Bachelor thesis

Artificial Intelligence

Radboud University

Automatic extraction of characterizing features for non-native Dutch read speech

Author:
Rosa Bosland
s1010114

First supervisor:
dr. W.A.J. Strik
`Centre for Language Studies`
`helmer.strik@ru.nl`

Second supervisor:
dr. K.A. van der Heijden
`Donders Institute for Brain,`
`Cognition and Behaviour`
`kiki.vanderheijden@donders.ru.nl`

January 25, 2022

Abstract

Although finding characteristic features for English atypical speech has been the topic for many researches, not much research has been done for atypical Dutch speech. In 2008, the JASMIN speech corpus was completed, a spoken Dutch corpus containing children, elderly and non-native Dutch speakers [1]. In this thesis, non-native Dutch read speech from JASMIN is compared to native read speech to find out which features are characteristic for non-native Dutch speech. By automatically computing 103 word level Praat and eGeMAPS features from speech recordings and transcriptions, ranking these features with a Recursive Feature Elimination (RFE) method, classifying them with binary comparisons using a Support Vector Machine (SVM), and finally evaluating them using statistical tests, this research succeeded in automatic extraction of characteristic features for non-native Dutch read speech. Through binary comparisons with native speech, 93 out of 103 features were found to be significantly different. Two characteristic and partly overlapping sets of features were found; the first set based on the RFE ranking, the second based on an individual effect size ranking. Both sets support the hypotheses that a lower speaker volume and lower order Mel-Frequency-Cepstral-Coefficients are characteristic for non-native Dutch speech, and show indications of a slower reading pace for non-natives. Moreover, formant related features were prevalent in both rankings, indicating a different shape of the vocal tract owing to deviations in non-native pronunciation compared to native speakers.

Contents

Chapter 1

Introduction

Speech recognition applications are developing rapidly and taking an increasingly more important role in our society. Since speech recognition systems are generally developed for the average population with standard speech (i.e. native adult speech), they have reduced performance for subgroups with different speaking patterns. Speech patterns from these subgroups are referred to as atypical speech, and include children, elderly, non-native Dutch speakers, and pathological speech. In 2008, the JASMIN speech corpus was completed, a spoken Dutch corpus containing children, elderly and non-native Dutch speakers [1]. JASMIN was made as an addition to the Spoken Dutch Corpus (CGN), which was restricted to the speech of adult, native speakers of Dutch in the Netherlands and Flanders. JASMIN extends the CGN along three dimensions: age (children and elderly), first language (non-natives) and communication setting (human machine interaction). The data from JASMIN can support research to higher level technology that considerably increases non-natives opportunities and their participation in our society. For instance, characteristic features for non-native speech can help improve Automatic Speech Recognition (ASR) applications for this subgroup and hence support their participation in the Information Society [2]. ASR can also facilitate non-natives acquisition of the Dutch language through pronunciation training. On the other hand, language courses can improve their students pronunciation if the features characterizing the speech of a native speaker are known.

Over the past years, automatic speech analysis has become a prominent approach to study atypical speech. This field of research is facilitated by the development of effective research instruments such as the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [3]. eGeMAPS is a set of standardized acoustic features with shown theoretical relevance and potential to index relevant aspects of speech production. These features can be automatically extracted from speech recordings.

A previous research by Van Bemmel showed successful classification through automatically extracted acoustic features of COPD patients in exacerbated and stable condition [8]. An interesting follow-up is to apply the pipeline of this research to other atypical speech. In the current research, I will analyze non-native Dutch read speech using the same methods; word level features will be computed using the segmentations obtained by a forced aligner. Through a binary comparison with Recursive Feature Elimination (RFE) and statistical analyses, I aim to get insight in characteristic features of non-native Dutch speech. The obtained features can be used for further research and application development for non-native speakers.

The aim of this research is to find through automatic extraction a characteristic set of features for non-native Dutch read speech. First I will elaborate on related work (Chapter 2), then I will explain the pipeline used in this research (Chapter 3). In Chapter 4 the results will be shown, which will be analyzed in the discussion section (Chapter 5). The chapter ends with limitations of this thesis, and suggestions for future research.

Chapter 2

Related Work

This section first summarizes research by Van Bemmel [11] which forms the basis for the pipeline of my research. This is followed by a brief overview of other studies that extracted characteristic features for atypical speech. Based on this background information, I will formulate my research question and hypotheses at the end of the chapter.

2.1 Automatic selection of characterizing features for atypical speech

In her 2021 study, Van Bemmel explored the possibilities of automatic selection of the most characterizing acoustic features for different types of Dutch atypical speech [11]. Automatic analysis is advantageous over human analysis in the aspects of limiting human bias, being quick and consistent, and not requiring the time and effort of a (medical) professional. Since speech contains a lot of information about a person and one's health, and since recording speech is non-invasive, there have been numerous endeavors to develop applications for remote monitoring or diagnosis for various pathologies that are apparent in speech. It is key to find markers in speech that indicate whether the speech is typical or atypical to ensure quality of these applications.

Each type of atypical speech differs from typical speech in a distinct way. This specific research analyzed three types of Dutch atypical speech. The first dataset comprises speech from people suffering from Chronic Obstructive Pulmonary Disease (COPD) in both stable and exacerbated condition. Secondly, there is speech from people suffering from Parkinson's Disease (PD) recorded at before the experiment, before treatment, and after treatment. The last atypical speech type is non-native speech mainly consisting of Dutch idioms spoken by native German speakers who are currently learning Dutch.

This research aimed to find the most characterizing acoustic features per type of Dutch atypical speech by first automatically extracting Praat and eGeMAPS features from speech on three different linguistic levels (full, word and phoneme), and then ranking these features with Recursive Feature Elimination (RFE). This ranking is necessary, because using the entire feature set frequently yields a lower accuracy compared to using a subset, a concept known as the curse of dimensionality. Lastly, classification and statistical analysis is done to analyse and evaluate the selected features. Classification is performed with multiple classifiers, Leave-One-Subject-Out (LOSO) cross-validation and evaluated with the Matthews Correlation Coefficient (MCC). The statistical test chosen in her research is a Multivariate ANOVA (MANOVA) for COPD and non-native speech as a between-subject analysis, for PD a repeated measures GLM for within-subject approach.

In her research, the top ten characteristic features were found to be significant in the multivariate tests for all three types of atypical speech, indicating that the combination of these features is indeed characteristic of that type of speech. In terms of segmentation level, the full-segmentation features yielded the highest classification score, and phoneme-segmented features the lowest. This is plausible, because full level segmentation features provide more linguistic information compared to the other two segmentations. However, full-segmentation level selected features are prone to overfitting. Since my data contains different prompts and participants for the two sample groups, a full level analysis will probably not be informative. For this reason, features will be extracted and analyzed at word level in my research.

Regarding classification, the SVM classifier with linear kernel generally outperformed the other classifiers. Regarding feature selection, results showed that RFE is a suitable method for selecting the most characterizing features; RFE consistently outperformed classification using the entire feature set.

For all three atypical speech groups, they found similarities and differences when comparing the most characterizing features with significant features found in previous studies. For more detailed information on this with regards to COPD and PD data, please refer back the the original paper [11]. For non-native speech, previous literature suggested that generally rate of speech is lower, and duration is higher compared to native speech. However, Van Bemmel's results for non-native data selected no duration features in the top 10 from RFE. A found distinction between German speakers and reference speech is that the volume is less loud; a lower volume could indicate confidence in speaking, which is logical for non-native speakers.

The methods used in this study were selected to be generally applicable to other types of atypical speech with little data available. Based on the methods and findings from this research, I will analyze non-native read speech data on word level.

2.2 Characteristic features

Previous studies have already found that there are objective features that characterize certain atypical speech. However, these features differ per language, speech task and atypical speech [17] [18].

Research by Cucchiarini et al. investigated perceived fluency in both read and spontaneous Dutch speech of non-natives [18]. Fluency ratings by experts and objective measures obtained by means of a continuous speech recognizer (CSR) of the same speech fragments were compared. There were striking differences between the two types of speech (read and spontaneous). Also, the automatically calculated variables varied as a function of non-natives proficiency level. The results indicate that for read speech (which is the type of speech analyzed in this thesis), all primary variables are relevant for perceived fluency, with the most influential ones being *rate of speech* and *number of silent pauses per minute.* They concluded that both human and automatically calculated fluency scores can be employed for fluency assessment, but that the selection and interpretation of variables should take into account the type of speech task and the type of speech.

Research by Yarra et al. explored the possibilities of automatic identification of native (L1) and non-native (L2) English speakers from eleven L1 backgrounds [17]. The study was performed with data from the ETS speech corpus, containing over 50 hours of spontaneous speech from 4099 non-native English speakers. Three different types of features were computed based on ASR decoded text: acoustic and segmental features, prosodic features and topic-model features, and baseline OpenSMILE features. For example for acoustic features, nativity discriminating factors mentioned are phoneme specific variations (e.g. phoneme /w/ is typically replaced by /v/ in Indian English) and speaking rate. Likewise for prosody based features, every language has its own speaking patterns and the voice of L1 speakers speaking another language is influenced by their native prosodic patterns. An SVM classifier with linear kernel was used for identifying 11 nativities. They concluded that performance of a classifier based on acoustic and prosodic feature selection strategy improves over the baseline technique of openSMILE features. Hence, it is important to keep in mind that acoustic and prosodic features can be important for characterizing an L2.

