 | 3.8 | 2.8 | **85.6** ± **0.2** |
| F-score | **0.93** | **0.91** | **0.86** | **0.88** | **0.90** | **0.82** | **0.86** |

Frame level accuracies for each AF are given in bold

best for ‘burst+release’ (Bur): the F-score increased from 0.73 for *Baseline* to 0.82 for *Both*. This was to be expected, since bursts are events of very short duration. The *Short* and *Both* classifiers perform best for ‘fricative’ (Fric). Most importantly, the *Both* classifier seems to be able to combine the classification power of the *Short* and *Long* classifiers, its F-score increased from 0.84 for the *Baseline* acoustic set to 0.88, an overall improvement in F-score of 0.04. The *Both* acoustic parameter set thus seems to be best able to capture both very short and longer acoustic events.

### *3.2.2 Discussion*

Comparing the classification accuracies with those reported in the literature is not straightforward since, as indicated in Sect. 2.3, different authors used slightly different sets of *manner* features. For example, Scharenborg et al. (2007) used the feature values ‘vowel’, which we did not consider, and they modeled ‘plosives’ as a whole. Including the performance for ‘vowel’ classification, Scharenborg et al. (2007) obtained an average accuracy of 84%. In order to compare our results with those of Scharenborg et al. (2007), we re-analyzed the data of that study and calculated average classification accuracies over all other manner values, which gave accuracies between 66.8 and 75.6% for the different classification methods they used. Our classification accuracy of 87.7% thus outperforms Scharenborg et al. (2007).

Chang et al. (2005) used a similar set of *manner* values as Scharenborg et al. (2007), but they did not study 'approximant' and 'retroflex' consonants. Their baseline system is different from ours: Classifiers are trained on NTIMIT (Jankowski et al. 1990), which has telephone bandwith quality, and the acoustic frames have a length of 25 ms and are shifted in 10 ms steps. Their average classification accuracies for the *manner* values are lower than ours: 70% excluding 'vowel', 75% including 'vowel' (Chang et al. 2005). One reason for their lower performance accuracy might also be the lower quality of the recordings used (NTIMIT vs. TIMIT).

Salomon et al. (2004) developed a set of temporal measures for increasing the time resolution for manner classification. They distinguished the values 'sonorant', 'fricative', 'stop', and 'silence'. They observed that the use of only the temporal measures yielded the same overall accuracy as the use of MFCCs (70.1%, using 20 ms windows and 5 ms shift for all parameters). For the combination of MFCCs and the temporal measures they reported an accuracy of 74.8%, which is lower than our accuracy of 87.7% for *Both*. The relative improvement (4.7%) they report is comparable to the improvement we obtained (4.3% in accuracy or 0.04 in F-scores). In conclusion, our combination of acoustic parameters derived from short an long windows show satisfactory results for TIMIT, both in comparison with our baseline parameters and with previous results for multi-value classification experiments from the literature.

## 3.3 AF classification of Switchboard

### *3.3.1 Results*

We subsequently evaluated the best performing acoustic parameter sets on conversational speech. Following the procedure for read speech, we trained classifiers with the acoustic parameter set (*Baseline* and *Both*) on the complete STP material. Classifiers were trained using 50K frames with the *Baseline* feature (10 ms shift) and the corresponding 200K frames for *Both* (2.5 ms shift). In order to compare the results with those obtained on TIMIT, the classifiers were tested on automatically created AF labeled material.

Table 4 shows the frame-level classification accuracies and the performances in terms of F-scores obtained on Switchboard. Our results show that for conversational speech, our set of acoustic parameters *Both* did not yield an improvement in comparison to *Baseline* (i.e., F = 0.65 vs. F = 0.66). Thus, whereas on TIMIT the F-score improvement was 0.04, for Switchboard there was no improvement at all. Moreover, the additional temporal information did not yield the rise in performance for the short events which was observed for TIMIT. The F-score improvement in performance for 'burst+release' was 0.04 for Switchboard against 0.09 for TIMIT. Apparently, improvements

**Table 4** Frame-level confusion matrices for the AF classifiers trained on the conversational speech of STP and tested on the automatically derived AF labels from SV-AF

