The challenge of *manner* classification in conversational speech

Barbara Schuppler[1], *Joost van Doremalen*[2], *Odette Scharenborg*[3], *Bert Cranen*[2], *Lou Boves*[2]

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[2]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
[3]Centre 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 [12]), even though a major reason for using APFs is that this representation might have more potential to capture pronunciation variability. To our knowledge, only the work done during the 2004 Johns Hopkins Summer Workshop [13, 1, 14] and Pruthi and Espy-Wilson [15] used Switchboard for APF research.

In the field of APF classification, researchers have focused on finding the ideal set of acoustic parameters for building either multi-value or binary classifiers (e.g., [16, 4, 17, 18]) or the ideal statistical classification method (e.g., a comparison of ANNs with SVMs by [19]).

Previous studies have in common that it is not known whether improved classification in read speech generalizes to spontaneous speech. Also, classifiers are trained and tested on the basis of APF labels that were automatically generated from broad phonetic transcriptions. APF labels created in that way change synchronously at phone boundaries, which obviously violates the observation that articulators move independently and asynchronously. The effect of this automatic mapping on APF classification performance may well be much larger in spontaneous speech than in read speech.

This paper presents two studies which demonstrate that methods yielding improvements in specifically designed databases do not automatically do so in real-life data. The first study aims at developing acoustic parameters for accurate classification of stationary sounds (e.g., nasals) as well as short acoustic events, such as bursts in plosives. The second study presents experiments from a data selection approach for improving the training material for APF classifiers. Both studies compare the performance achieved on read speech vs. conversational speech, and present analyses of observed discrepancies.

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 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.

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

Phone	Manner APF 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, all vowels	NIL

Table 2: **Phone-to-APF mapping**: Mapping of the values in SV-APF ('Dg1' = Degree of forward constriction) to our set of APF values.

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

2.3. Read speech corpus: TIMIT

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

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].

SVitchboard-APF (SV-APF) consists of 78 utterances (a total of 119 s of speech, excluding silences) [14]. There is no overlap between STP and SV-APF. 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).

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 (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. APF classification of TIMIT

For the optimization of the C and γ 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 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 best performing classifier for each APF is highlighted in bold.

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 represented the classes recognized by the classifier. Average F-scores: *BL* = 0.84; *Short* =0.86; *Long* = 0.87; *Both* = 0.88.

BL	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	$93.2_{\pm.2}$	0.3	0.2	0.8	3.0	0.6	2.0
Liq	0.4	$89.0_{\pm.3}$	2.4	2.8	2.8	1.0	1.4
Gli	0.8	12.8	$77.0_{\pm.7}$	2.8	3.3	1.3	2.0
Nas	1.6	2.4	0.6	$86.3_{\pm.4}$	4.2	0.5	4.4
Fric	2.7	1.1	0.4	1.6	$89.5_{\pm.2}$	1.6	3.1
Bur	2.8	2.9	0.6	1.9	12.8	$65.2_{\pm.6}$	13.8
Clo	4.3	0.9	0.3	3.5	4.9	2.4	$83.6_{\pm.3}$
Sum	105.8	109.4	81.5	99.7	120.5	72.6	110.3
F	**0.93**	0.88	0.81	0.84	0.88	0.73	0.83

Short	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	$92.5_{\pm.1}$	0.2	0.1	0.9	3.6	0.6	2.1
Liq	0.6	$89.1_{\pm.2}$	2.8	3.7	2.1	1.0	0.8
Gli	0.6	13.8	$78.1_{\pm.4}$	3.2	2.1	1.4	0.8
Nas	1.7	2.7	0.8	$87.6_{\pm.2}$	3.2	0.4	3.5
Fric	3.0	0.8	0.4	2.0	$88.5_{\pm.1}$	2.5	2.9
Bur	2.5	1.5	0.5	0.8	11.6	$76.4_{\pm.3}$	2.9
Clo	4.9	0.6	0.2	3.2	4.1	2.1	$84.8_{\pm.2}$
Sum	105.8	108.7	82.9	101.4	115.2	84.4	101.6
F	0.92	0.89	0.81	0.85	0.88	0.80	0.85

