Graz University of Technology

Master Thesis

Automatic Speech Recognition for dysarthric Speakers

conducted at the
Signal Processing and Speech Communication Laboratory

by
Susanne Rexeis

Supervisors:
Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin
Dipl.-Ing. Dr.techn. Stefan Petrik

Graz, Austria, October 2011

Abstract

Dysarthria is a speech impairment caused by neuro-muscular damages of various cause that can also lead to reduced dexterity or paralysis of other body parts, e.g. the limbs. For these patients the use of speech technology as interface to an environmental control system or to a computer can be a valuable assistance in everyday life. However, due to the various speaker-dependent disturbances typical for dysarthric speech the performance of standard automatic speech recognition (ASR)-systems is limited.

This work investigates different approaches to improve the performance of speech recognizers for German-speaking males suffering from moderate to severe dysarthria. The speech data was recorded in cooperation with the Simon project.

Evaluations on a small-vocabulary connected digits task showed that speaker-independent (SI) acoustic models adapted to dysarthric speech using maximum likelihood linear regression (MLLR) could achieve better results than speaker-dependent (SD) acoustic models for a patient suffering from severe dysarthria. For two out of the five dysarthric speakers word recognition rates of over 90% could be achieved using MLLR-adaptation. On a task using a larger vocabulary of 69 command words, however, only a maximum word recognition rate of 70% could be achieved using acoustic adaptation.

In the utterances of the dysarthric speakers mispronunciations of certain phonemes could be identified. Two data-driven approaches to adapt the pronunciation dictionaries of the recognition systems to dysarthric speech were proposed and evaluated: phonological rules and finite state transducer (FST) networks. The pronunciation errors of the speakers were modeled based on the evaluation of the speech recognizers on a rhyme-test. Lexical adaptation with phonological rules achieved promising results on the rhyme-test evaluation. In contrast the improvement of the recognition rate in the command word task was barely measurable, as a high number of new confusions occurred after adaptation. Two methods to prune the generated pronunciation variants based on their probability did not succeed to lower the number of confusions. Lexical adaptation with FSTs failed to improve results on both the rhyme-test and the command word task. The number of new recognition errors after adaptation was again very high, although a score to measure the confusability of the newly generated variants was used for pruning. The information extracted from the phone confusions of the rhyme-test seems to be too sparse to score the confusability of the new pronunciations correctly in this approach.

Kurzfassung

Dysarthrie ist eine Sprachstörung, die durch neuro-muskuläre Schädigungen hervorgerufen wird. Diese können auch zu Einschränkungen der Beweglichkeit oder gar zu Lähmungen ganzer Körperteile, z.B. der Extremitäten, führen. Für Dysathrie-Patienten kann die Verwendung von Sprachtechnologien, als Schnittstelle zu einfachen Regelungssystemen (z.B. der Heizung), speziell aber auch zur Steuerung von Computern, eine wertvolle Erleichterung im Alltagsleben bedeuten. Voraussetzung dafür sind eine einfache Bedienbarkeit und Zuverlässigkeit der Systeme. Durch die Sprachstörungen der Patienten ist die Verwendung von Standard Spracherkennern, wie sie heute in vielen elektronischen Geräten eingebaut sind, kaum möglich. Die Systeme müssen auf die Sprecher angepasst werden.

In dieser Arbeit wurden verschiedene Ansätze untersucht, um die Erkennungsrate von Spracherkennungssystemen für fünf männliche, deutschsprachige Dysarthrie-Patienten zu verbessern. Die Sprachdaten wurden vom Team des Simon Projektes zur Verfügung gestellt.

Die Evaluierung der Daten mit einer einfachen „Ziffernfolgen Erkennungsaufgabe“ zeigte, dass ein sprecherunabhängiges akustisches Modell mittels akustischer Adaption mit MLLR („maximum likelihood linear regression“) für einen Patienten mit schwerer Dysarthrie bessere Ergebnisse erzielen kann als ein mit Sprachdaten des Patienten trainiertes akustisches Modell. Für zwei der Sprecher wurde mittels akustischer Adaption eine Wort Erkennungsrate von über 90% erzielt. Bei einer zweiten Erkennungsaufgabe mit einem größeren Vokabular (69 Kommandoworte) konnte lediglich eine Wort Erkennungsrate von maximal 70% erzielt werden. Dabei blieben die adaptieren Systeme deutlich hinter den akustischen Modellen, die direkt mit den Sprachdaten der Sprecher trainiert wurden, zurück.