| | Sil | Liq | Gli | Nas | Fric | Bur | Clo |
|---|---|---|---|---|---|---|---|
| BL | | | | | | | |
| Sil | **91.3 ± 0.6** | 0.4 | 0.2 | 1.0 | 5.0 | 0.7 | 1.5 |
| Liq | 6.9 | **73.0 ± 3.4** | 6.9 | 6.3 | 5.1 | 0.2 | 1.6 |
| Gli | 2.8 | 16.8 | **64.0 ± 3.8** | 10.2 | 5.2 | 0.7 | 0.2 |
| Nas | 6.9 | 6.1 | 4.4 | **77.7 ± 2.8** | 4.5 | 0.1 | 0.4 |
| Fric | 13.7 | 1.7 | 1.0 | 7.2 | **68.4 ± 2.2** | 4.9 | 3.1 |
| Bur | 25.6 | 2.2 | 1.6 | 3.5 | 25.0 | **34.8 ± 4.5** | 0.7 |
| Clo | 16.2 | 0.6 | 0.8 | 8.4 | 21.0 | 2.8 | **50.3 ± 2.9** |
| F-score | **0.91** | **0.70** | **0.69** | **0.68** | **0.63** | **0.41** | **0.60** |
| Both | | | | | | | |
| Sil | **87.4 ± 0.3** | 1.2 | 0.9 | 2.1 | 5.2 | 1.1 | 2.1 |
| Liq | 8.8 | **77.9 ± 1.8** | 4.4 | 3.5 | 3.5 | 0.4 | 1.5 |
| Gli | 2.3 | 12.6 | **67.1 ± 2.3** | 11.1 | 6.2 | 0.8 | 0.0 |
| Nas | 7.9 | 6.8 | 3.6 | **74.7 ± 1.5** | 5.0 | 1.1 | 1.0 |
| Fric | 13.4 | 1.1 | 1.4 | 5.4 | **70.9 ± 1.2** | 4.4 | 3.3 |
| Bur | 28.9 | 2.0 | 1.0 | 3.4 | 19.7 | **36.2 ± 1.8** | 8.9 |
| Clo | 18.9 | 1.3 | 0.6 | 7.0 | 18.8 | 6.3 | **47.1 ± 1.5** |
| F-score | **0.89** | **0.71** | **0.69** | **0.66** | **0.65** | **0.37** | **0.55** |

Frame level accuracies for each AF are given in bold obtained for read speech do not yield similar improvements in conversational speech.

### 3.3.2 *Discussion*

In general, we observed that classification performance for Switchboard is substantially worse (F = 0.66) than for TIMIT (F = 0.88). This is in line with earlier results by Pruthi and Espy-Wilson (2007) for detecting vowel nasalization. For SVMs with RBF kernels they report chance-normalized accuracies of 77.9% for TIMIT and 69.6% for Switchboard. The best results on Switchboard were achieved by Frankel et al. (2007a), who parameterized the acoustic waveforms with 12 PLP cepstra plus energy every 10 ms with 25 ms Hamming windows and trained MLP classifiers. For the classification of *degree* (having the values 'silence', 'vowel', 'fricative', 'closure', 'approximant', and 'flap'), they report a frame-level accuracy of 77.8% (34.3% chance), which is higher than our results for conversational speech (F-score = 0.66). Our results compare favorably to those recently reported by Manjunath and Sreenivasa Rao (2016) for conversational speech (56.25% accuracy with HMMs and 65.65% with FFNNs).

It is not surprising that in the present experiment, AF classification is overall worse for conversational speech than for read speech. Conversational speech shows more variability than read speech, because, among other things, words can be strongly reduced [e.g., see for American English (Johnson 2004)] and articulatory gestures may heavily overlap. This may result in more labeling errors introduced with the canonical mapping from phone labels to AF labels for conversational speech than for read speech. This hypothesis is investigated in the following subsection. It is, however, surprising that the improvements of our set of acoustic parameters obtained for read speech do not result in similar improvements for conversational speech. This result is one of the main findings of this paper: Not only the overall classification performance in conversational speech is lower, but also the relative improvement.

## 3.4 Impact of inaccurate labeling

In order to estimate the impact of the accuracy of the labeling of the frames on the performance scores on the test set, the classifiers trained on the conversational speech of STP from the previous subsection were tested on the 53,115 frames of manual-SV-AF, which supposedly contains more accurate AF labels since these labels are created manually and do not change synchronously at phone boundaries. The results show a substantial improvement of the *Both* classifier compared to the *Baseline* classifier (F-scores: 0.60 for Baseline vs. 0.65 for *Both*); Thus, whereas when tested on the automatically generated labels of STP the improvement from *Both* over *Baseline* was only 0.01, on the manually created labels we reach an improvement of 0.05. The overall performance, however, still is the same as on the automatically created labels (F = 0.65).

In order to estimate the impact of labeling accuracy in the training set, we estimated the number of erroneous labels in the training set on the basis of the disagreement in frame labels between automatic-SV-AF and manual-SV-AF. We computed the number of speech samples where the labeling in automatic-SV-AF differed from that of manual-SV-AF. Overall, we observed a disagreement for 19.9% of the frames. Of these, 29.4% of the samples carrying the label 'liquid' (Liq) in the automatic-SV-AF set did not contain a liquid according to the human labelers. Extrapolating these results suggests that a substantial part of the labels used for training do not actually represent the putative acoustic feature. The question that arises is whether AF classifiers trained on better labeled frames will yield improved results for conversational speech. This is investigated in the next section (Table 5).