Long	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	$93.3_{\pm.1}$	0.2	0.1	0.7	2.9	0.7	2.1
Liq	0.5	$90.3_{\pm.2}$	2.9	2.2	2.2	1.0	0.9
Gli	0.8	9.8	$83.1_{\pm.3}$	2.0	2.2	1.2	1.0
Nas	1.7	1.8	0.7	$89.0_{\pm.2}$	2.5	0.4	3.8
Fric	2.4	0.8	0.4	1.3	$90.1_{\pm.1}$	2.3	2.7
Bur	2.4	1.4	0.5	0.5	10.3	$78.0_{\pm.3}$	7.0
Clo	4.3	0.7	0.2	3.0	4.0	2.7	$85.1_{\pm.2}$
Sum	105.4	105.0	87.9	98.7	114.2	86.3	102.6
F	**0.93**	0.90	0.85	**0.88**	**0.90**	0.81	0.85

Both	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	$93.5_{\pm.1}$	0.2	0.1	0.7	2.6	0.7	2.2
Liq	0.5	$91.0_{\pm.2}$	2.5	2.0	2.0	1.0	0.9
Gli	0.6	9.3	$84.2_{\pm.3}$	1.8	1.9	1.3	0.9
Nas	1.6	1.7	0.7	$89.4_{\pm.2}$	2.3	0.3	3.9
Fric	2.3	0.7	0.3	1.2	$90.8_{\pm.1}$	2.2	2.5
Bur	2.0	1.3	0.5	0.6	8.9	$79.6_{\pm.2}$	7.0
Clo	4.1	0.6	0.2	2.9	3.8	2.8	$85.6_{\pm.2}$
Sum	104.6	104.2	88.5	98.6	112.3	87.9	103.0
F	**0.93**	**0.91**	**0.86**	**0.88**	**0.90**	**0.82**	**0.86**

Table 4: Frame-level confusion matrices for classifiers trained on **STP** and tested on **automatic-SV-APF**. Average F-scores: *BL* = 0.66; *Both* = 0.65.

BL	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	$91.3_{\pm.6}$	0.4	0.2	1.0	5.0	0.7	1.5
Liq	6.9	$73.0_{\pm3.4}$	6.9	6.3	5.1	0.2	1.6
Gli	2.8	16.8	$64.0_{\pm3.8}$	10.2	5.2	0.7	0.2
Nas	6.9	6.1	4.4	$77.7_{\pm2.8}$	4.5	0.1	0.4
Fric	13.7	1.7	1.0	7.2	$68.4_{\pm2.2}$	4.9	3.1
Bur	25.6	2.2	1.6	3.5	25.0	$34.8_{\pm4.5}$	0.7
Clo	16.2	0.6	0.8	8.4	21.0	2.8	$50.3_{\pm2.9}$
Sum	163.4	100.8	78.9	114.3	134.2	44.2	57.8
F	0.91	0.70	0.69	0.68	0.63	0.41	0.60

Both	Sil	Liq	Gli	Nas	Fric	Bur	Clo
Sil	$87.4_{\pm.3}$	1.2	0.9	2.1	5.2	1.1	2.1
Liq	8.8	$77.9_{\pm1.8}$	4.4	3.5	3.5	0.4	1.5
Gli	2.3	12.6	$67.1_{\pm2.3}$	11.1	6.2	0.8	0.0
Nas	7.9	6.8	3.6	$74.7_{\pm1.5}$	5.0	1.1	1.0
Fric	13.4	1.1	1.4	5.4	$70.9_{\pm1.2}$	4.4	3.3
Bur	28.9	2.0	1.0	3.4	19.7	$36.2_{\pm1.8}$	8.9
Clo	18.9	1.3	0.6	7.0	18.8	6.3	$47.1_{\pm1.5}$
Sum	167.6	102.9	79.0	107.2	129.2	50.3	64.0
F	0.89	0.71	0.69	0.66	0.65	0.37	0.55

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 accuracy in the training set, we calculated the amount of erroneous labels. We computed the number of speech samples in which the labeling in the automatic-SV-APF transcriptions differed from the manual-SV-APF labels. Overall, we observed a disagreement for 19.9% of the samples. A more detailed analysis for all APF values separately showed that 29.4% of the samples carrying the label 'liquid' (Liq) in the automatic-SV-APF set did not contain a liquid according to the human labelers. Hence, a substantial part of labels 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.