In den Sprachdaten der Dysarthrie-Patienten konnten Fehler in der Aussprache bestimmter Phoneme festgestellt werden. Zwei Ansätze, die diese aufbauend auf dem Erkennungsergebnis eines Reimtests modellieren und automatisiert im Aussprachemodell des Erkenners abbilden, wurden evaluiert. Im ersten Ansatz, basierend auf phonologischen Regeln, konnten vielversprechende Ergebnisse auf einer Auswertung erzielt werden. Bei den Kommandoworten konnte jedoch kaum eine Verbesserung der Erkennungsrate gemessen werden, da das neue Aussprachemodell zahlreiche neue Erkennungsfehler verursachte. Zwei Methoden, um die Zahl der Aussprachevarianten zu reduzieren, führten ebenfalls zu keiner Verbesserung. Die lexikale Adaption mit FST („finite state transducer“) Netzwerken konnte aufgrund neuer Erkennungsfehler auf keinem der beiden Experimente zufriedenstellende Ergebnisse erzielen, obwohl ein Maß für die Verwechslungswahrscheinlichkeit der Aussprachevarianten eingeführt wurde. Dieses konnte jedoch aufgrund der wenigen Daten im Reimtest die Verwechselbarkeit nicht akkurat abbilden.

Acknowledgments

During my studies and while writing this thesis many people have helped me, in both technical and personal ways.

First and foremost I would like to thank my parents and grandparents for supporting me in every thinkable way not only during my studies but through my entire life. I am grateful to my boyfriend Stefan Berger, my sister Verena Rexeis and my friends, especially Ulrike Aldrian for being always there for me and for correcting parts of my thesis. I warmly thank my colleagues at the DSP-Lab, especially Martin Schickbichler, Florian Krebs, Anna Fuchs and Wolfgang Jäger for their input and the great time at the laboratory. In this context I also want to thank the Signal Processing and Speech Communication Laboratory of the Graz University of Technology and professor Gernot Kubin for providing students such a great place to work.

I am heartily thankful to my supervisor Dr. Stefan Petrik for his support, constructive input and feedback, as well as for being the reference speaker in this work. In addition I want to thank the team of the Simon project for organizing and for carrying out the recordings of the dysarthric speakers, whom I also owe a very special thanks for their time and effort. Without them this thesis would not have been possible. I also would like to thank Dr. Harald Romsdorfer for his helpful input.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Place, Date Signature

Contents

List of Figures

List of Tables

1. Introduction

1.1. Definition of Dysarthria

Dysarthria is a collective term for speech impairments 'resulting from disturbances over the speech mechanism due to damage of the central or peripheral nervous system'[11]. These damages can be caused by a traumatic brain injury [20, 29] or stroke [11, 12, 29]. They can also be the result of neuromuscular diseases like cerebral palsy, Parkinson's disease [11, 12, 20] or Multiple Sclerosis [12]. In addition Dysarthria is mentioned in connection with Huntington's disease [11] and Down Syndrome [29].

Depending on the patient's particular damage in the nervous system speech organs like 'lungs, larynx, oro- and nasopharynx, soft palate and articulators (lips, lounge, teeth and jaw)'[12] can be affected to a different degree. Consequently the type and severity of the individual speech impairment varies strongly from patient to patient. Phenomena that are commonly observed in dysarthric speech are 'imprecise consonants and distorted vowels, irregular articulatory breakdowns, excessive or equal stress to all syllables and a slow rate of speech with a phonatory-prosodic insufficiency described as harsh, monotonous and monoloudness'[11].

1.2. Motivation

The use of automatic speech recognition (ASR) systems has become a part of everyday life in the recent years and systems with high recognition rates are widely available. With the increasing recognition rate for standard speech also the development of ASR-systems that achieve acceptable performance for speakers with atypical speech, e.g. speakers with foreign accent or speech impairments like dysarthria has made progress and has become one of the focuses in current research [28].

Dysarthric speakers often also suffer from motor disabilities of other body parts, e.g. the limbs that are resulting from the same disease or injury as the dysarthric symptoms. Limitations are ranging from an decreased level of dexterity to complete paralysis of the affected body part. For these patients the use of speech technology, e.g. in environmental control systems, can help to improve the quality of life in many different areas [8] allowing them to autonomously control things in their home-environment, which would otherwise require the help of another person.

Another widely studied application of ASR is to provide a user-interface to a computer for motor-impaired persons, who cannot easily use a keyboard or other supportive devices. This way ASR-systems can provide access not only to very basic activities such as writing or listening to music, but also to all amenities of the digital age, e.g. information access and modern communication forms. However, for patients with severe dysarthria these systems have so far only been successful when speech recognizers with a very small vocabulary were used [36].

Patients suffering from dysarthria can improve the control over their speech organs with speech therapy, although a lot of training is necessary. The training is usually monitored by a

speech therapist who gives feedback and corrects the patient. ASR-systems in computer aided systems for speech therapy, e.g. for pronunciation verification [29] can help the patient to train his voice autonomously without a speech therapist being present as extension to traditional speech therapy.

However, all systems described will only be accepted by dysarthric speakers if they are easy to use and work reliably, which of course strongly depends on the recognition rate of the speech recognizer. Otherwise using the system will be frustrating for an impaired speaker. As the nature of dysarthric speech has a strong impact on the recognition rate of modern speech recognizers, the development of technologies that improve recognition performance for dysarthric speakers is a challenging topic.

