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

For additional information about this publication click this link.
http://hdl.handle.net/2066/116264

Please be advised that this information was generated on 2017-07-26 and may be subject to change.
The challenge of manner classification in conversational speech

Barbara Schuppler¹, Joost van Doremalen², Odette Scharenborg³, Bert Cranen², Lou Boves²

¹Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
²Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
³Centre for Language Studies, Radboud University Nijmegen, The Netherlands
b.schuppler@tugraz.at, {j.vandoremalen, o.scharenborg, b.cranen, l.boves}@let.ru.nl

Abstract

In recent years, acoustic-phonetic features (APF) have received great interest as a replacement for phones in automatic speech recognition (ASR) systems. Many studies have focused on improving feature sets and acoustic parameters to describe the APFs. Invariably, these are developed and tested on a limited number of well-researched databases containing read speech. When tested on conversational speech data, these improved APFs and acoustic parameter sets, however, do not show the same improvement. In two experiments, we show that this approach does not work because some of the basic assumptions (here: segmentation in terms of phones) that work well for read speech do not work for conversational speech. More generally speaking, our studies suggest that we need to take the nature of our application data into account already when building the concepts, when defining the basic assumptions of a method, and not only when applying the method to the application data.

Index Terms: acoustic-phonetic feature classification, conversational speech, support vector machines

1. Introduction

Acoustic-phonetic features (APFs) have received substantial interest in the field of speech science and technology as basic unit of representation (e.g., [1, 2, 3, 4, 5, 6]). The most popular corpus for APF classification is beyond a doubt TIMIT [7], a corpus of read American English, (e.g., [8, 9, 11]). No other corpus of comparable size comes with equally accurate and detailed phonetic transcriptions. Hardly any work on APF classification, however, has been done on spontaneous, conversational speech (e.g., Switchboard). APFs are obtained using the conventional approach of automatic mapping from phone transcriptions to APFs. The elitist approach was proposed 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 its probability for being correctly classified. For training the final model, only those frames are selected, whose probability for correct classification is bellow a set threshold.

In both studies, Support Vector Machines (SVMs) are used. The SVMs are trained and tested using the LibSVM package [21]. We adopt the one-versus-one method and use the soft-margin approach. The parameters $C$ and $\gamma$ are optimized for each study separately (and explained below) using a grid search.

2. Materials and Methods

2.1. The two studies

Study I presents experiments aimed at optimizing the acoustic parameters for manner classification using parameters that provide both a high frequency and a high time resolution. We train and test classifiers for read speech (TIMIT) and spontaneous speech (Switchboard). APFs are obtained using the conventional approach of automatic mapping from phone transcriptions to APFs.

Study II investigates the so-called elitist approach [20] on the task of APF classification of conversational speech. The elitist approach was proposed 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 its probability for being correctly classified. For training the final model, only those frames are selected, whose probability for correct classification is bellow a set threshold.

In both studies, Support Vector Machines (SVMs) are used. The SVMs are trained and tested using the LibSVM package [21]. We adopt the one-versus-one method and use the soft-margin approach. The parameters $C$ and $\gamma$ are optimized for each study separately (and explained below) using a grid search.

2.2. Acoustic-phonetic feature values

Table 1 shows our set of APF values. In order to develop acoustic parameters for accurate classification of stationary sounds, plosives are represented as a sequence of closure and release. Affricates are a sequence of a plosive and a fricative. Since not all of our available speech material comes with boundaries between these two parts, affricates were excluded from our experiments. As manner of articulation is only defined for consonants, vowels were excluded too.
and 97% in word-initial position). Reported an agreement for plosives of 47% in word-medial position. This agreement is in the range of what has been reported in the literature, e.g., 63% in the manual-SV-APF corpus. The agreement of labels obtained with the automatic labeling method and the manually obtained labels of the plosive segments in the manual-SV-APF corpus was 63%. This agreement is in the range of what has been reported in the literature, e.g., (20, 5, 4, 6, 11]). With the conventional 25 ms window shifted with 10 ms, good results are obtained for fairly stationary features. In order to accurately detect short acoustic events, such as bursts in plosives, shorter window lengths and shifts are needed, e.g., [18] used 5 ms windows shifted with 1 ms steps.

Our goal, however, is to capture both very short (e.g., bursts) and longer acoustic events (e.g., nasality). To that end, we investigate MFCCs derived from 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 all sets, the input speech is first divided into overlapping Hamming windows of 25 ms or 5 ms with a 10 ms 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 $7 \times 39 = 273$. 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. Study I: Improving acoustic parameters for APF 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 APF classification tasks, however, mostly MFCCs have been used (e.g., [20, 5, 4, 6, 11]). With the conventional 25 ms window shifted with 10 ms, good results are obtained for fairly stationary features. In order to accurately detect short acoustic events, such as bursts in plosives, shorter window lengths and shifts are needed, e.g., [18] used 5 ms windows shifted with 1 ms steps.