A subgroup of features from eGeMAPS are the Mel-Frequency-Cepstral-Coefficients (MFCC), which are closely related to spectral shape parameters [3]. Several studies on automatic feature selection, such as Schuller et al. [24], suggest that lower order MFCCs are more characteristic for affect and paralinguistic voice analysis tasks than higher order MFCCs. Paralinguistics (including prosody) refers to the non-lexical part of communication such as volume, speed, intonation and other non-verbal cues. For the comparison made in my research, the difference is expected to be in these paralinguistic aspects, which in turn will have an effect on the lower order MFCCs.

In short, selected features for non-native speech not only differ per atypical speech, but also per language and speech task. Atypical speech from languages other than English has not been analyzed much yet, also owing to the lack of useful speech data. My research will contribute to gaining more knowledge on Dutch atypical speech by analyzing non-native speech from JASMIN.

2.3 Research question and hypotheses

Based on aforementioned literature, hypotheses can be formulated based on the research question: which automatically extracted acoustic features are characteristic for non-native read Dutch speech (compared to native speech)? There are three hypotheses:

1. Features related to loudness are characteristic for non-native Dutch read speech, because non-natives have less confidence in reading. This lower confidence level will result in a lower reading volume.
2. Features related to rate of speech are characteristic for non-native Dutch read speech, since non-natives generally have a slower reading pace.
3. Lower order MFCCs are characteristic for non-native Dutch read speech, since paralinguistic factors such as volume and intonation are likely to be different between native and non-native speech.

Chapter 3

Methods and Materials

The pipeline used in this research is based on previous research by Van Bemmel [11]. A visualization of the altered pipeline is shown in Figure 3.1.

The pipeline contains six different steps:

1. Speech recording and transcribing the recordings
2. Using a Forced Aligner to align the data on word level
3. Extracting Praat and eGeMAPS features
4. Running Recursive Feature Elimination (RFE) to obtain a ranking of features
5. Running an SVM classifier to evaluate the RFE ranked features
6. Performing statistical tests to obtain significance and effect scores for all features

Step (5) and (6) will result in a ranked subset of characteristic features; (5) based on RFE and (6) based on effect size. We draw our conclusions by comparing these two ranked subsets. The separate steps are explained in more detail below.

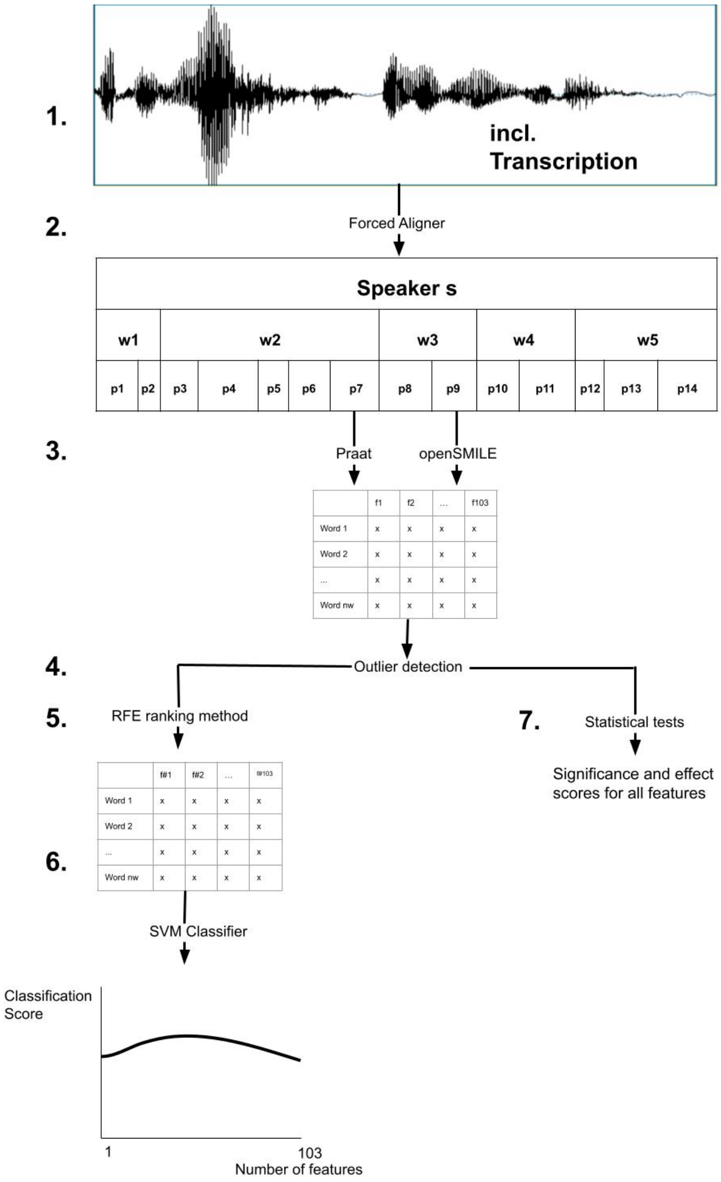
3.1 Data

The data used for this research comprise two data sets of adult read speech: JASMIN [1] for non-native Dutch speech, and for reference Dutch speech the reading of the story "De Koning" by Dutch natives. Different text is read by the two groups, but the stories are phonetically rich and can hence be used for comparison. Table 3.1 gives a short overview of the data, and a more detailed overview per dataset is given below.

3.1.1 JASMIN

The non-native adult read speech section from JASMIN comprises speech recordings and additional annotations of texts from *Code 1* and *Code 2* from Thieme Meulenhoff Publishers, a widely used method for learning Dutch as a second language. These phonetically rich texts are selected as to be suitable for learners of levels A1 and A2. The data contains speech recordings of 45 non-native adults, which were recruited through language schools that teach Dutch classes for foreigners. The language spoken at home and country of origin vary greatly, giving a realistic reflection the situation in the L2 Dutch classes. Other important factors included for the non-native adults are gender, age (ranging from 18 to 60), birth place, proficiency level in Dutch (ranging from A1 to B2), length of stay and time spent on learning Dutch.

Figure 3.1: A visualization of the methods of this thesis. In this example, there are nw total words. Altered version of pipeline visualization in Van Bemmel's research [11].



The speech recordings are in .wav format at 16 kHz sampling frequency. All speech recordings were orthographically transcribed (included in the corpus in .TextGrid format) according to the same conventions adopted in the Spoken Dutch Corpus (CGN) by using the Praat program. Each speech recording was spell-checked and double checked by a trained transcriber. Since this corpus also contains speech by non-native speakers, special conventions were required, for instance, for transcribing words realised with non-native pronunciation. For a more detailed account of JASMIN, please refer back to [1].

3.1.2 Native read speech

This dataset contains read speech from 30 native Dutch adults aged between 30 and 39 years. The gender is balanced (15 women, 14 men). Little other demographic information was available for the speakers, but this cannot be changed due to the retrospective nature of the experiment. The recordings were made at Radboud UMC in 2019 and contain the out loud reading of the phonetically balanced Dutch text “de Koning” [10]. In total the story contains 289 words on average per speaker. Similar to the JASMIN data, speech recordings are in .wav format at 16 kHz sampling frequency, and each recording has a manual orthographic transcription in .TextGrid format.

Table 3.1: Overview of the data

	Nr of read words		**Type of read speech**	**Nr of speakers**
	Original	After outlier removal		
JASMIN	42267	32345	Non-native	45
readings from “De Koning”	15775	11118	Native	30

3.2 Data pre-processing

Thanks to previous research by Van Bemmel [11], forced aligned files with corresponding features for the native data were already available. The same procedure will be applied to the JASMIN data, which will be discussed in more detail in the following sections. For further details on pre-processing and feature extraction of native data, please refer back to [11].

A Forced Aligner (Centre for Language Studies, https://webservices.cls.ru.nl/forcedalignment2) will be used to create a word and phoneme level segmentation. For this research, only word level segmentations will be used. Please note that the forced aligned files do not have perfect boundaries, especially with regard to some beginnings of sentences. However, found errors were not taken out manually, since it is not possible to check all data and the purpose of this research is to find out if it is possible to *automatically* extract characterizing features. Through later outlier detection and removal of missing feature values, hopefully most forced aligner flaws will be removed. As shown in Table 3.2, there was a striking difference in number of filler words in the two samples; non-natives use filler words more because of their lower reading proficiency. Filler words are not part of the prompt, but do effect feature averages like duration. Therefore, filler word instances ‘uh’, ‘uhm’, ‘eh’ and ‘ehm’ were removed from both datasets to account for this difference.

Table 3.2: Number of removed filler words per group.