## 3.5 Impact of sound quality

We were aware of the double mismatch between TIMIT and Switchboard: TIMIT is read speech, while Switchboard is conversational speech; and TIMIT was recorded in noise-free condition, whereas Switchboard was recorded over the landline telephone network. In order to be able to estimate, whether the decrease in classification performance in Switchboard is really mainly due to the speaking style, and not due to the lower sound condition, we performed a classification experiment with the *Both* AF classifier trained on TIMIT and tested on manual–SV-AF. The results (c.f. Table 6) show that both the F-scores and the overall accuracies of this classifier are much lower than in the matched condition (i.e., tested on on TIMIT 85.1 vs. 52.2%).

Part of the decrease in classification performance may be due to mismatch in the recording conditions. To investigate this hypothesis we analyzed the distributions of the MFCC coefficients C0 and C1. Figure 1 show the distributions calculated from 20,772 frames for each set. It can be seen that cepstral mean normalization does not compensate for all differences in recording conditions. The distance between the first and second hump in the C0 distribution illustrates that there is a substantial difference in signal to noise ratio (SNR) between the TIMIT and the SV-AF set. Also, the difference in shape of the C1 distribution suggests that there remains a considerable mismatch between the range of spectral slopes in the different data sets. Another indicator for the impact of the recording conditions on the classification accuracy is that silence and closures are more often (incorrectly) classified as fricative in the manual-SV-AF

**Table 5** Frame-level confusion matrices for the AF classifiers trained on the conversational speech of STP and tested on the manual AF labels of SV-AF

| | Sil | Liq | Gli | Nas | Fric | Bur | Clo |
|---|---|---|---|---|---|---|---|
| BL | | | | | | | |
| Sil | **92.3 ± 0.6** | 0.5 | 0.2 | 1.2 | 4.3 | 0.5 | 1.1 |
| Liq | 5.1 | **73.1 ± 3.5** | 5.9 | 6.7 | 6.8 | 0.2 | 2.2 |
| Gli | 5.8 | 28.3 | **45.6 ± 4.0** | 12.7 | 5.6 | 1.3 | 0.7 |
| Nas | 6.8 | 5.1 | 5.4 | **73.5 ± 3.0** | 7.9 | 0.6 | 0.8 |
| Fric | 20.6 | 1.7 | 1.4 | 7.3 | **61.9 ± 2.3** | 3.8 | 3.3 |
| Bur | 11.4 | 1.9 | 0.2 | 4.4 | 35.0 | **36.4 ± 4.6** | 10.7 |
| Clo | 16.4 | 3.3 | 1.3 | 1.2 | 26.1 | 5.7 | **35.2 ± 2.8** |
| F-score | **0.91** | **0.65** | **0.54** | **0.63** | **0.58** | **0.41** | **0.46** |
| Both | | | | | | | |
| Sil | **91.0 ± 0.3** | 0.5 | 0.1 | 1.4 | 4.0 | 0.9 | 2.0 |
| Liq | 8.4 | **75.9 ± 1.8** | 3.3 | 3.9 | 5.0 | 1.1 | 2.3 |
| Gli | 0.6 | 23.5 | **50.7 ± 2.3** | 9.9 | 7.1 | 1.4 | 1.4 |
| Nas | 9.7 | 6.7 | 3.6 | **73.3 ± 1.5** | 4.8 | 1.1 | 0.8 |
| Fric | 19.6 | 1.3 | 1.2 | 5.0 | **60.4 ± 1.2** | 7.3 | 5.2 |
| Bur | 12.2 | 1.3 | 0.1 | 0.9 | 23.1 | **55.4 ± 1.8** | 7.0 |
| Clo | 16.6 | 3.1 | 0.8 | 9.9 | 17.6 | 6.9 | **45.1 ± 1.5** |
| F-score | **0.91** | **0.67** | **0.60** | **0.68** | **0.57** | **0.57** | **0.52** |

Frame level accuracies for each AF are given in bold

**Table 6** Confusion matrices for the AF classifier trained on TIMIT and tested on manual–SV-AF

| Both | Sil | Liq | Gli | Nas | Fric | Bur | Clo |
|---|---|---|---|---|---|---|---|
| Sil | **53.0 ± 0.6** | 0.7 | 0.1 | 1.6 | 37.9 | 2.6 | 4.1 |
| Liq | 0.6 | **67.7 ± 1.8** | 4.7 | 8.6 | 13.3 | 3.6 | 1.5 |
| Gli | 1.4 | 39.0 | **23.0 ± 1.7** | 13.6 | 14.7 | 4.5 | 3.8 |
| Nas | 2.1 | 11.2 | 3.0 | **52.7 ± 1.7** | 26.4 | 1.5 | 3.0 |
| Fric | 4.3 | 2.7 | 0.5 | 4.9 | **73.8 ± 1.1** | 12.2 | 1.7 |
| Bur | 1.2 | 2.3 | 0.2 | 1.9 | 36.3 | **55.2 ± 2.4** | 3.0 |
| Clo | 10.3 | 4.1 | 0.3 | 9.7 | 41.2 | 3.3 | **31.1 ± 1.4** |
| F-score | **0.68** | **0.54** | **0.33** | **0.52** | **0.32** | **0.41** | **0.37** |

Frame level accuracies for each AF are given in bold

test material than in the TIMIT test. One can observe that there is a much stronger bias towards classifying frames as ‘fricative’ in manual-SV-PF than in the TIMIT material: the F-score for ‘fricative’ is much lower than one might expect from the classification accuracy.