1.3. Problem setting

The ideal case to implement an ASR-system for a dysarthric speaker would be to train a speaker-dependent (SD) acoustic model from a huge set of data. But as speaking for long periods of time can be very tiring for patients suffering from dysarthria [30] this method can not be considered as optimal solution in most practical cases. Training speech recognizers with a small dataset has the drawback that parts of the acoustic model might not be well trained, leading to reduced recognition performance.

Another possible solution is the use of a well trained speaker-independent (SI) acoustic model that is adapted to a dysarthric speaker. Different state-of-the art approaches for adaptation on acoustical [14, 38] and lexical [3, 4, 35] level exist in the literature and have been studied for different problem settings. Many of them have also been applied to dysarthric speech for English, Spanish and other languages. A brief summary will be given in section 1.4. However especially for speakers with severe dysarthria there is still a lot of room for improvement.

This work investigates the recognition performance of ASR-systems using both SD and SI acoustic models focusing on the potential improvements achievable on SI-models with acoustic adaptation using maximum likelihood linear regression (MLLR) and lexical adaptation based on phonological rules and finite state transducer (FST)-networks on different small vocabulary tasks. Evaluations are done on speech data provided by the Simon project [31] of five male speakers suffering from moderate to severe dysarthria.

1.4. Related Work

Many speech recognition applications have been developed in the recent years for dysarthric speakers. This section gives a brief overview about the developments in the different fields described in section 1.2 and the technologies used. Some of the results achieved in these works are also discussed in more detail in chapter 3 and chapter 4, where also the underlying technologies are described in depth.

The EU-project ENABL [18, 25] aims to provide computer access to patients with motor-speech disorders using SI hidden Markov model (HMM) speech recognizers and MLLR adaptation. The STARDUST project [5] also had the goal to provide access to assistive technology for dysarthric speakers, focusing on patients suffering from severe dysarthria. The work of the STARDUST project was based on SD HMM recognizers and was continued in two follow-up projects [7, 20]. The open-source software developed by the Austrian Simon project [31] also aims to provide a user interface for interaction with a computer for dysarthric speakers based on SD HMM models. A small speech corpus was recorded by members of the project [31] of

four dysarthric speakers (two male, two female), containing samples of isolated digits, 21 command words as well as a few words and short sentences from speech therapy settings. Speech recognizers based on both SI and SD acoustic models were also evaluated on Dutch-speaking dysarthric speakers in [26].

An expansive set of game-like utilities for speech and language therapy was presented in [29] addressing several symptoms of dysarthria. While some tools are based on acoustic features of the speech signal itself, e.g. formants, pitch and sonority, also two utilities were presented for pronunciation training that are based on ASR. A speech corpus from 14 Spanish-speaking children (7 male/7 female) suffering from dysarthria was recorded and is described in depth in [22]. From each speaker four sessions of a 57 word vocabulary were recorded. The phonetically rich words were taken from a popular Spanish handbook for speech therapy. An SI acoustic HMM-model trained from a large Spanish speech database was adapted to dysarthric speech using maximum a posteriori (MAP). The same speech database and acoustic models have also been evaluated on a simple lexical adaptation task described in [27]. In this work a phone recognizer was used to generate transcriptions of the recorded speech samples using a leave-out strategy for evaluation of the found pronunciation variants. The phone transcriptions of three utterances of each word were added to the pronunciation dictionary as possible transcription and the new dictionary was evaluated on the fourth utterance. The work presented in [28] focuses on adaptation on feature level using a Vocal Tract Length Normalization algorithm to improve the performance on whole word HMM speech recognizers. The same corpus of 57 words as described previously was recorded for 19 Spanish-speaking dysarthric speakers for evaluation of this work, but the age distribution of the speakers was greater ranging form 15 to 60 years.

Another work [20] dealing with lexical adaptation for dysarthric speakers is based on FSTs to generate the new pronunciation variants. Evaluation was done using speech data from the Nemours database [19] which contains speech data of 11 male dysarthric speakers. For each speaker 74 nonsense sentences were recorded that had the form 'The X is Ying the Z.' with $X \neq Z$ being element of a set of 74 monosyllabic nouns and Y selected out of 37 monosyllabic verbs. The selection of the nouns and verbs was done 'to provide closed set phonetic constraints (e.g. place, manner and voicing constraints) within an associated set of four to six words'[19].

At the University of Illinois a speech database was developed for large vocabulary tasks on dysarthric speech. The Universal Access database (UA-database) [13] contains data from patients suffering from spastic dysarthria, sometimes mixed with other forms. Recordings are grouped in different word categories: digits, computer commands, radio alphabet letters, common words (taken from Brown corpus of written English) and uncommon words selected from Children's novels. By the time of writing parts of the database were publicly available for research on the homepage of the Statistical Speech Technology Group [34] at the University of Illinois. In [30] HMM speech recognizers based on different subword modeling types were evaluated on different subsets of the UA-database. By the time of writing no information was available about the existence of a publicly available German corpus of dysarthric speech.