For all sets, the input speech is first divided into overlapping Hamming windows of 25 ms or 5 ms with a 10 ms 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 the 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 $7 \times 39 = 273$. 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. APF classification of TIMIT

For the optimization of the $C$ and $\gamma$ parameters, two independent subsets of 5000 feature vectors (one for training and one for testing) were extracted from the original TIMIT training set. For training the SVM classifiers with the Baseline parameters, 100k vectors were extracted from randomly chosen files from the TIMIT training set. For the Short, Long, and Both parameters, the same audio data was 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 APF classification accuracy in terms of percentage correctly classified frames on the TIMIT test set.
Comparing the three new acoustic parameters with the baseline shows that the *Both* classifier performed best for ‘burst+release’ (Bur): the F-score increases from 0.73 to 0.82. This was to be expected, since bursts are events of very short duration. The *Short* and *Both* classifiers perform best for ‘fricative’ (Fric). The average frame level accuracies are: 83.4% for Baseline, 85.3% for *Short*, 87.0% for *Long*, and 87.7% for *Both*. Most importantly, the *Both* classifier seems to be able to combine the classification power of the *Short* and *Long* classifiers.

Since studies from the literature all tend to use slightly different sets of APF values, a fair comparison with the performance of our classifiers is not possible. However, the accuracy of 87.7% reached by our new set of acoustic parameters show satisfactory results for TIMIT, both in comparison with our baseline and with previous results for multi-value classification experiments from the literature (e.g., for manner of articulation - excluding vowels) [19] achieved an accuracy of 75.6%, [20] of 70%, and [17] of 74.8%.

### 3.3. APF classification of Switchboard

We trained classifiers with the best performing set acoustic parameters (*Baseline* and *Both*) on the complete STP material using the same procedure as for TIMIT. Classifiers were trained using 50k frames with the *Baseline* feature (10 ms shift) and the corresponding 200k frames for *Both* (2.5 ms shift). The classifiers were tested on automatic-SV-APF, which consists of 53,115 frames.

Table 4 shows that overall the frame-level classification accuracy and the performance in terms of F-scores obtained with Switchboard is much lower than the results for TIMIT, i.e., F = 0.65 vs. F = 0.88 for *Both*. Comparing *Baseline* and *Both* on Switchboard, does not show the same improvement for *Both* that was found for TIMIT (*Both*: F = 0.65 vs. BL: F = 0.66). Moreover, the additional temporal information does not yield the rise in performance for the short events (F-scores for ‘burst+release’ (Bur): 0.41 for *Baseline* vs. 0.37 for *Both*) which was found in the TIMIT experiments. Apparently, improvements obtained for read speech do not generalize to spontaneous speech.

Also previous studies reported that classification performance is substantially worse for Switchboard than for TIMIT. For instance, Pruthi and Espy-Wilson [15] report accuracies of 77.90% for detecting vowel nasalization in TIMIT, but only 69.58% for Switchboard.

### 3.4. Impact of labeling accuracy

Conversational speech shows more variability (e.g., [26]) than read speech, and articulatory gestures may heavily overlap. We hypothesize that the canonical mapping from phone labels to APF labels introduces more errors in spontaneous than in read speech. This might explain why the improvement found for *Both* for TIMIT was not found for Switchboard.

In order to estimate the impact of the labeling accuracy of the test set, the classifiers trained on STP (cf. section 3.3) were additionally tested on the 53,115 frames of manual-SV-APF, which contains more accurate APF labels. The results showed that the overall classification performance is still lower than for TIMIT (*BL*: 0.60 vs. 0.84 for TIMIT) but here the *Both* classifier performs better than the *Baseline* classifier (F-scores: 0.65 vs. 0.60). This experiment shows that the labeling accuracy of the test set does have an impact on the classification performance.

In order to estimate the impact of labeling errors in the training set, we calculated the amount of erroneous labels. We
The results showed that combining MFCCs derived from a long speech (TIMIT) and spontaneous speech (Switchboard). In both cases, the acoustic parameters corresponding to the winning APF are designed databases do not automatically do so for real-life data. Therefore, the acoustic parameters corresponding to the winning APF are used for training does not actually represent the putative acoustic feature.

4. Study II: The elitist approach for APF classification of Switchboard

For this study, first an SVM classifier with the Both acoustic parameters is trained on the 200k frames in the STP data set. Then, this classifier is used to predict the posterior probabilities for each frame of the STP training material (following the method presented in [27]). Finally, we train classifiers on only those frames for which 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).