From Tables 4 and 5 it can be seen that the frame-level classification results for a classifier trained on STP yields an F-score of 0.65 in the matched case (i.e., when tested on automatic-SV-AF) and the mismatched-case (i.e., manual–SV-AF). The TIMIT classifier tested in the matched case (i.e., on TIMIT) yielded a much higher F-score of 0.88. However, classifiers trained on TIMIT and tested on manual-SV-APF (the mismatched case) yielded a much lower F-score of 0.64, which is however only slightly lower than the STP-trained classifiers. There are two ways in which one can look at these numbers. On the one hand, we see a large detrimental effect of the mismatch between recording conditions and speaking style between training and test. But at the same time we also see an equally large difference between matched results on TIMIT and on Switchboard. Although the acoustic conditions in Switchboard are likely to affect the classification performance somewhat, we still think that speech style is much more important.

## 4 Study 2: the elitist approach for AF classification

Chang et al. (2005) proposed the so-called elitist approach as a solution for dealing with mislabeled frames in the training data. In this approach, initial models are trained on the complete training set and each frame is assigned a probability for being correctly classified.

For training the final model, only those frames are selected whose probability for correct classification is above a predefined threshold. We follow their procedure.

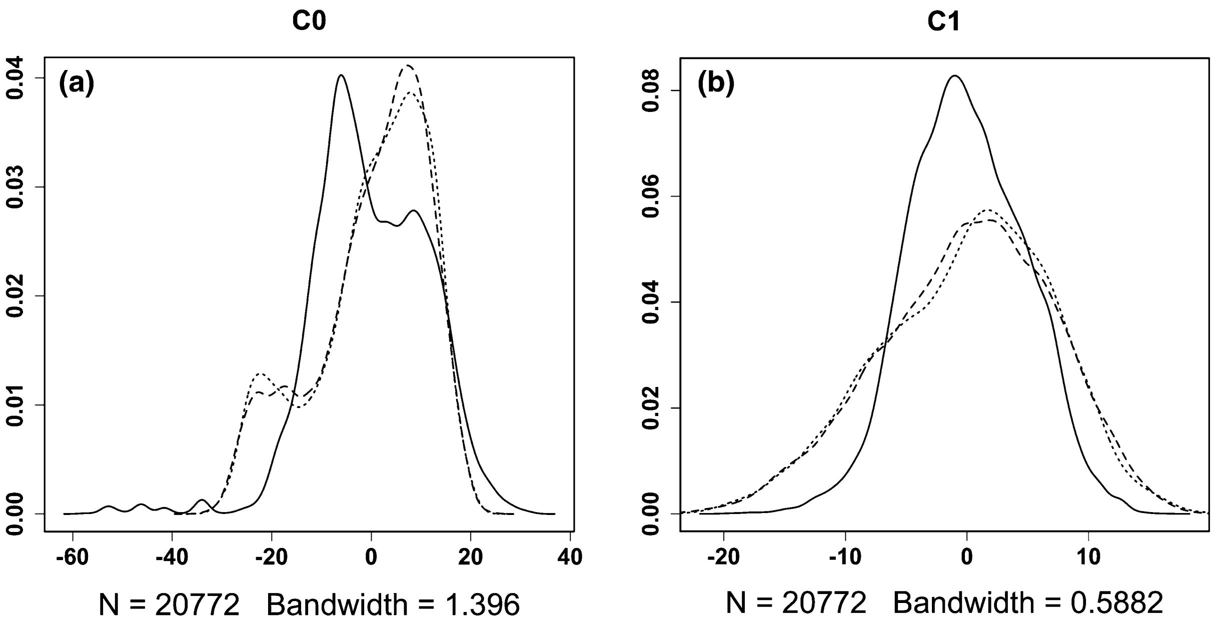

**Fig. 1** Distribution of the C0-coefficient (**a**) and the C1-coefficient (**b**) for the TIMIT test set (*dashed line*) the TIMIT train set (*dotted line*) and manual-SV-AF (*solid line*)

First an SVM classifier with the *Both* acoustic parameters is trained on the 200K frames of the STP data set. Subsequently, this classifier is used to predict the posterior probabilities for each frame of the STP training material. Finally, classifiers are trained only on those frames where the probability of the winning class is larger than a certain threshold. We compare the classification performance of five different threshold settings: 0.95, 0.90, 0.70, 0.50, and 0.00 (original training set).