Recently tools for assistive writing [9, 36] have been developed that rely on neural network (NN) and hybrid HMM/NN speech recognizers. Other works focus on speech enhancement devices [11, 12] that aim to improve intelligibility for a human listener by re-synthesis of dysarthric speech.

1.5. Thesis Organization

The thesis is organized as follows: In chapter 2 an acoustic analysis is done on the speech corpus recorded from five males suffering from dysarthria for this work to illustrate the differences between dysarthric and normal speech. Chapter 3 gives a brief introduction in state-of-the art speech recognition technology and describes the speech sources used in this work in detail as well as the training of the acoustic models. In addition different approaches for acoustic adaptation are presented and corresponding results achieved in related works on dysarthric speech are discussed. In Chapter 4 different lexical adaptation approaches are introduced and a detailed description of the implementations used in this work is given. The tasks on which the ASR-systems and proposed adaptation approaches were evaluated on dysarthric speech are presented in chapter 5, as well as the interpretations of the achieved results. Chapter 6 summarizes the findings of this work and gives a short outlook.

2. Acoustic analysis of dysarthric speech

Several studies have investigated the influence of different characteristics of dysarthria on ASR-systems. In [28] a smaller phonetic distance between the five Spanish vowels was measured in speech utterances of 30 dysarthric speakers compared to a reference corpus of 19 unimpaired speakers, which is a potential source of error for an ASR-system.

In [29] a correlation between the intelligibility of dysarthric speech for a human listener and the recognition rate of an ASR-system could be shown for patients suffering from a different degree of dysarthria.

On the other hand many works, e.g. [30] have also shown that the recognition rate of a speech recognizer depends strongly on the individual speaker when patients with a similar level of human intelligibility are compared. This indicates that the individual symptoms of the impaired speaker influence an ASR-system to a different degree.

An acoustic analysis of 12 Swedish-speaking patients suffering from dysarthria was done for the ENABL project [18]. In that work it could be shown that speaking rate and frequency of pauses of the individual speaker are of potential importance for a speech recognition system. In addition articulatory deviations were detected in the speech data, that were expected to cause severe problems for a speech recognizer. To overcome the latter 'specially adjusted phonetic transcriptions of the words' [18] were proposed.

As already mentioned in chapter 1 speech data from five male dysarthric speakers `M002`, `M063`, `M066`, `M067` and `M068` was recorded by members of the Simon project [31] for this work. In this chapter a brief analysis of the speech samples is done regarding their intelligibility, as well as distinctive features like disfluencies and mispronunciations. In addition the dysarthric speech samples are compared to speech data recorded from an unimpaired reference speaker `M000`.

2.1. Observations

A comparison of the speech samples from the individual speakers showed that a great variety in loudness between different utterances could be observed for the speakers `M002` and `M063`. Due to the decreased control of the speech organs some phones are uttered harshly, while other utterances are spoken in a mute tone. The average power E_s in dB

$$E_s = 10 \log_{10}(\frac{1}{T} \sum_{t=1}^{T} x[t]^2) \tag{2.1}$$

shows the variation of the loudness in the speech signals. For example `M002`'s utterance of 'gell' [gɛl] has an average power E_s of around -15 dB, while E_s of the utterance of 'Hain' [h͡aɪn] is around -31 dB. The average power of the same words uttered by the reference speaker is between -29 and -30 dB. For all calculations the speech signals were clipped with a rectangular window to exclude the parts of the signals containing only silence. The speech samples recorded from `M066` and `M068` generally seem to have a more constant loudness than the utterances of `M002` and `M063`, but some utterances also sound harsh. `M067` also speaks with a constant

loudness, e.g. for the words 'gell' and 'Hain' uttered by M067 E_s is around -19 dB and -22 dB respectively. But in contrast to the other speakers M067's utterances do not sound harsh and it seems that he speaks with a muter voice.

The speech of all dysarthric speakers analyzed in this work is slurred to a certain degree. An example for the word 'Wald' [valt] uttered by two dysarthric speakers M002 and M067, as well as the unimpaired reference speaker M000 illustrates some of the differences in speech. A comparison between the utterances of M002 in figure 2.1 and M000 in figure 2.2 shows that the utterance by the dysarthric speaker is almost twice as long. M002 stretches the vowel /a/ and the following lateral approximant /l/, as well as the transition between both phones, which itself sounds like the diphthong /a͡i/. In contrast the utterance of M067 shown in figure 2.3 has about the same length as the utterance of M000, but the labiodental fricative /v/ is pronounced like the approximant /w/ and is stretched. The duration of the approximant in M067's utterance is longer than the durations of both /a/ and /l/.