Table 5: Elitist approach: Frame-level F-scores and overall accuracy for the APF classifiers trained on STP and tested on manual-SV-APF.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Slt</th>
<th>Liq</th>
<th>Glt</th>
<th>Nas</th>
<th>Ftr</th>
<th>Bur</th>
<th>Cli</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.91</td>
<td>0.67</td>
<td>0.60</td>
<td>0.68</td>
<td>0.57</td>
<td>0.57</td>
<td>0.52</td>
<td>64.5</td>
</tr>
<tr>
<td>0.50</td>
<td>0.91</td>
<td>0.68</td>
<td>0.60</td>
<td>0.66</td>
<td>0.59</td>
<td>0.49</td>
<td>0.53</td>
<td>64.1</td>
</tr>
<tr>
<td>0.70</td>
<td>0.90</td>
<td>0.66</td>
<td>0.59</td>
<td>0.66</td>
<td>0.59</td>
<td>0.49</td>
<td>0.53</td>
<td>63.9</td>
</tr>
<tr>
<td>0.90</td>
<td>0.91</td>
<td>0.67</td>
<td>0.60</td>
<td>0.65</td>
<td>0.59</td>
<td>0.48</td>
<td>0.53</td>
<td>64.1</td>
</tr>
<tr>
<td>0.95</td>
<td>0.91</td>
<td>0.66</td>
<td>0.60</td>
<td>0.64</td>
<td>0.58</td>
<td>0.47</td>
<td>0.54</td>
<td>64.2</td>
</tr>
</tbody>
</table>

Comparing all different threshold settings (see Table 5), the highest average accuracy (64.5%) is obtained with the original training set. Thus, there is no increase in classification accuracy when training the classifiers with a subset of frames. For the individual APF values, different threshold settings are optimal. Whereas silence, glide, nasal, and burst do not profit from the training on a selection of the best frames, other APFs profit from data selection (liquid, fricative, and closure).

Our results on spontaneous speech do not confirm the findings of Chang et al. [20], who achieved an 8% absolute improvement in classification accuracy on read speech (NTIMIT [28]). Although they used a slightly different APF set, we think that the difference in performance improvement is mainly due to the difference in speech style, and to the fact that we did not remove potentially ambiguous frames from the test. In spontaneous speech, there are fewer frames than in carefully read speech which correspond to a ‘pure’ APF value. Therefore, the acoustic parameters corresponding to the winning APF are likely to overlap with the parameters of other APFs.

5. General discussion and conclusions

This paper presents two studies which demonstrate that methods for APF classification yielding improvements in specifically designed databases do not automatically do so for real-life data. The first study presented a set of acoustic parameters with a high time and a high frequency resolution which were tested on read (TIMIT) and spontaneous speech (Switchboard). In both cases, we applied the conventional mapping from phone to APF 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) overall outperforms MFCCs derived from a window of 25 ms shifted with 10 ms (F=0.84). For spontaneous speech, however, the overall performance dropped to F=0.66 for Baseline and, what is more, there was no gain in performance for the new set of acoustic parameters (Both: F=0.65) over Baseline. In the second study, we applied the elitist approach, which earlier showed a performance improvement of 8% in NTIMIT [20], to conversational speech material (Switchboard). In our experiments, however, this method did not improve classification performance over all APFs, and for fricatives and closure by 2%.

Analysis of the labeled material as well as of our experiments with manually and automatically created APF labels showed that the labeling method has a great impact on classification performance (Section 3.4). It is questionable whether a segmentation in terms of phones is equally suitable for the two speech styles in the first place. Due to the high pronunciation variability in spontaneous speech (e.g., [26, 29]), a segmentation in terms of phones is extremely difficult. This difficulty is reflected in the inter-human labeling disagreement of phonetic transcriptions (5.6% for read speech vs. 21.2% for spontaneous speech) [30, 31]. Therefore, the accuracy of the phonetic segmentation in TIMIT is surely higher than in STP. For all these reasons, we argue that the ‘canonical’ mapping from phone labels to APF labels may still result in relatively good training material for read speech, while it does not for spontaneous speech. This is especially apparent for features that are inherently difficult to define. For example, confusions of glides and liquids are much more frequent in spontaneous 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 [32]. 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 actual speech.

Performance drops when going from carefully articulated data (TIMIT) to real-life data (Switchboard) have also been reported for ASR, where for TIMIT word accuracies are typically >95%, while for Switchboard they are only in the 50 - 70% range [13]. Hence, it is not surprising that our classification performance is overall worse for spontaneous speech than for read speech. It is surprising, however, that the relative performance improvements due to our new methods do not transfer from read to spontaneous speech.

In speech science, methods are mostly developed and improved using read speech corpora (e.g., TIMIT) and only afterwards they are adapted to spontaneous speech. Our studies suggest that the nature of the application data needs to be taken into account already when defining the concepts (here: 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.

6. Acknowledgements

The work by Barbara Schuppler was supported by a Hertha-Firnberg grant from the FWF (Austrian Science Fund). Joost van Doremalen was supported by the project DISCO, funded by the Dutch-Flemish programme STEVIN. The research by Odette Scharenborg is supported by a Vidi-grant from the Netherlands Organization for Scientific Research (NWO).
7. References