	Uh	Uhm	Eh	Ehm	Total
Non-native	946	24	1	1	972
Native	0	2	0	0	2

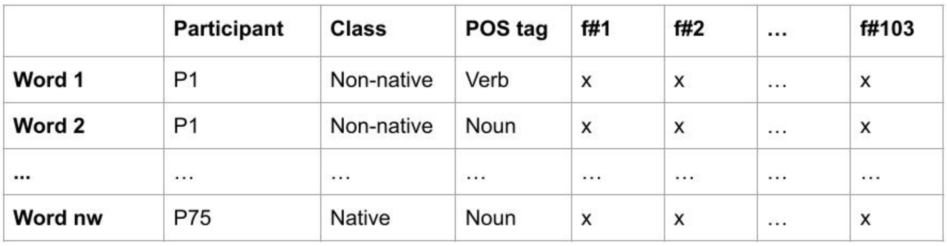
3.3 Acoustic features

Feature extraction will be performed using both Praat [9] and openSMILE [4]. Both tools take two inputs per participant: the audio file and the matching TextGrid file with the orthographic transcription. For this research, they output all features on word level: Praat results in 15 features, and openSMILE in 88 features.
A script made by Kerkhoff [12] will be used to extract Praat features for each word. These features include duration, pitch (minimum, maximum, mean, standard deviation and variability), intensity (minimum, maximum, mean and standard deviation), formants (f1, f2, f3, f4) and the centre of gravity. The script leaves out segmentations that do not contain any words. The resulting Praat features are organized and converted to an excel file using a script by Van Bemmel [15].

The remaining 88 features come from the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). This feature set was put together with the aim of providing a common baseline for evaluation of future research and eliminating differences caused by varying parameter sets. The parameters were selected based on their potential to signal affective physiological changes in voice production, on their shown value in previous research, on their automatic extractability and theoretical significance [3]. The feature set contains frequency related, energy related, spectral and temporal parameters. For further details on these features, please refer back to [3]. eGeMAPS features can be extracted using the publicly available openSMILE toolkit [4]. A script made by Van Bemmel [13] will be used to compute eGeMAPS features on word level.

All 103 features are combined into one word level feature file using a script from Van Bemmel [14]. Participant information and Part Of Speech tags are also added to the feature file. Figure 3.2 shows an overview of the format of the feature file.

Figure 3.2: Schematic overview of the feature file. In this example, there are nw total words. f#1 stands for the first feature, f#2 for the second feature, et cetera.



	Participant	**Class**	**POS tag**	**f#1**	**f#2**	**...**	**f#103**
Word 1	P1	Non-native	Verb	x	x	...	x
Word 2	P1	Non-native	Noun	x	x	...	x
...	...	...	...	...	...	...	...
Word nw	P75	Native	Noun	x	x	...	x

3.4 Outlier detection

Since automatic feature extraction is based on forced aligned speech recordings, not all boundaries are aligned perfectly, and thus not all feature values are calculated correctly. To account for this inaccuracy, a multivariate outlier detection procedure will be performed. A script made by Kwee [16] will be used at Chi-square cutoff probability 0.95 on Mahalanobis distances (MD). The Mahalanobis distance is frequently used for multivariate outlier detection, which detects unusual combinations of two or more variables (in this case, 103 variables). More specifically, in order to find outliers by Mahalanobis distance, distances between every point and the center in n-dimensional data are calculated. The distance between two points is calculated by considering how many standard deviations they are away from each other. So based on the calculated MDs, the numbers that fall above the Chi-square cutoff score for a Chi-square of 0.95 with 103 degrees of freedom (= number of extracted features) will be identified. These identified outliers and words containing one or more empty values (NaN) are removed from the feature file.

3.5 Recursive Feature Elimination and Classification

The first method to reduce the number of features and find a characterizing set of features is Recursive Feature Elimination (RFE) based on Support Vector Machine (SVM) classification with a linear kernel. This method was first used by Guyon et al. [22], and is now widely used in many fields of research, including the pipeline used by Van Bemmel [8]. RFE is a backward elimination method; starting out with the entire feature set, one feature is deleted iteratively until only one feature is left. This results in a list of ranked features. Backwards feature elimination takes into account the combined effect of features, and iteratively eliminates the feature that contributes least to the combined effect.

For evaluating the obtained feature ranking of RFE, an SVM classifier with a linear kernel was chosen. Note that classification on itself is not the goal here, it is merely meant to finding a group of most characterizing features. A linear kernel is used instead of a more complex kernel to reduce the risk of overfitting. Leave-One-Subject-Out (LOSO) cross-validation is used within classification to avoid overfitting on small datasets and to prevent identity-confounding [23]. The data is normalized per test and train set as preprocessing for classification. The Matthews Correlation Coefficient (MCC) will be used as evaluation measure instead of the standard accuracy measure. MCC is mostly employed when there is imbalanced data to prevent overfitting on one class. As shown in Table 3.1, this data contains imbalanced speech classes. The MCC is a score ranging from -1 (poor classification) to 1 (perfect classification). It is calculated using True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives(FN) from the binary classification. The MCC equation is shown below in Equation 3.1 .

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{3.1}$$

3.6 Statistical analysis

The second method to find characterizing features is statistical analysis. The statistical analysis is carried out via Welch's t-test and effect size (Cohen's distance). The Welch's t-test is an adaptation of the independent samples t-test. It is recommended over a standard independent t-test when the variances are not homogeneous and sample sizes are unequal [21]. The heterogeneity of variances is compensated for by adjusting the degrees of freedom [20]. Welch's t-test will indicate which features are significantly different, but it does not show the size of this difference. Therefore, effect size will be calculated for all significantly different features; effect size is an objective and standardized measure of the magnitude of the observed effect [19]. Cohen's distance (Cohen's d) calculates the difference between two means divided by a pooled standard deviation for the data.

$$d = \frac{(u_1 - u_2)}{s} \tag{3.2}$$

$$s = \sqrt{\frac{(n_1 - 1) \cdot s_1 + (n_2 - 1) \cdot s_2}{n_1 + n_2 - 1}} \tag{3.3}$$

In Equation 3.2, d is Cohen's distance, u_1 is the mean of the first sample, u_2 is the mean of the second sample and s is the pooled standard deviation of both samples. Equation 3.3 gives the formula for the pooled standard deviation for unequal sample sizes. In this Equation, n_1 and n_2 are the size of the first and second sample respectively, and s_1 and s_2 are the variance for the first and second sample. The subtractions serve as adjustments for the degrees of freedom.

The significantly different features will be ranked according to highest effect size (Cohen's d rank), and this ranking will be compared to the RFE ranking. Given that none of these two methods explicitly take into account possible correlations between the features, a correlation matrix for both rankings will be computed to see how pairwise features correlate with each other.

Chapter 4

Results

First, I discuss the results of the statistical tests on the complete set of features, along with the resulting effect size ranking. Then the classification results on RFE ranked features are shown, along with the resulting RFE ranking. The highest pairwise correlation scores for both rankings are included, ending with box plots and summary statistics for both rankings.

The assumptions of the data for Welch's t-test are that the independent variable (IV) is categorical on at least two levels (i), and that the dependent variable (DV) is continuous (ii) and satisfies normality conditions (iii). Welch's t-test is recommended over a standard independent t-test when the variances are not homogeneous and sample sizes are unequal [21], which are both the case for these data. Homogeneity of variances is tested with Levene's test, results of which are shown in Appendix B; as visible in the table, this assumption has been violated by most features. Only the features in the rows marked in grey have homogeneous variances. Welch's t-test makes a correction when the homogeneity of variances assumption is violated, and works as a standard t-test when this assumption is not violated. Therefore, I have chosen to perform Welch's t-test on all features.

Assumption (i) holds as the independent variable has the two levels *Non-native* and *Reference*. All dependent variables are continuous and so assumption (ii) holds too. The normality assumption (iii) is violated by most dependent variables as tested with Shapiro-Wilk test. However, with large sample sizes it is very easy to get significant results from small deviations from normality. According to the Central Limit Theorem a distribution of many data points will always approach normality. So in this case, normality can be assumed even though the Shapiro-Wilk test states otherwise.

According to Welch's t-test, 93 of 103 features are significantly different (p-value < 0.05) for native versus non-native Dutch speech on word level. Appendix A shows the table with significant features ordered by descending effect size. The first 14 features from the table can be interpreted as having a very large effect size ($d > 0.8$), the following 18 features still have a large effect size ($0.5 < d < 0.8$), and only the last 22 features from the ranking have a small effect size ($d < 0.2$). Table 4.1 shows the top 14 effect sizes features, which have a Cohen's distance that is bigger than 0.8.