Threshold	Sil	Liq	Gli	Nas	Fri	Bur	Clo	Acc.
0.00	0.91	0.67	**0.60**	**0.68**	0.57	**0.57**	0.52	64.5
0.50	0.91	**0.68**	0.60	0.66	0.59	0.49	0.53	64.1
0.70	0.90	0.66	0.59	0.66	0.59	0.49	0.53	63.9
0.90	0.91	0.67	0.60	0.65	**0.59**	0.48	0.53	64.1
0.95	**0.91**	0.66	0.60	0.64	0.58	0.47	**0.54**	64.2

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 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

[1] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: Report of the 2004 John Hopkins Summer Workshop," in *Proceedings of ICASSP*, 2005, pp. 213–216.

[2] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland, College Park, MD, USA, 2004.

[3] A. Juneja and C. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1154–1168, 2008.

[4] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, pp. 333–353, 2000.

[5] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefield, 1999.

[6] K. Schutte and J. Glass, "Robust detection of sonorant landmarks," in *Proceedings of Interspeech*, 2005, pp. 1005–1008.

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

[8] F. Pernkopf, T. V. Pham, and J. A. Bilmes, "Broad phonetic classification using discriminative bayesian networks," *Speech Communication*, vol. 51, pp. 151–166, 2009.

[9] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, pp. 225–239, 2004.

[10] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in *Proceedings of IEEE ASRU Workshop*, 2007, pp. 566–569.

[11] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, vol. 1, 1992, pp. 517–520.

[12] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and Ö. Çetin, "Articulatory feature classifieres trained on 2000 hours of telephone speech," in *Proceedings of Interspeech*, 2007, pp. 2485–2488.

[13] K. Livescu, A. Bezman, N. Borges, L. Yung, Ö. Çetin, J. Frankel, S. King, M. Magimai-Doss, X. Chi, and L. Lavoie, "Manual transcriptions of conversational speech at the articulatory feature level," in *Proceedings of ICASSP*, vol. 1, 2007, pp. 953–956.

[14] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," in *Proceedings of Interspeech*, 2007, pp. 1925–1928.

[15] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks." *Computer Speech and Language*, vol. 21, no. 4, pp. 620–640, 2007.

[16] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1296–1305, 2004.

[17] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," in *Proceedings of ICASSP*, 1999, pp. 425–428.

[18] O. Scharenborg, V. Wan, and R. K. Moore, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication - Special Issue on Intrinsic Speech Variation and Speech Recognition*, vol. 49, pp. 811–826, 2007.

[19] S. Chang, M. Wester, and S. Greenberg, "An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language," *Speech Communication*, vol. 47, pp. 290–311, 2005.

[20] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm (last viewed 28/05/2010), 2001.

[21] B. Schuppler, J. van Doremalen, O. Scharenborg, B. Cranen, and L. Boves, "Using temporal information for improving articulatory-acoustic feature classification," in *Proceedings of IEEE ASRU Workshop*, 2009, pp. 70–75.

[22] S. Greenberg, "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, 1997.

[23] B. Schuppler, "Automatic analysis of acoustic reduction in spontaneous speech," Ph.D. dissertation, Radboud University of Nijmegen, Nijmegen, The Netherlands, 2011.

[24] A. Khasanova, J. Cole, and M. Hasegawa-Johnson, "Assessing reliability of automatic burst location," in *Proceedings of Interspeech*, 2009.

[25] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous Speech: Data and Analysis. The National International Institute for Japanese Language.*, K. Yoneyama and K.Maekawa, Eds., Tokyo, Japan, 2004, pp. 29–54.

[26] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[27] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwith speech database," in *Proceedings of ICASSP*, 1990, pp. 109–112.

[28] K. J. Kohler, "Articulatory dynamics of vowels and consonants in speech communication," *Journal of the International Phonetic Association*, vol. 31, pp. 1–16, 2001.

[29] A. Kipp, M. Wesenick, and F. Schiel, "Automatic detection and segmentation of pronunciation variants in German speech corpora," in *Proceedings of ICSLP*, 1996, pp. 106–109.

[30] ——, "Pronunciation modeling applied to automatic segmentation of spontaneous speech," in *Proceedings of Eurospeech*, 1997, pp. 1023–1026.

[31] C. Y. Espy-Wilson, "Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English." *Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 736–751, 1992.