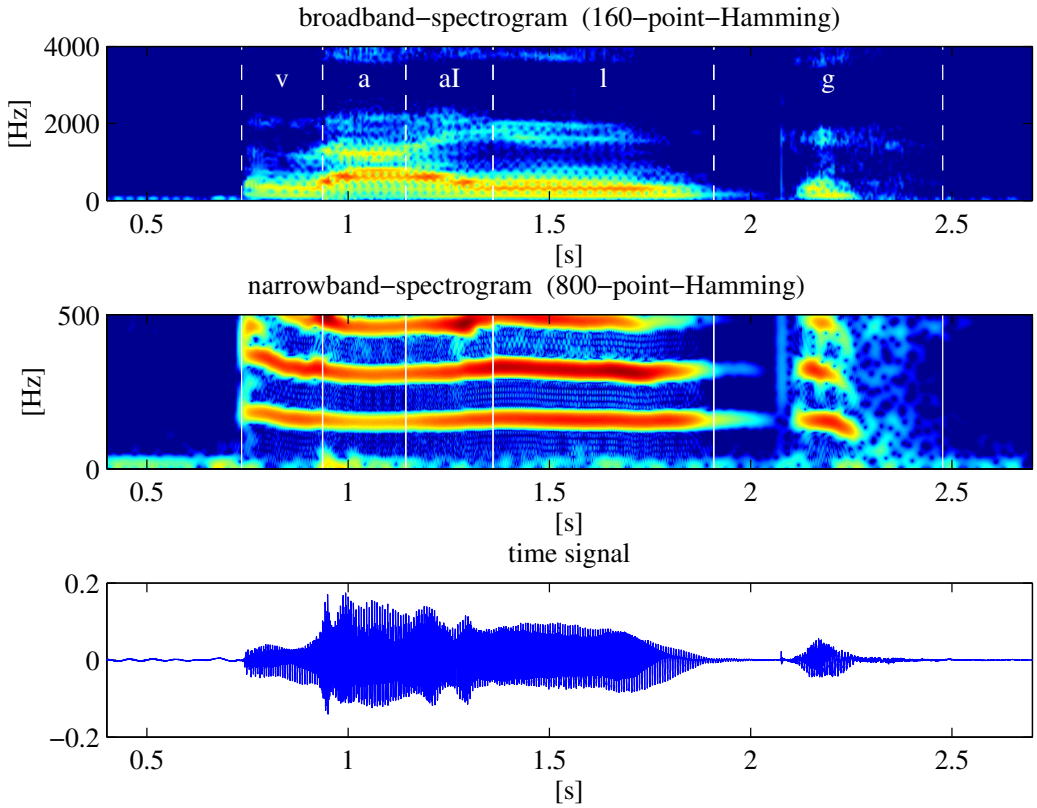

Figure 2.1.: The spectrogram of the word 'Wald' [valt] uttered by M002. The SAMPA-labels of the phones in the spectrogram were assigned manually.

From the narrowband-spectrogram of the examples shown in figure 2.1 and figure 2.3 one can also see that the number of harmonics is lower than in the same utterance of M000 shown in figure 2.2, which shows that the prosody of the dysarthric speakers is more monotonous than for M000. The other dysarthric speakers show the same features in the narrowband-spectrogram.