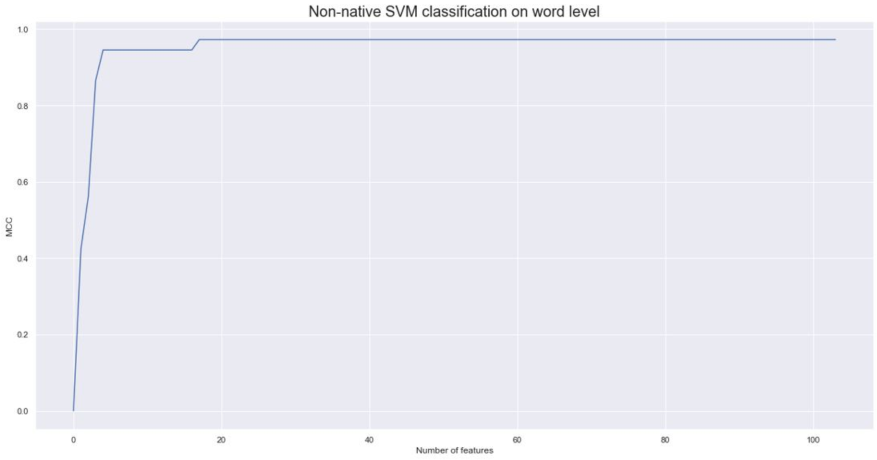
Figure 4.1 shows MCC classification scores per number of features for the ranked RFE feature set using an SVM. After the first seventeen features, the MCC score does not improve anymore, so this subset is used for further comparison. Table 4.2 shows this subset of seventeen features as ranked by RFE, together with their p-value and effect size. All features in this top 17 are significant as calculated by Welch's t-test. Table 4.3 shows the highest pairwise correlations of both the top 17 RFE and top 14 effect size features as calculated with Pearson's r.

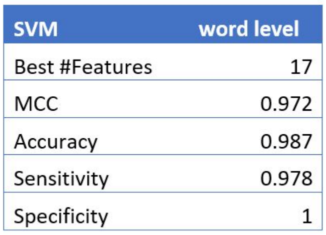
Boxplots and summary statistics for both top rankings are also included; Figure 4.2 and 4.3 show the boxplots of the top ranked features from RFE and effect size respectively. Figure 4.4 and 4.5 show the summary statistics of the top ranked features from RFE and effect size respectively.

Table 4.1: Top 14 features (significant, $p < 0.05$) for non-native Dutch speech ranked by effect size (Cohen's distance d), very large effect size ($d > 0.8$). Overlapping features with top 17 of RFE ranked features are marked in grey.

Feature	**t-value**	**p-value**	**Cohen's distance**
F1frequency_sma3nz_amean	128.37	0.00e+00	1.58
spectralFlux_sma3_amean	-87.23	0.00e+00	1.44
F3frequency_sma3nz_amean	115.70	0.00e+00	1.40
F2frequency_sma3nz_amean	117.53	0.00e+00	1.39
spectralFluxV_sma3nz_amean	-79.05	0.00e+00	1.30
spectralFluxUV_sma3nz_amean	-78.33	0.00e+00	1.27
mfcc2_sma3_amean	93.76	0.00e+00	1.05
f3	92.66	0.00e+00	1.01
f2	87.00	0.00e+00	0.95
mfcc1V_sma3nz_amean	-77.05	0.00e+00	0.90
loudness_sma3_percentile50.0	-59.713	0.00e+00	0.87
loudness_sma3_amean	-58.27	0.00e+00	0.87
slopeUV0-500_sma3nz_amean	58.04	0.00e+00	0.83
mfcc2V_sma3nz_amean	73.78	0.00e+00	0.83

Figure 4.1: Classification results with SVM on word level using the RFE ranking. The table shows classification scores using the lowest number of features that reaches the highest MCC score.





SVM	word level
Best #Features	17
MCC	0.972
Accuracy	0.987
Sensitivity	0.978
Specificity	1

Table 4.2: Top 17 features as ranked by RFE on word level. Overlapping features with top 14 of the effect size ranking are marked in grey.

RFE rank	Feature	p-value	Cohen's d	Cohen's d rank
1	spectralFlux_sma3_amean	0.00e+00	1.44	2
2	loudness_sma3_percentile80.0	0.00e+00	0.74	17
3	mfcc2_sma3_amean	0.00e+00	1.05	7
4	F1frequency_sma3nz_amean	0.00e+00	1.58	1
5	F0semitoneFrom27.5Hz_sma3nz_percentile50.0	0.00e+00	0.50	34
6	F0semitoneFrom27.5Hz_sma3nz_amean	0.00e+00	0.46	38
7	mfcc1V_sma3nz_amean	0.00e+00	0.90	10
8	mfcc1_sma3_amean	2.86e-259	0.42	43
9	hammarbergIndexUV_sma3nz_amean	0.00e+00	0.47	37
10	F3amplitudeLogRelF0_sma3nz_amean	1.18e-72	0.21	68
11	spectralFluxUV_sma3nz_amean	0.00e+00	1.27	6
12	hammarbergIndexV_sma3nz_amean	5.00e-59	0.18	76
13	alphaRatioV_sma3nz_amean	7.89e-139	0.27	61
14	mfcc1V_sma3nz_stddevNorm	8.63e-97	0.25	64
15	F1amplitudeLogRelF0_sma3nz_amean	5.67e-55	0.19	74
16	F1amplitudeLogRelF0_sma3nz_stddevNorm	1.33e-05	0.06	90
17	VoicedSegmentsPerSec	0.00e+00	0.67	23

Table 4.3: Highest pairwise correlations (Pearson's $r > 0.85$) for both rankings

	Feature 1	Feature 2	Pearson's r
RFE	F0semitoneFrom27.5Hz_sma3nz_percentile50.0	F0semitoneFrom27.5Hz_sma3nz_amean	0.99
	F3amplitudeLogRelF0_sma3nz_amean	F1amplitudeLogRelF0_sma3nz_amean	0.98
	F1amplitudeLogRelF0_sma3nz_amean	F1amplitudeLogRelF0_sma3nz_stddevNorm	0.88
Effect size	spectralFlux_sma3_amean	spectralFluxV_sma3nz_amean	0.93
	loudness_sma3_percentile50.0	loudness_sma3_amean	0.93
	mfcc2_sma3_amean	mfcc2V_sma3nz_amean	0.92

Figure 4.2: Boxplot of top 17 RFE features on word level. The boxes indicate the median and quartiles per group and feature.