### 4.1 Results

When comparing all different threshold settings (see Table 7), the highest average overall accuracy (64.5%) is obtained with the original training set. Thus, there is no increase in overall classification accuracy when training the classifiers using a subset of supposedly better labeled frames. For individual AF values, however, improvements over the baseline setting can be observed. Here, different threshold settings are optimal. Whereas ‘silence’, ‘glide’, ‘nasal’, and ‘burs’t do not profit from training on a selection of the best labeled frames, ‘liquid’, ‘fricative’, and ‘closure’ do profit from data selection. For the feature values of the first group it holds that they are likely to spread, so that including more frames may be profitable. For the second group it may be the other way around, also because these are often short-lived.

### 4.2 Discussion

Our results on conversational speech do not replicate the findings of Chang et al. (2005), who achieved an 8% absolute improvement in classification accuracy on read speech(NTIMIT) when training on only the best labeled frames. In comparing these experiments, the recording quality is not a possible source for the difference in gain due to data selection, since both NTIMIT and Switchboard is telephone speech. Although in Chang et al. (2005) a slightly different AF set was used, we think that the difference in performance is mainly due to the difference in

**Table 7** Classification results for the elitist approach: frame-level class-dependent F-scores and overall accuracy (Acc.) for the AF classifiers trained on STP and tested on SV-AF

| Threshold | Sil | Liq | Gli | Nas | Fri | Bur | Clo | Acc. | #SV |
|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 0.91 | 0.67 | **0.60** | **0.68** | 0.57 | **0.57** | 0.52 | 64.5 | 105,325 |
| 0.50 | 0.91 | **0.68** | 0.60 | 0.66 | 0.59 | 0.49 | 0.53 | 64.1 | 90,646 |
| 0.70 | 0.90 | 0.66 | 0.59 | 0.66 | 0.59 | 0.49 | 0.53 | 63.9 | 84,912 |
| 0.90 | 0.91 | 0.67 | 0.60 | 0.65 | **0.59** | 0.48 | 0.53 | 64.1 | 71,034 |
| 0.95 | **0.91** | 0.66 | 0.60 | 0.64 | 0.58 | 0.47 | **0.54** | 64.2 | 51,161 |

Frame level accuracies for each AF are given in bold speech style, i.e., read speech (NTIMIT) versus conversational speech (Switchboard). In conversational speech, there are less frames than in carefully read speech that can be unambiguously assigned to one particular AF value, so training models on representative frames improves classification accuracy less in conversational speech than in read speech. As a result there may be too few unambiguous frames left in conversational speech (due to highly overlapping sounds and reduced segments) for training a reliable classifier. In addition, Chang et al. (2005) applied data selection also on the test set; thus, they optimized not only the frames on which classifiers were trained, but they also selected 'representative' frames for the evaluation of the classifiers. Selecting an optimal subset for testing the classifiers certainly raises the classification accuracy; however, it is questionable whether these classifiers also *perform* better.

From the rightmost column in Table 7 it can be seen that the number of support vectors decreases as the selection of the training samples becomes more strict. Thus, selection does have an effect, however this seems to surface in the form of more compact classifiers rather than in higher classification accuracy. This finding suggests that the classification task in the conversational speech in Switchboard is inherently very difficult. Finally, we need to mention another possible source for the lack of improvement due to the data selection approach: by filtering the training set, the number of frames available for training is decreased. Summing up, even though data selection showed to be a well working approach on the large read speech data from TIMIT, it does not yield similar improvements on the conversational Switchboard data.

## 5 General discussion

The main aim of this paper was to show that it can not be assumed that methods yielding improved classification results on read speech also yield a similarly high improvement in spontaneous, conversational speech. Whereas most previous studies assumed that the classification performance in conversational speech would simply be overall lower than in read speech, we showed on the basis of two studies that it is not only lower, but that improvements reached on read speech do not transfer to other speaking styles. The aim of the first study was to find a set of acoustic parameters with a high time and a high frequency resolution. Classifiers with different sets of acoustic parameters were tested on read (TIMIT) and conversational speech (Switchboard). In both cases, the AF labels were obtained through an automatic mapping from phone to AF labels. The results showed that combining MFCCs derived from a long window of 25 ms and from a short window of 5 ms both shifted with 2.5 ms steps (F = 0.88) outperformed a baseline system where the MFCCs were derived from a window of 25 ms shifted with 10 ms (F = 0.84) for read speech. For conversational speech, however, the overall performance dropped to F = 0.66 for the *Baseline* system and, importantly, there was no gain in performance for our acoustic parameters (*Both*: F = 0.65) over *Baseline*.