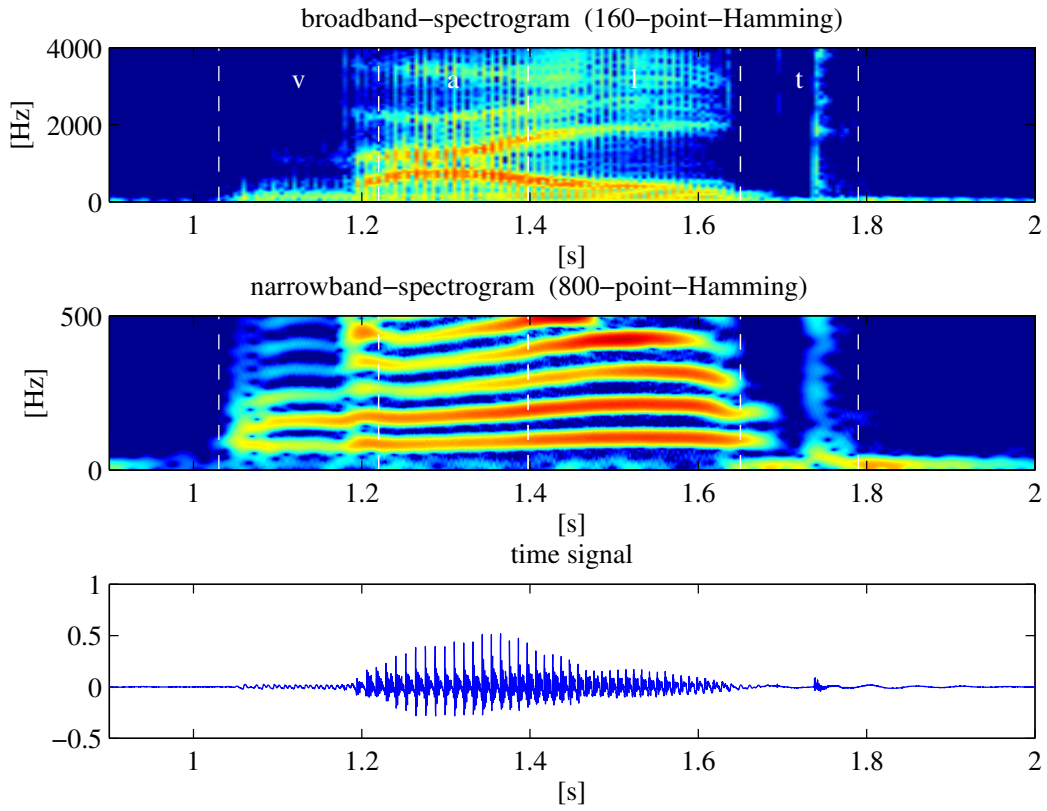

Figure 2.2.: The spectrogram of the word 'Wald' [valt] uttered by M000. The SAMPA-labels of the phones in the spectrogram were assigned manually.

2.2. Disfluencies

In case of M002 non-linguistically motivated pauses occur in phone transitions that are difficult for the speaker to pronounce. The occlusion phase in plosives is also often longer than in case of the reference speaker, which can also be observed in figure 2.1. It could be shown that both the insertion of pauses and the duration of the occlusion phase cause errors in speech recognition systems for this speaker (see section 5.4.1).

Looking at the broadband-spectrogram of figure 2.1 one can also see that the transitions between phones are not as fluent as for M000 as the lack of control over the speech organs leads to jumps in the spectrogram. During the evaluation it could be shown that these disfluencies also lead to recognition errors in ASR. An analysis of the resulting errors from such disfluencies is done in section 5.2 on an example utterance of M066 (figure 5.1 shows the corresponding spectrogram).

2.3. Phone confusions

Audible phone confusions are present in the speech samples of all analyzed dysarthric speakers, although the number of confusions varies strongly from speaker to speaker.

A common source of error are plosives which are often pronounced incorrectly by M002, M063 and M066. While M002 tends to utter /t/ and /d/ fronted as /k/ and /g/, M066 sometimes pronounces unvoiced plosives as voiced. Both M002 and M066 sometimes insert the nasal /m/

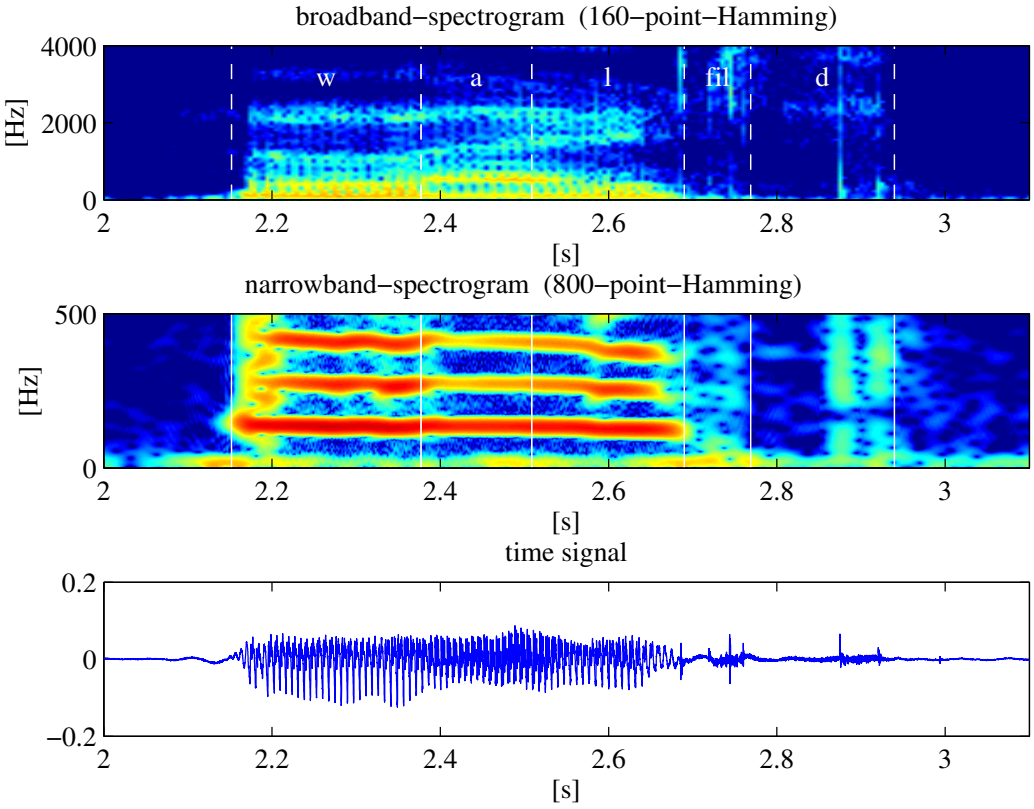

Figure 2.3.: The spectrogram of the word 'Wald' [valt] uttered by M067. The SAMPA-labels of the phones in the spectrogram were assigned manually.

before plosives at the beginning of a word. For M002 this insertion could only be observed in front of utterances of /p/, while for M063 the insertion occurred in front of utterances of /p/ and /t/.