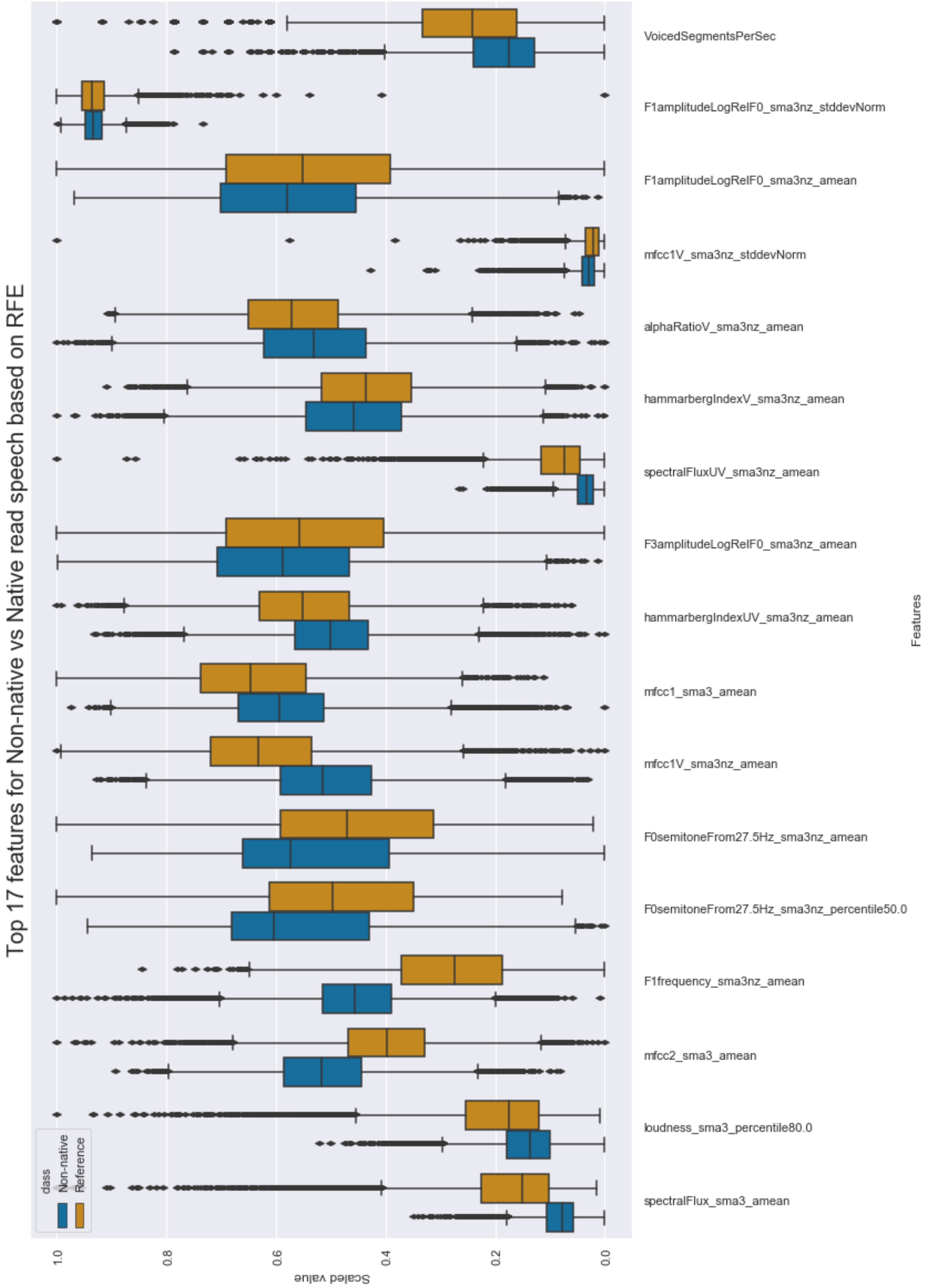
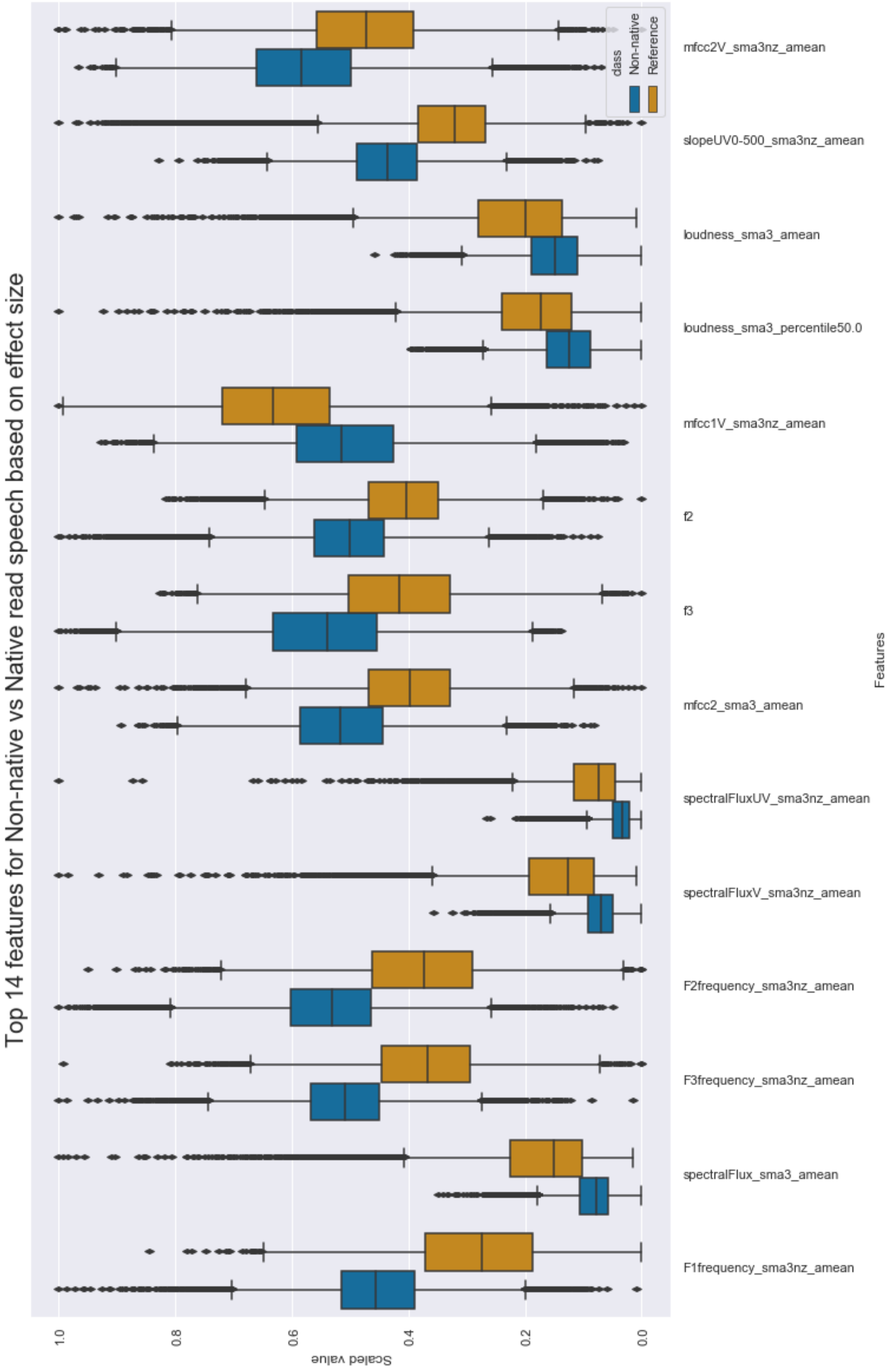


Figure 4.3: Boxplot of top 14 effect size (Cohen's distance) features on word level. The boxes indicate the median and quartiles per group and feature.



Feature (RFE Top 17)	Mean		Standard deviation		Median		25th Percentile		75th Percentile	
	Non-native	Native	Non-native	Native	Non-native	Native	Non-native	Native	Non-native	Native
spectralFlux_sma3_amean	0.24	0.48	0.10	0.29	0.22	0.41	0.17	0.28	0.29	0.60
loudness_sma3_percentile80.0	0.92	1.25	0.34	0.64	0.88	1.10	0.68	0.80	1.12	1.55
mfcc2_sma3_amean	12.85	3.74	8.53	8.94	13.16	3.57	7.34	-2.09	18.77	9.29
F1frequency_sma3nz_amean	705.50	554.04	88.79	112.99	710.35	544.66	648.75	464.44	764.23	633.52
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	32.12	29.58	4.96	5.57	33.46	29.91	27.70	25.02	36.01	33.75
F0semitoneFrom27.5Hz_sma3nz_amean	31.92	29.57	4.92	5.52	33.16	29.96	27.53	24.99	35.86	33.72
mfcc1V_sma3nz_amean	26.92	32.81	6.32	7.16	27.37	33.45	22.90	28.43	31.30	37.87
mfcc1_sma3_amean	21.43	24.50	7.02	8.28	21.99	25.12	17.09	19.05	26.41	30.50
hammarbergIndexUV_sma3nz_amean	18.38	22.25	7.76	9.33	18.68	22.41	13.47	16.12	23.58	28.46
F3amplitudeLogRelF0_sma3nz_amean	-92.62	-99.32	29.95	34.79	-91.28	-97.08	-113.30	-124.47	-70.10	-72.89
spectralFluxUV_sma3nz_amean	0.15	0.35	0.10	0.26	0.13	0.29	0.08	0.18	0.19	0.45
hammarbergIndexV_sma3nz_amean	24.66	23.61	5.91	5.86	24.61	23.61	20.57	19.71	28.68	27.38
alphaRatioV_sma3nz_amean	-16.33	-14.90	5.48	5.06	-16.26	-14.56	-20.07	-18.03	-12.54	-11.38
mfcc1V_sma3nz_stddevNorm	0.28	0.23	0.18	0.21	0.24	0.18	0.16	0.10	0.35	0.31
F1amplitudeLogRelF0_sma3nz_amean	-85.44	-91.82	32.65	38.41	-84.01	-89.57	-108.42	-120.35	-60.66	-62.63
F1amplitudeLogRelF0_sma3nz_stddevNorm	-1.02	-1.04	0.31	0.47	-0.99	-0.98	-1.21	-1.28	-0.81	-0.73
VoicedSegmentsPerSec	3.53	4.82	1.69	2.47	3.19	4.41	2.33	2.94	4.35	6.06

Figure 4.4: Summary statistics for the top 17 RFE features

	Mean		Standard deviation		Median		25th Percentile		75th Percentile	
Feature (Effect Size Top 14)	Non-native	Native	Non-native	Native	Non-native	Native	Non-native	Native	Non-native	Native
F1frequency_sma3nz_amean	705.50	554.04	88.79	112.99	710.35	544.66	648.75	464.44	764.23	633.52
spectralFlux_sma3_amean	0.24	0.48	0.10	0.29	0.22	0.41	0.17	0.28	0.29	0.60
F3frequency_sma3nz_amean	2680.95	2515.69	111.46	135.70	2682.14	2510.01	2610.79	2421.55	2752.71	2604.51
F2frequency_sma3nz_amean	1708.33	1518.20	130.45	152.48	1703.57	1512.01	1622.47	1411.90	1788.52	1620.45
spectralFluxV_sma3nz_amean	0.29	0.58	0.13	0.37	0.27	0.48	0.20	0.32	0.36	0.73
spectralFluxUV_sma3nz_amean	0.15	0.35	0.10	0.26	0.13	0.29	0.08	0.18	0.19	0.45
mfcc2_sma3_amean	12.85	3.74	8.53	8.94	13.16	3.57	7.34	-2.09	18.77	9.29
f3	3998.42	3590.85	402.45	399.31	3981.00	3601.00	3719.00	3327.00	4272.00	3865.00
f2	2934.63	2600.21	352.10	348.78	2912.00	2574.00	2707.00	2375.00	3131.00	2799.00
mfcc1V_sma3nz_amean	26.92	32.81	6.32	7.16	27.37	33.45	22.90	28.43	31.30	37.87
loudness_sma3_percentile50.0	0.54	0.78	0.21	0.39	0.53	0.71	0.40	0.51	0.67	0.96
loudness_sma3_amean	0.60	0.83	0.20	0.40	0.58	0.75	0.46	0.55	0.72	1.02
slopeUV0-500_sma3nz_amean	-0.01	-0.03	0.02	0.03	-0.01	-0.03	-0.02	-0.05	0.00	-0.02
mfcc2V_sma3nz_amean	10.96	1.97	10.79	11.17	11.67	1.81	4.06	-5.52	18.55	9.31

Figure 4.5: Summary statistics for the top 14 effect size features

Chapter 5

Discussion

5.1 Conclusions

The aim of this thesis was to find out which automatically extracted acoustic features are characteristic read speech by non-native Dutch speakers. I succeeded in automatically extracting 103 acoustic features on word level using Praat and openSMILE. Using Recursive Feature Elimination and subsequent evaluation using a Support Vector Machine classifier, I obtained the first set of characteristic features. Performing Welch's t-test and making effect size rankings resulted in a second set of characteristic features. These two rankings will be discussed and compared in more detail below.