The aim of the second study was to test whether a data selection approach for creating the training material is an equally powerful method with conversational speech as with read speech material. Previous work using this *elitist* approach by Chang et al. (2005) showed a performance improvement of 8% for read speech (NTIMIT). Our results on conversational speech (Switchboard), however, did not show such an improvement in overall classification performance when only the best-labeled frames were selected for training; only a small improvement of 2% was found for fricatives and closures and of 1% for liquids. In order to interpret these results, we need to look at the distribution of the classification confidence of the frames within a segment. In read speech, the highest confidence is in the middle of the phone and the lowest confidence is at the boundaries (Schuppler et al. 2009b). Thus, with the elitist approach mainly frames close to the phone-boundaries are removed from the training material. In conversational speech, however, where segment durations are much shorter and the overlap of features is higher, frames are also removed from the center of the phone. The question arises whether in conversational speech, unambiguous frames, i.e., frames with a single *manner* value, exist at all and whether it is at all possible to train classifiers for a single *manner* value using conversational speech as training material. In the future, one might train classifiers on read speech (where frames of well defined class membership exist) and test these classifiers on conversational speech. In doing so, however, test and training material will not be matching and thus again not result in a classification improvement.

Performance drops when going from read speech (TIMIT) to conversational speech (Switchboard) have also been reported for ASR, where word accuracies for TIMIT typically exceed 95%, while for Switchboard they tend to be in the 50–70% range (Godfrey et al. 1992). This difficulty is also reflected in the inter-human labeling disagreement of phonetic transcriptions (5.6% for read speech vs. 21.2% for conversational speech; Kipp et al. 1996, 1997). Hence, it is not surprising that our classification performance is overall worse for conversational speech than for read speech. Whereas such overall performance drops from read to conversational speech have previously been shown (e.g., Manjunath and Sreenivasa Rao 2016), it has not previously been shown that relative performance improvements (here: due to our set of acoustic parameters and due to the elitist approach) does not transfer from read to conversational speech.

We suggest that one reason for this lack of transfer is that a segmentation in terms of phones, which is the basis for the automatically created AF labels, is not equally suitable for the two speech styles. As observed above, due to the high pronunciation variability in conversational speech (e.g., Johnson 2004; Kohler 2001), segmenting speech in terms of phones is extremely difficult. Therefore, the accuracy of the phonetic segmentation of read speech is surely higher than that of conversational speech. Consequently, the canonical mapping from phone labels to AF labels may still result in relatively good training material for read speech, while it does not for conversational speech. These problems with phone to AF label mappings are especially apparent for features that are inherently difficult to define. For example, confusions of glides and liquids are much more frequent in conversational than in read speech (22.8 vs. 8.6%). An explanation may be that in American English word final /l/ tends to be velarized, making the second formant similar to that of /w/, which we label a glide (Espy-Wilson 1992). Thus, some confusions are not due to low performance of the classifier, but rather and more fundamentally to inextricable overlap between the manner features in conversational speech and the resulting errors that are made when automatically converting transcription symbols to AF values. More insights will be found not until a much larger spontaneous speech corpus with reliable AF transcriptions becomes available (i.e., currently only 78 utterances from Switchboard are transcribed manually on the AF level).

## 6 Conclusions

To sum up, the work presented in this paper has extended previous research on AF classification with an emphasis on conversational speech. We have presented a set of acoustic parameters with a high frequency and time resolution, which reached an improvement of the performance of 4% of the classifiers when tested and trained on TIMIT. On the conversational telephone speech in Switchboard, the new set of acoustic parameters only yielded an improvement compared to the baseline parameters when using manually labeled testing material. In general, in all Switchboard experiments, we have observed lower performances than in TIMIT experiments. Therefore, it appears that the overlap between the acoustic parameters corresponding to AFs in spontaneous speech is much larger than in carefully read speech. This acoustic overlap is due to a larger degree of articulatory variability in spontaneous speech which is not captured by the phone-level segmentations, which were the starting point for the experiment.

In speech science, methods are mostly developed and improved using read speech corpora (e.g., TIMIT) and only afterwards they are adapted to conversational speech. Similarly, in image recognition methods are mostly developed and improved on some standard databases (e.g., MNIST) without having the application characteristics in mind from the beginning. Our studies suggest that the nature of the application data needs to be taken into account already when defining the concepts (here: starting point of a segmentation in terms of phones) and the basic assumptions of a method. Applying concepts and methods that were designed for a different speech style to the application data may fail due to the inherent differences between read and conversational speech.

**Acknowledgements** Open access funding provided by Graz University of Technology.

## References

Abdelatti Ali, A. M., Van der Spiegel, J., & Mueller, P. (2001). Acoustic-phonetic features for the automatic classification of stop consonants. *IEEE Transactions on Audio, Speech and Language Processing*, *9*(8), 833–841.

Baum, M. F. (2003). *Improving automatic speech recognition for pluricentric languages exemplified on varieties of German*. PhD thesis, Graz University of Technology, SPSC Laboratory.

Bitar, N. N., & Espy-Wilson, C. Y. (1996). Knowledge-based parameters for HMM speech recognition. In *Proceedings of ICASSP*, pp. 29–32.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines. Retrieved May 28, 2010, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chang, S., Wester, M., & Greenberg, S. (2005). An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. *Speech Communication*, *47*, 290–311.