In contrast both M067 and M068 are able to pronounce plosives correctly. However, in the utterance shown in figure 2.3 M067 inserts a filled pause with a audible breathing or smacking like sound before the occlusion phase of the plosive. This filled pause could also be observed in the same context in other utterances of the speaker. Both speakers have problems with the pronunciation of other phones. For example in case of M068 utterances of the phones /r/ and /l/ are pronounced as /j/. As already mentioned M067 replaces the fricative /v/ with the approximant /w/.

M063 had the lowest intelligibility in the analysis as he often swallows consonants at the beginning and end of a word that are difficult for him to pronounce. Many phones are generally not uttered by the speaker at the beginning of words, e.g. /f/, /r/, /s/, /t͡s/ and /ʃ/. Also combinations of plosives and trills, e.g. /d r/ are swallowed at the beginning of words. M063 also sometimes truncates the end of a word, e.g. if it contains a plosive. In some cases utterances are truncated in both the beginning and end of the word, making it almost impossible to recognize the original word, even for a human listener. For example the utterance of 'flink' shown in figure 2.4 with the orthographic transcription [flɪŋk] is truncated by the speaker in a way that the recorded sample actually sounds like [ɪmf] with an additional distortion at the end of the utterance, where the sound sample is overdriven. In comparison to the consonants the vowels are uttered more clearly by M063, although phone confusions occur, e.g. the vowels

/a/ and /o/ are pronounced as the diphthong /a͡ʊ/ in some cases.

The truncations of words by M063 cause severe problems for both for human intelligibility and speech recognition. During the evaluation recognition results for M063 were always the lowest. While tasks with a very small vocabulary containing only digits acceptable results were achieved, recognition tasks with a 69 word command words vocabulary failed for this speaker in contrast to the other dysarthric speakers. A further analysis of the impact of M063's speech on ASR-systems can be found in section 5.3.2.

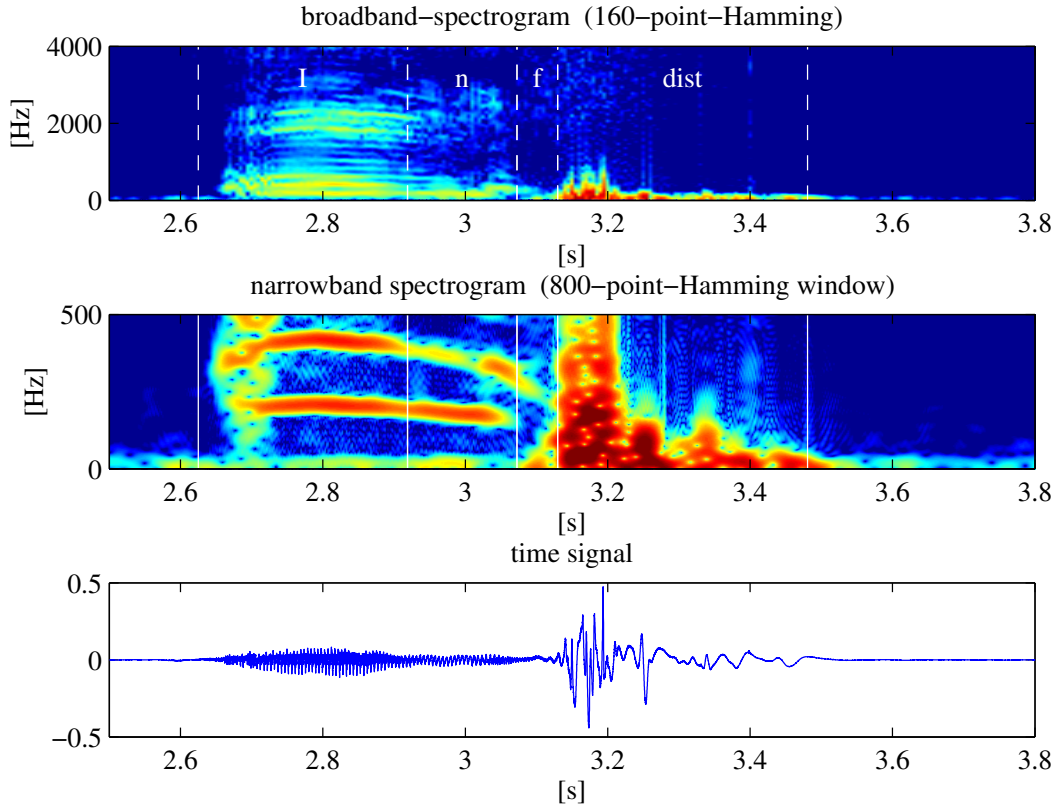

Figure 2.4.: The spectrogram of the word 'flink' [flɪŋk] uttered by M063. The SAMPA-labels of the phones in the spectrogram were assigned manually.