Overall, the ranking based on effect size overlapped partly with the RFE ranking, but there were some big differences too. These differences can be explained by the fact that both methods have completely different roots; RFE is a machine learning approach, whereas effect size is a statistical approach. RFE takes into account the combined effect of features, whereas the statistical analysis looks only at individual effects of features. Correlation scores were added to give more insight into how different features influence each other. It is interesting to compare how both rankings overlap, but it is no surprise that they show big differences. It is hard to say whether statistical tests or or a classification based ranking are more expressive for characteristic acoustic features. Individual statistically significant features allow for easier interpretation of their effect, whereas for classification one has to look at their combined effect. This is more difficult to interpret compared to individual features. Below is a comparison of the two rankings.

The difference between non-native and native speech from 93 out of the initial 103 acoustic features is found to be significant by Welch's t-test. The SVM resulted in almost perfect classification (MCC = 0.972) after using the subset of 17 features from the RFE ranking. These two findings indicate that the non-native acoustic word features are clearly different from the native features. As visible in the graph of Figure 4.1, there is a steep increase of MCC for the first four subsets of features, after which the MCC score stagnates until the subset containing 17 features. When comparing this to "Cohen's d rank" in Table 4.2, we can see that the first four features in the RFE ranking are relatively high in the effect size ranking compared to the later features. The last feature, 'VoicedSegmentsPerSec', then has a relatively high Cohen's d ranking again, which is in line with the final increase of MCC score to 0.972. Using subsets of more than 17 features for classification did not improve the MCC score anymore (as can be seen in Figure 4.1), indicating that this subset of 17 features is characteristic for this collection of non-native Dutch read speech. However, with a small dataset and high MCC scores like this, it is clear that overfitting occurred.

The selection made by RFE and linear SVM classifier contained significantly different features only. The features ranked high in the effect size ranking do not fully overlap the RFE ranking, as just explained above. However, 'loudness_sma3_percentile80' was high in ranking for both methods, and other loudness related features (such as 'slopeUV0-500_sma3nz_amean', 'intensity_mean', 'loudness_sma3_percentile50', 'loudness_sma3_amean', 'intensity_max') showed very large effect sizes too. Specifically, 'hammarbergIndex' is in the top 17 twice and also characteristic for loudness; it is less sensitive to recording differences (e.g. distance of the speaker to microphone) than other loudness-related features. Boxplots in Figure 4.2 and Figure 4.3 show that non-natives have a lower speaking volume. This confirms the first part of my hypothesis; non-native speakers read with a lower volume, which could indicate lower confidence.

I also hypothesized that speaking rate is lower for non-native speech compared to native speech. Since all features are measured at word level, since the used feature set does not contain many speech rate features and since prompts are different for both groups, it is difficult to make conclusions about this aspect. Interestingly, ‘spectralFlux_sma3_amean’ is high in both rankings, and there are also two other ‘spectralFlux’ features in the top effect size ranking. This feature is based on the difference of the spectra of two consecutive frames. A higher speaking rate generally results in more spectral fluctuations. As visible in the boxplots in Figure 4.3, natives have a higher value for this feature compared to non-natives, which is an indication for a higher speech rate. The feature ‘VoicedSegmentsPerSec’ is also indicative of speech rate; the boxplot in Figure 4.2 shows that non-natives have a lower value, indicating a lower speech rate. These results are indications that speech rate for non-natives could be lower, but these results are not conclusive for accepting this second hypothesis.

Although they are in neither of the two top rankings, the features ‘VoicedSegmentsPerSec’ and ‘dur’ (duration) show large effect sizes, and boxplots show that word durations are indeed longer for non-native Dutch speakers. These three features are indications that speech rate for non-natives could be lower, but these results are not conclusive for accepting this hypothesis.

The third hypothesis stated that lower order MFCCs are characteristic for non-native Dutch read speech. This hypothesis can be accepted; both ‘mfcc1V_sma3nz_amean’ and ‘mfcc2_sma3_amean’ are high up in the two rankings. This can be interpreted as that paralinguistic speech factors, such as intonation, are significantly different for non-natives compared to natives.

Although not hypothesized, formant related features are high in both rankings; ‘F1frequency_sma3nz_amean’ is at rank 1 and 4 of the effect size ranking and RFE ranking respectively. The RFE top 17 also contains ‘F1amplitudeLogRelF0_sma3nz_amean’ and ‘F1amplitudeLogRelF0_sma3nz_stddevNorm’ (related to the first formant), and ‘F3amplitudeLogRelF0_sma3nz_amean’ (related to the third formant). The effect size top 14 contains additional features ‘F3frequency_sma3nz_amean’ and ‘f3’ (related to the third formant), and ‘F2frequency_sma3nz_amean’ and ‘f2’ (related to the second formant). The difference of formant related features between non-natives and natives makes sense; formants are the result of vibrating air in the vocal tract. This air vibrates at different frequencies depending on the vocal tract opening’s shape and size. These frequencies are called formants, and they can be changed by changing the shape and size of the vocal tract. The first formant (F1) responds to the degree to which the mouth is open during speech (a higher F1 indicates more open mouth/jaw), whereas the second formant (F2) is usually associated with the place where the tongue creates a narrowing in the vocal tract (a higher F2 means tongue more to the back). The third formant (F3) is especially influenced by lengthening of space between lips and teeth which is accomplished by rounding the lips (which leads to a lower F3) [25]. A general explanation for the difference in formant related features for non-natives compared to natives could be that non-natives have a deviating pronunciation, which must be associated with a different shape and size of the vocal tract.

Table 4.3 shows that there are highly correlated features in both rankings. This is not surprising given that for both eGeMAPS and Praat there are multiple features related to the same acoustic principle. For example, the first correlated features in the table both are related to ‘F0’, also referred to as the fundamental frequency. The only two highly correlated features that are not about the exact same principle are ‘F3amplitudeLogRelF0_sma3nz_amean’ and ‘F1amplitudeLogRelF0_sma3nz_amean’ ($r = 0.98$). However, as explained above, the first formant and the third formant are both influenced by the position and movement of the mouth, which accounts for their high correlation. When defining a set of characteristic features for non-native speech, it would only be necessary to select one feature out of each of these highly correlated feature pairs.

To summarize, I succeeded in automatic extraction of features for non-native Dutch read speech. Through binary comparisons with native speech, 93 out of 103 features were found to be significantly different. Through binary comparisons with native speech, 93 out of 103 features were found to be significantly different. Two characteristic and partly overlapping sets of features were found; the first set based on the RFE ranking, the second based on an effect size ranking. Both sets support the hypotheses that a lower speaker volume and lower order Mel-Frequency-Cepstral-Coefficients are characteristic for non-native Dutch speech, and show indications of a slower reading rate for non-natives. Moreover, formant related features were prevalent in both rankings, indicating a different shape of the vocal tract owing to deviations in non-native pronunciation compared to native speakers.

5.2 Limitations

As with all research, this thesis has several limitations. First of all, non-native speech data for Dutch language is limited. For the JASMIN corpus, it comprises 45 speakers with varying backgrounds and proficiency levels. Previous literature suggested that characteristic features for non-native speech depend on their first language and proficiency. Unfortunately, due to the limited number of speakers and the fact that mother tongue was not included in JASMIN corpus details, it was not feasible to make a distinction or comparison between these groups. This would be an interesting addition for future research when more non-native data is available. The restricted amount of data makes the classification prone to overfitting; to reduce overfitting, a simple linear SVM classifier was chosen with LOSO-cross validation. However, overfitting is inevitable, and with more data also a more sophisticated classifier could have been used.

Another limitation to this research is the availability of comparable native read speech. Due to time limitations, I opted for the already analyzed readings of “De Koning”. Although this story is phonetically balanced, it differs from the JASMIN prompts. In addition to word level feature extraction and analysis, I also performed it on speaker level (meaning every feature is extracted based on the full recording of one speaker). However, classification based on RFE ranking resulted in perfect classification with a subset of two features already, and effect sizes were very big for the majority of the features. This can be explained by the fact that some full level eGeMAPS features are highly prompt specific, and therefore comparison between these two groups on full level is not informative. For future research, comparable read speech could be searched for in other native corpuses like CGN, or the same JASMIN prompts could be read by native Dutch speakers for better comparison. It would be interesting to see if the found characteristic features then differ from the ones in this research.

5.3 Future research

In this thesis, I performed exploratory research on automatic extraction of characteristic features for non-native Dutch read speech. In addition to those mentioned in the Limitations section, there are several interesting options for further research. For example, it would be interesting to compare spontaneous non-native Dutch speech to reference Dutch speech; JASMIN contains recordings from non-native human-machine interaction that would be suitable for such research.