Chen, M. Y. (2000). Nasal detection module for a knowledge-based speech recognition system. *In Proceedings of ICSLP*, Vol. 4, pp. 636–639.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English. *Journal of the Acoustical Society of America*, *92*(2), 736–751.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: SAGE publications.

Fosler-Lussier, E., Greenberg, S., & Morgan, N. (1999). Incorporating contextual phonetics into automatic speech recognition. In *Proceedings of the International Congress of Phonetic Sciences*, San Francisco, California.

Frankel, J. (2003). *Linear dynamic models for automatic speech recognition*. PhD thesis, University of Edinburgh.

Frankel, J., Magimai-Doss, M., King, S., Livescu, K., & Çetin, Ö. (2007a). Articulatory feature classifiers trained on 2000 hours of telephone speech. In *Proceedings of Interspeech*, pp. 2485–2488.

Frankel, J., Wester, M., & King, S. (2007b). Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language*, *21*(4), 620–640.

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

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *In Proceedings of ICASSP*, Vol. 1, pp. 517–520.

Greenberg, S. (1997). The Switchboard Transcription Project. In *Research Report #24, Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University.

Greenberg, S. (1999). Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*, 159–176.

Greenberg, S. & Chang, S. (2000). Linguistic dissection of Switchboard-corpus automatic speech recognition systems. In *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, pp. 195–202.

Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., et al. (2005). Landmark-based speech recognition: Report of the 2004 John Hopkins Summer Workshop. In *Proceedings of ICASSP*, pp. 213–216.

Jankowski, C., Kalyanswamy, A., Basson, S., & Spitz, J. (1990). NTIMIT: A phonetically balanced, continuous speech, telephone bandwith speech database. In *Proceedings of ICASSP*, pp. 109–112.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis* (pp. 29–54). Tokyo: The National International Institute for Japanese Language.

Juneja, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks*. PhD thesis, University of Maryland, College Park, MD, USA.

Juneja, A., & Espy-Wilson, C. (2008). A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *The Journal of the Acoustical Society of America*, *123*(2), 1154–1168.

Jyothi, P. (2013). *Discriminative and articulatory feature-based pronunciation models for conversational speech recognition*. PhD thesis, The Ohio State University, Ohio, USA.

Kessens, J. M., Cucchiarini, C., & Strik, H. (2003). A data-driven method for modeling pronunciation variation. *Speech Communication*, *40*, 517–534.

Khasanova, A., Cole, J., & Hasegawa-Johnson, M. (2009). Assessing reliability of automatic burst location. In *Proceedings of Interspeech*.

King, S., Frankel, S., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, *121*(2), 723–742.

King, S., & Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language*, *14*, 333–353.

Kipp, A., Wesenick, M., & Schiel, F. (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. In *Proceedings of ICSLP*, pp. 106–109.

Kipp, A., Wesenick, M., & Schiel, F. (1997). Pronunciation modeling applied to automatic segmentation of spontaneous speech. In *Proceedings of Eurospeech*, pp. 1023–1026.

Kirchhoff, K. (1998). Combining articulatory and acoustic information for speech recognition in noisy and reveberant environments. In *Proceedings of ICSLP*, pp. 891–894.

Kirchhoff, K. (1999). *Robust speech recognition using articulatory information*. PhD thesis, University of Bielefield.

Kirchhoff, K., Fink, G., & Sagerer, G. (2000). Conversational speech recognition using acoustic and articulatory input. In *Proceedings of ICASSP*, pp. 1435–1438.

Kirchhoff, K., Fink, G. A., & Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, *37*, 303–319.

Kohler, K. J. (2001). Articulatory dynamics of vowels and consonants in speech communication. *Journal of the International Phonetic Association*, *31*, 1–16.

Lehtinen, G. & Safra, S. (1998). Generation and selection of pronunciation variants for a flexible word recognizer. In *Proceedings of the ESCA Workshop: Modeling Pronunciation Variation for ASR*, pp. 67–71.

Lin, H., Deng, L., Yu, D., Gong, Y.-f., Acero, A., & Lee, C.-H. (2009). A study on multilingual acoustic modeling for large vocabulary ASR. In *Proceedings ICASSP-2009*, pp. 4333–4336.

Livescu, K., Bezman, A., Borges, N., Yung, L., Çetin, Ö., Frankel, J., et al. (2007). Manual transcriptions of conversational speech at the articulatory feature level. *In Proceedings of ICASSP*, Vol. 1, pp. 953–956.

Manjunath, K., & Sreenivasa Rao, K. (2016). Articulatory and excitation source features for speech recognition in read, extempore and conversation modes. *International Journal of Speech Technology*, *19*, 121–134.

Metze, F. (2005). *Articulatory features for conversational speech recognition*. PhD thesis, Universität Fridericiana zu Karlsruhe, Karlsruhe, Germany.

