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 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 γ 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 below 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 γ 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). \[24\] reported an agreement for plosives of 47% in word-medial position. This 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 (CO-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 mid-point 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]).

### 2.4. Spontaneous speech corpus: Switchboard

Switchboard is a corpus of telephone bandwidth speech from spontaneous conversations speech from 500 speakers of American English [12].

**Switchboard-APF (SV-APF)** consists of 78 utterances (a total of 119 s of speech, excluding silences) [14]. There is no overlap between STP and SVAPF: SV-APF contains phone labels along with APF labels. The original set of APF labels was manually adapted to our set of labels [25], starting from the tier ‘Dg1’ (Degree of forward constriction). Table 2 shows the mapping between these two sets. The resulting transcriptions are further referred to as **manual-SV-APF**. Additionally, in order to be able to make direct comparisons between the Switchboard and TIMIT results, the APF labels (and boundaries) from the 78 **manual-SV-APF** utterances were generated automatically. These transcriptions are referred to as **automatic-SV-APF**.

The **Switchboard Transcription Project (STP)** [22] contains 72 minutes of speech from the Switchboard corpus (taken from 618 conversations by 370 different speakers) that were manually transcribed phonetically. 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 [23]. 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., [24] reported an agreement for plosives of 47% in word-medial and 97% in word-initial position).

### Table 1: Mapping of TIMIT phone symbols to the manner APF values.

<table>
<thead>
<tr>
<th>Phone</th>
<th>Manner APF Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sil, pau, h#</td>
<td>silence</td>
</tr>
<tr>
<td>l, el, r</td>
<td>liquid</td>
</tr>
<tr>
<td>w, y</td>
<td>glide</td>
</tr>
<tr>
<td>em, en, eng, m, n, ng, nx</td>
<td>nasal</td>
</tr>
<tr>
<td>dh, f, hh, s, sh, th, v, z, zh, hv</td>
<td>fricative</td>
</tr>
<tr>
<td>b, d, g, p, t, k, q</td>
<td>burst-release</td>
</tr>
<tr>
<td>bcl, dcl, gcl, pcl, tcl, kcl</td>
<td>closure</td>
</tr>
<tr>
<td>ch, jh, dh, ep, all vowels</td>
<td>silence</td>
</tr>
</tbody>
</table>

### Table 2: Phone-to-APF mapping

<table>
<thead>
<tr>
<th>Phone Dg1</th>
<th>Our APF Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>I, el, r</td>
<td>closure</td>
</tr>
<tr>
<td>w, y</td>
<td>approximan</td>
</tr>
<tr>
<td>dh, f, hh, s, sh, th, v, z, zh, hv</td>
<td>fricative</td>
</tr>
<tr>
<td>b, d, g, p, t, k, q</td>
<td>fricative</td>
</tr>
<tr>
<td>bcl, dcl, gcl, pcl, tcl, kcl</td>
<td>closure</td>
</tr>
<tr>
<td>silence</td>
<td>silence</td>
</tr>
</tbody>
</table>

#### 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.

### Table 3: Mapping of the values in SV-APF (‘Dg1’ = Degree of forward constriction) to our set of APF values.

<table>
<thead>
<tr>
<th>Phone</th>
<th>Dg1</th>
<th>Our APF Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>w, y</td>
<td>glide</td>
<td>silence</td>
</tr>
<tr>
<td>l, el, r</td>
<td>closure</td>
<td>silence</td>
</tr>
<tr>
<td>em, en, eng, m, n, ng, nx</td>
<td>nasal</td>
<td></td>
</tr>
<tr>
<td>dh, f, hh, s, sh, th, v, z, zh, hv</td>
<td>fricative</td>
<td></td>
</tr>
<tr>
<td>b, d, g, p, t, k, q</td>
<td>fricative</td>
<td></td>
</tr>
<tr>
<td>bcl, dcl, gcl, pcl, tcl, kcl</td>
<td>closure</td>
<td></td>
</tr>
<tr>
<td>silence</td>
<td>silence</td>
<td></td>
</tr>
</tbody>
</table>

### 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. 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 (CO-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 mid-point were concatenated, resulting in feature vectors of length $273 + 147 = 420$.
Table 3: Frame-level confusion matrices for the APF classifiers trained and tested on TIMIT. The row labels represent the true classes, the column labels represent the classes recognized by the classifier. Average F-scores: BL = 0.84; Short = 0.86; Long = 0.87; Both = 0.88.

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 3.4) 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: 0.65 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 erroneousness in the training set, we calculated the amount of erroneous labels. We
The results showed that combining MFCCs derived from a long (TIMIT) and spontaneous speech (Switchboard). In both cases, time and a high frequency resolution which were tested on read speech. The first study presented a set of acoustic parameters with a high odds for APF classification yielding improvements in specifically 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 segmentations 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