Furthermore, once speech intelligibility (SI) ratings are obtained from a listening experiment of JASMIN non-native read speech, the obtained objective features from this work can be used to get insight in human intelligibility of non-native Dutch speech. Using Automatic Speech Recognition (ASR) systems a comparison can be made between the actual read speech and the ASR output to obtain ’computer intelligibility ratings’ [6]. The obtained features from this work can then be used to get more insight into characteristic features for non-native ASR purposes; this could be used to improve ASR application for the subgroup of non-native Dutch speakers.

On a different note, by using the already available Part Of Speech tags, it would be possible to compare features of content and function words for non-native speakers. Due to the lack of speech rate features in the feature set used in this thesis, it was not possible to draw conclusions on speech rate. De Jong and Wempe have developed a method in Praat for automatically detecting syllable nuclei in order to measure speech rate [26]. Adding this feature to the feature set will lead to more insight into speech rate for atypical speakers.

Bibliography

[1] Cucchiarini, C., Driesen, J., Van Hamme, H., and Sanders, E. (2008). Recording speech of children, non-natives and elderly people for HLT applications: The JASMIN-CGN corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pp. 1445–1450. European Language Resources.

[2] Cucchiarini, C., Van hamme, H. (2013). The JASMIN Speech Corpus: Recordings of Children, Non-natives and Elderly People. In *Spyns P., Odijk J. (eds) Essential Speech and Language Technology for Dutch. Theory and Applications of Natural Language Processing.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_3

[3] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., ... Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), pp. 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[4] Eyben, F., Wöllmer, M. and Schuller, B (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462.

[5] Eynde, F. (2004). Part of speech tagging en lemmatisering van het corpus gesproken nederlands.

[6] Elffers, B., Van Bael, C., and Strik, H. ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions. Internal report, Department of Language Speech, University of Nijmegen.

[7] Ganzeboom, M., Bakker, M., Cucchiarini, C., and Strik, H. (2016). Intelligibility of Disordered Speech: Global and Detailed Scores. 2503-2507. 10.21437/Interspeech.2016-1448.

[8] Van Bemmel, L., Harmsen, W., Cucchiarini, C., and Strik, H. (2021). Automatic selection of the most characterizing features for detecting COPD in speech.

[9] Boersma, P., Weenik, D. (2020). Praat: doing phonetics by computer (version 6.1.16). http://www.praat.org.

[10] Haasnoot, R. (2012). De ontwikkeling van de fonetisch uitgebalanceerde standaardtekst “de koning”.

[11] Van Bemmel, L. (2021). Automatic selection of the most characterizing features for different types of atypical speech.

[12] Kerkhoff, J. (2015). “Labeledsegmentanalysis.praat” (script).

[13] Van Bemmel, L. (2021). “ExtractingFeatures_opensmile.py” (script).

[14] Van Bemmel, L. (2021). “combining_praat_gemaps.py” (script).

[15] Van Bemmel, L. (2021). “organizing_Praatfeatures.py” (script).

[16] Kwee, V. (2021). “Outlierdetection 3.7.1.R” (script).

[17] Yarra, C., Rao, A., and Ghosh, P. K. (2018). Automatic native language identification using novel acoustic and prosodic feature selection strategies. In *2018 15th IEEE India Council International Conference (INDICON)*, pp. 1–6, IEEE.

[18] Cucchiarini, C., Strik, H., and Boves, L. (2002). Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech. In *the Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873.

[19] Field, A., Miles, J., and Field, Z. (2012). Discovering statistics using R. SAGE Publications.

[20] Rietveld, T., Van Hout, R. (2015). The t test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology. In *Journal of Communication Disorders, 58*, pp. 158-68.

[21] Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. In *British Journal of Mathematical Psychology, 57*, pp. 173–181.

[22] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. In *Machine learning*, vol. 46, no. 1, pp. 389–422.

[23] Neto, E. C., Pratap, A., Perumal, T. M., Tummalacherla, M., Snyder, P., Bot, ... Omberg, L. (2019). Detecting the impact of subject characteristics on machine learning-based diagnostic applications. In *NPJ digital medicine*, vol. 2, no. 1, pp. 1–6.

[24] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L. ... Aharonson, V. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc., Aug. 2007*, pp. 2253–2256.

[25] Bloothooft, G. (2008). *Spraakakoestiek*. Universiteit Utrecht. Retrieved January 10, 2022, from https://www.phil.uu.nl/tst/2012/Slides/Spraakakoestiek.pdf

[26] De Jong, N.H., Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. In *Behavior Research Methods 41*, pp. 385–390. https://doi.org/10.3758/BRM.41.2.385

Appendix A

Effect size ranking

Table A.1: Significant features (Welch's t-test, $p < 0.05$) for non-native Dutch speech ranked by effect size (Cohen's distance d).

Feature	t-value	p-value	Cohen's distance
F1frequency_sma3nz_amean	128.375	0.000	1.585
spectralFlux_sma3_amean	-87.230	0.000	1.439
F3frequency_sma3nz_amean	115.697	0.000	1.399
F2frequency_sma3nz_amean	117.526	0.000	1.394
spectralFluxV_sma3nz_amean	-79.051	0.000	1.301
spectralFluxUV_sma3nz_amean	-78.327	0.000	1.268
mfcc2_sma3_amean	93.756	0.000	1.054
f3	92.656	0.000	1.015
f2	87.003	0.000	0.952
mfcc1V_sma3nz_amean	-77.055	0.000	0.900
loudness_sma3_percentile50.0	-59.713	0.000	0.868
loudness_sma3_amean	-58.271	0.000	0.866
slopeUV0-500_sma3nz_amean	58.043	0.000	0.833
mfcc2V_sma3nz_amean	73.782	0.000	0.825
equivalentSoundLevel_dBp	-61.220	0.000	0.758
intensity_mean	-59.963	0.000	0.745
loudness_sma3_percentile80.0	-51.025	0.000	0.744
loudness_sma3_meanRisingSlope	-53.289	0.000	0.743
slopeV0-500_sma3nz_amean	58.577	0.000	0.726
pitch_var	-41.709	0.000	0.709
loudness_sma3_percentile20.0	-47.452	0.000	0.695
intensity_min	-57.011	0.000	0.684
intensity_max	-53.440	0.000	0.674
VoicedSegmentsPerSec	-51.234	0.000	0.674
dur	70.744	0.000	0.668
HNRdBACF_sma3nz_amean	59.466	0.000	0.665
loudnessPeaksPerSec	-46.540	0.000	0.601
MeanVoicedSegmentLengthSec	58.741	0.000	0.576

Continued on next page

Table A.1: Significant features (Welch's t-test, $p < 0.05$) for non-native Dutch speech ranked by effect size (Cohen's distance d).

Feature	**t-value**	**p-value**	**Cohen's distance**
jitterLocal_sma3nz_stddevNorm	52.889	0.000	0.572
pitch_std	-33.353	0.000	0.548
F3frequency_sma3nz_stddevNorm	46.647	0.000	0.536
loudness_sma3_pctlrange0-2	-35.465	0.000	0.503
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	42.514	0.000	0.498
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	42.626	0.000	0.496
shimmerLocaldB_sma3nz_stddevNorm	41.633	0.000	0.490
F2bandwidth_sma3nz_stddevNorm	46.278	0.000	0.485
hammarbergIndexUV_sma3nz_amean	-39.224	0.000	0.472
F0semitoneFrom27.5Hz_sma3nz_amean	39.809	0.000	0.463
spectralFluxV_sma3nz_stddevNorm	39.297	0.000	0.453
F1bandwidth_sma3nz_stddevNorm	40.694	0.000	0.449
shimmerLocaldB_sma3nz_amean	-32.792	0.000	0.447
hammarbergIndexV_sma3nz_stddevNorm	37.952	0.000	0.433
mfcc1_sma3_amean	-35.007	0.000	0.417
F3bandwidth_sma3nz_stddevNorm	37.858	0.000	0.416
spectralFlux_sma3_stddevNorm	36.843	0.000	0.409
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	36.130	0.000	0.407
f1	38.643	0.000	0.402
loudness_sma3_stddevFallingSlope	-25.815	0.000	0.372
F1bandwidth_sma3nz_amean	-32.382	0.000	0.361
pitch_min	31.775	0.000	0.348
StddevVoicedSegmentLengthSec	40.059	0.000	0.338
logRelF0-H1-H2_sma3nz_amean	32.835	0.000	0.323
F2frequency_sma3nz_stddevNorm	25.507	0.000	0.308
mfcc4_sma3_amean	-26.816	0.000	0.297
StddevUnvoicedSegmentLength	30.665	0.000	0.295
pitch_max	-19.046	0.000	0.285
loudness_sma3_meanFallingSlope	-18.374	0.000	0.281
intensity_std	23.368	0.000	0.277
slopeV500-1500_sma3nz_amean	23.353	0.000	0.274
alphaRatioV_sma3nz_amean	-25.273	0.000	0.267
pitch_mean	22.755	0.000	0.267
jitterLocal_sma3nz_amean	-20.925	0.000	0.266
alphaRatioUV_sma3nz_amean	21.757	0.000	0.257
mfcc1V_sma3nz_stddevNorm	21.013	0.000	0.250
slopeUV500-1500_sma3nz_amean	-21.367	0.000	0.248
F2bandwidth_sma3nz_amean	-20.890	0.000	0.225
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	22.899	0.000	0.222
F3amplitudeLogRelF0_sma3nz_amean	18.114	0.000	0.214

Continued on next page

Table A.1: Significant features (Welch's t-test, $p < 0.05$) for non-native Dutch speech ranked by effect size (Cohen's distance d).