Metze, F. (2007). Discriminative speaker adaptation using articulatory features. *Speech Communication*, *49*, 348–360.

Naess, A. B., Livescu, K., & Prabhavalkar, R. (2011). Articulatory feature classification using nearest neighbors. In *Proceedings of Interspeech 2011*, pp. 2301–2304.

Niyogi, P., Burges, C., & Ramesh, P. (1999). Distinctive feature detection using support vector machines. In *Proceedings of ICASSP*, pp. 425–428.

Niyogi, P., & Sondhi, M. (2001). Detecting stop consonants in continuous speech. *Journal of the Acoustical Society of America*, *111*, 1063–1076.

Ogden, R. (2009). *An introduction to English phonetics* (pp. 138–153). Edinburgh: Edinburgh University Press.

Ostendorf, M. (1999). Moving beyond the ‘beads-on-a-string’ model of speech. In *Procedings of IEEE ASRU Workshop*, pp. 79–84.

Pernkopf, F., Pham, T. V., & Bilmes, J. A. (2009). Broad phonetic classification using discriminative Bayesian networks. *Speech Communication*, *51*, 151–166.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 3763.

Pruthi, T., & Espy-Wilson, C. Y. (2004). Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, *43*, 225–239.

Pruthi, T., & Espy-Wilson, C. Y. (2007). Acoustic parameters for the automatic detection of vowel nasalization. In *Proceedings of Interspeech*, pp. 1925–1928.

Saenko, K., Livescu, K., Siracusa, M., Wilson, K., Glass, J., & Darrell, T. (2005). Visual speech recognition with loosely synchronized feature streams. *In Proceedings of ICCV*, Vol. 2, pp. 1424–1431.

Salomon, A., Espy-Wilson, C. Y., & Deshmukh, O. (2004). Detection of speech landmarks: Use of temporal information. *Journal of the Acoustical Society of America*, *115*(3), 1296–1305.

Saraçlar, M., Nock, H., & Khudanpur, S. (2000). Pronunciation modelling by sharing gaussian densities across phonetic models. *Computer Speech and Language*, *14*, 137–160.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. *Journal of the Acoustical Society of America*, *127*(6), 3758–3770.

Scharenborg, O. & Cooke, M. (2008). Comparing human and machine recognition performance on a VCV corpus. In *Proceedings of the workshop on Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark.

Scharenborg, O., Wan, V., & Moore, R. K. (2007). Towards capturing fine phonetic variation in speech using articulatory features. *Speech Communication* , *49*, 811–826.

Schölkopf, B., Smola, A. J., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*(5), 1207–1245.

Schuppler, B. (2011). *Automatic Analysis of acoustic reduction in spontaneous speech*. PhD thesis, Radboud University Nijmegen, The Netherlands.

Schuppler, B., van Doremalen, J., Scharenborg, O., Cranen, B., & Boves, L. (2013). The challenge of manner classification in conversational speech. Workshop on speech production in automatic speech recognition, Sattelite Workshop of Interspeech 2013.

Schuppler, B., Adda-Decker, M., & Morales-Cordovilla, J. A. (2014). Pronunciation variation in read and conversational Austrian German. *In Proceedings of Interspeech 2014*, pp. 1453–1457.

Schuppler, B., van Dommelen, W., Koreman, J., & Ernestus, M. (2009a). Word-final [t]-deletion: An analysis on the segmental and sub-segmental level. In *Proceedings of Interspeech*, pp. 2275–2278.

Schuppler, B., van Doremalen, J., Scharenborg, O., Cranen, B., and Boves, L. (2009b). Using temporal information for improving articulatory-acoustic feature classification. In *Proceedings of IEEE ASRU Workshop*, pp. 70–75.

Schutte, K., & Glass, J. (2005). Robust detection of sonorant landmarks. In *Proceedings of Interspeech*, pp. 1005–1008.

Siniscalchi, S. M., & Lee, C.-H. (2014). An attribute detection based approach to automatic speech recognition. *Loquens*, *1*(1), e005. doi:10.3989/loquens.2014.005.

Siniscalchi, S. M., Svendsen, T., and Lee, C.-H. (2007). Towards bottom-up continuous phone recognition. In *Proceedings of IEEE ASRU Workshop*, pp. 566–569.

Siniscalchi, S. M., Svendsen, T., and Lee, C.-H. (2008). Towards a detector-based universal phone recognizer. In *Proceedings of ICASSP*, pp. 4261–4264.

Stüker, S., Metze, F., Schulz, T., and Waibel, A. (2003). Integrating multilingual articulatory features into speech recognition. In *Proceedings of Eurospeech*, pp. 1033 – 1036.

Yoon, S.-Y., Hasegawa-Johnson, M., & Sproat, R. (2010). Landmark-based automatic pronunciation error detection. In *Accepted for presentation at Interspeech 2010.*