Feature	**t-value**	**p-value**	**Cohen's distance**
MeanUnvoicedSegmentLength	19.204	0.000	0.210
mfcc4V_sma3nz_amean	-18.987	0.000	0.207
alphaRatioV_sma3nz_stddevNorm	-25.074	0.000	0.206
F1frequency_sma3nz_stddevNorm	17.374	0.000	0.201
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	22.638	0.000	0.189
F1amplitudeLogRelF0_sma3nz_amean	15.672	0.000	0.186
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	18.159	0.000	0.183
F2amplitudeLogRelF0_sma3nz_amean	14.952	0.000	0.178
hammarbergIndexV_sma3nz_amean	16.256	0.000	0.178
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	20.583	0.000	0.157
mfcc3V_sma3nz_amean	12.215	0.000	0.136
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	14.355	0.000	0.136
logRelF0-H1-A3_sma3nz_amean	11.477	0.000	0.127
loudness_sma3_stddevRisingSlope	-9.204	0.000	0.119
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	14.977	0.000	0.118
loudness_sma3_stddevNorm	9.602	0.000	0.115
HNRdBACF_sma3nz_stddevNorm	-9.188	0.000	0.114
mfcc3_sma3_amean	9.965	0.000	0.110
f0	-9.043	0.000	0.098
F3amplitudeLogRelF0_sma3nz_stddevNorm	-7.652	0.000	0.096
logRelF0-H1-A3_sma3nz_stddevNorm	12.358	0.000	0.091
F1amplitudeLogRelF0_sma3nz_stddevNorm	4.358	0.000	0.058
mfcc1_sma3_stddevNorm	-4.534	0.000	0.043
mfcc2_sma3_stddevNorm	2.850	0.004	0.042
mfcc2V_sma3nz_stddevNorm	2.441	0.015	0.028

Appendix B

Levene’s test

Table B.1: Results from Levene’s test for all features. If Levene’s p-value < 0.5, the variances are significantly different and the assumption of homogeneity of variances has been violated. Only the features in the rows marked in grey have homogenous variances.

feature	**Levene’s statistic**	**Levene’s p**
spectralFlux_sma3_amean	12045.89	0.00
loudness_sma3_percentile80.0	4926.53	0.00
mfcc2_sma3_amean	7.87	0.01
F1frequency_sma3nz_amean	1387.21	0.00
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	212.78	0.00
F0semitoneFrom27.5Hz_sma3nz_amean	209.44	0.00
mfcc1V_sma3nz_amean	218.60	0.00
mfcc1_sma3_amean	498.20	0.00
hammarbergIndexUV_sma3nz_amean	562.78	0.00
F3amplitudeLogRelF0_sma3nz_amean	489.74	0.00
spectralFluxUV_sma3nz_amean	8766.63	0.00
hammarbergIndexV_sma3nz_amean	5.10	0.02
alphaRatioV_sma3nz_amean	126.63	0.00
mfcc1V_sma3nz_stddevNorm	50.43	0.00
F1amplitudeLogRelF0_sma3nz_amean	582.98	0.00
F1amplitudeLogRelF0_sma3nz_stddevNorm	1414.57	0.00
VoicedSegmentsPerSec	2256.06	0.00
f2	1.06	0.30
F2amplitudeLogRelF0_sma3nz_amean	571.05	0.00
mfcc3_sma3_amean	0.00	0.98
slopeV0-500_sma3nz_amean	562.74	0.00
mfcc2V_sma3nz_amean	13.21	0.00
F1frequency_sma3nz_stddevNorm	196.85	0.00
loudness_sma3_pctlrange0-2	3595.88	0.00
loudness_sma3_percentile20.0	5779.25	0.00
f3	4.57	0.03
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	313.77	0.00

Continued on next page

Table B.1: Results from Levene's test for all features. If Levene's p-value < 0.5, the variances are significantly different and the assumption of homogeneity of variances has been violated. Only the features in the rows marked in grey have homogenous variances.

feature	**Levene's statistic**	**Levene's p**
F0semitoneFrom27.5Hz_sma3nz_percentile20.0	28.16	0.00
dur	762.21	0.00
MeanVoicedSegmentLengthSec	429.60	0.00
pitch_var	11900.05	0.00
F3bandwidth_sma3nz_amean	342.75	0.00
spectralFlux_sma3_stddevNorm	4.30	0.04
spectralFluxV_sma3nz_stddevNorm	166.47	0.00
F3frequency_sma3nz_stddevNorm	164.17	0.00
F2frequency_sma3nz_stddevNorm	751.82	0.00
intensity_mean	1155.93	0.00
intensity_max	1531.71	0.00
loudness_sma3_meanRisingSlope	3010.39	0.00
alphaRatioUV_sma3nz_amean	295.92	0.00
slopeUV500-1500_sma3nz_amean	74.36	0.00
F2bandwidth_sma3nz_stddevNorm	126.41	0.00
F2bandwidth_sma3nz_amean	42.95	0.00
F2frequency_sma3nz_amean	496.30	0.00
mfcc4V_sma3nz_amean	16.72	0.00
mfcc4_sma3_amean	0.06	0.81
slopeV500-1500_sma3nz_amean	231.69	0.00
slopeUV0-500_sma3nz_amean	2479.31	0.00
loudness_sma3_amean	6131.02	0.00
loudness_sma3_percentile50.0	4564.12	0.00
intensity_std	481.11	0.00
loudness_sma3_stddevNorm	506.54	0.00
hammarbergIndexV_sma3nz_stddevNorm	79.80	0.00
HNRdBACF_sma3nz_amean	13.59	0.00
pitch_mean	121.74	0.00
pitch_std	10671.26	0.00
pitch_min	115.83	0.00
F3bandwidth_sma3nz_stddevNorm	2.35	0.13
StddevUnvoicedSegmentLength	1204.85	0.00
StddevVoicedSegmentLengthSec	2921.87	0.00
intensity_min	866.97	0.00
f0	13.08	0.00
loudness_sma3_stddevRisingSlope	1241.42	0.00
loudness_sma3_meanFallingSlope	4823.37	0.00
F3frequency_sma3nz_amean	682.55	0.00
F2amplitudeLogRelF0_sma3nz_stddevNorm	1415.71	0.00

Continued on next page

Table B.1: Results from Levene's test for all features. If Levene's p-value < 0.5, the variances are significantly different and the assumption of homogeneity of variances has been violated. Only the features in the rows marked in grey have homogenous variances.

feature	**Levene's statistic**	**Levene's p**
F3amplitudeLogRelF0_sma3nz_stddevNorm	1093.31	0.00
f1	127.01	0.00
logRelF0-H1-H2_sma3nz_amean	619.97	0.00
logRelF0-H1-A3_sma3nz_amean	0.57	0.45
equivalentSoundLevel_dBp	1093.33	0.00
shimmerLocaldB_sma3nz_amean	2186.87	0.00
pitch_max	6978.73	0.00
mfcc3_sma3_stddevNorm	21.58	0.00
shimmerLocaldB_sma3nz_stddevNorm	330.44	0.00
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	67.95	0.00
mfcc3V_sma3nz_amean	3.60	0.06
mfcc4V_sma3nz_stddevNorm	0.01	0.92
alphaRatioV_sma3nz_stddevNorm	25.24	0.00
jitterLocal_sma3nz_stddevNorm	8.65	0.00
jitterLocal_sma3nz_amean	809.77	0.00
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	379.01	0.00
F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	267.39	0.00
logRelF0-H1-A3_sma3nz_stddevNorm	17.91	0.00
loudnessPeaksPerSec	1168.92	0.00
slopeV500-1500_sma3nz_stddevNorm	8.02	0.00
F1bandwidth_sma3nz_amean	6.54	0.01
loudness_sma3_stddevFallingSlope	4886.62	0.00
mfcc1_sma3_stddevNorm	5.13	0.02
mfcc2_sma3_stddevNorm	269.76	0.00
spectralFluxV_sma3nz_amean	11607.87	0.00
F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	653.23	0.00
grav_center	65.08	0.00
F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	147.77	0.00
F1bandwidth_sma3nz_stddevNorm	3.70	0.05
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	101.00	0.00
mfcc2V_sma3nz_stddevNorm	22.94	0.00
mfcc4_sma3_stddevNorm	7.69	0.01
mfcc3V_sma3nz_stddevNorm	12.98	0.00
slopeV0-500_sma3nz_stddevNorm	142.00	0.00
HNRdBACF_sma3nz_stddevNorm	421.09	0.00
MeanUnvoicedSegmentLength	60.92	0.00
logRelF0-H1-H2_sma3nz_stddevNorm	3.30	0.07