3. Speech recognition system

3.1. Components of a speech recognizer

A speech recognizer translates an acoustic input into the most likely sentence S of a predefined language. A sentence is composed of one or more words W from the pronunciation dictionary, each word consisting of a sequence of labels l from the label set L defined in the acoustic model. Each label typically represents a linguistic unit, e.g. a word or phoneme.

A simplified model of a speech recognizer is shown in figure 3.1. In the feature extraction step the acoustic signal is sampled into overlapping frames from which multidimensional observation vectors $\boldsymbol{o}_t$ are computed. The acoustic model calculates the observation likelihood for every label l of the label set for each given observation vector $\boldsymbol{o}_t$. The pronunciation dictionary holds the list of known words of the language and their pronunciations specified in terms of the label set L represented in the acoustic model [3]. The language model uses prior knowledge about all words in the language and their allowed combinations to assign each word the probability of appearing in the language in current context. The decoder combines these informations to find the most likely label sequence S for the given speech signal.

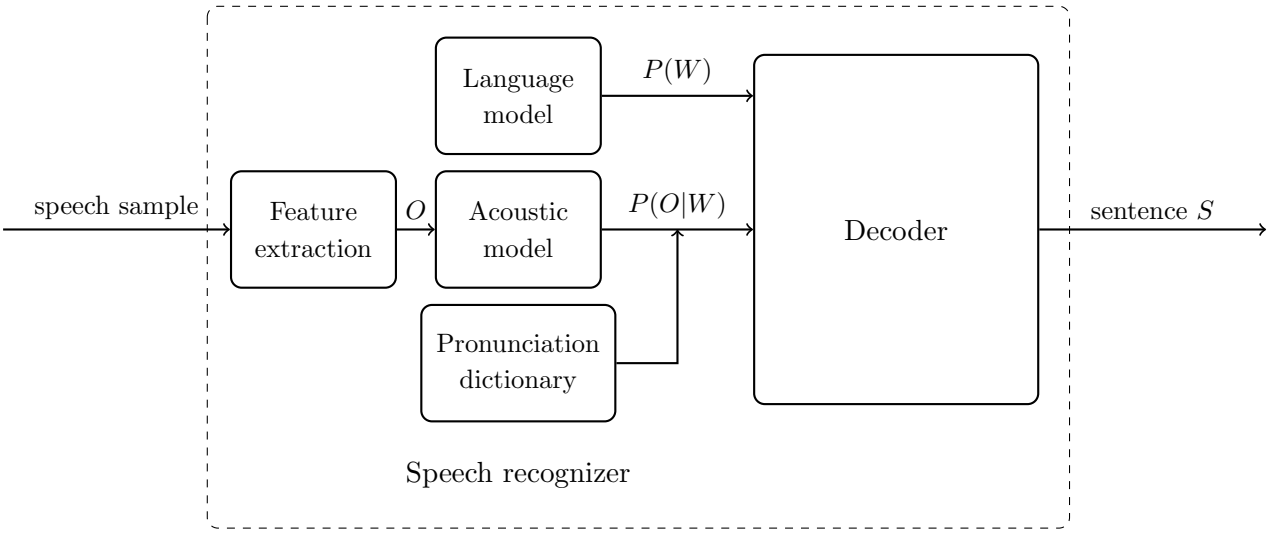

Figure 3.1.: Schematic architecture of a simplified speech recognition system.

3.1.1. Feature extraction

In the feature extraction step a parametric representation of the speech signal is computed. The Mel-frequency cepstral coefficients (MFCC) are widely used as acoustic features in speech recognition. In this work the 12 MFCCs and energy coefficient are used including the corresponding

velocity and acceleration coefficients. The 39-dimensional MFCC-vectors were extracted from the speech samples using the Hidden Markov Model Toolkit (HTK)-tool `HCopy`.

The computation of the MFCC-vectors is done in six steps [10]. In the first preemphasis step a high-pass filter is used to boost the energy of the higher frequencies of the sound signal. The pre-emphasis coefficient $0.9 \leq \alpha \leq 1.0$ of the filter equation $y[n] = x[n] - \alpha x[n-1]$ was set to $\alpha = 0.97$ in this work. In the second step the speech signal is windowed into small frames that can roughly be assumed to be stationary. In this work a Hamming window with a length of 25 ms was applied to the speech signal with a frame shift of 10 ms. The energy coefficient is computed directly from the time signal of each frame, as well as the fast Fourier transform (FFT). The spectrum is then filtered with a Mel-filter bank, which scales the frequency bands according to the sensitivity of the human ear. The cepstrum is then computed by taking the inverse fast Fourier transform (IFFT) of the log of the spectrum. The first 12 values of the cepstrum are used as MFCCs. The 13 delta features of the cepstral coefficients and the energy represent the changes of the corresponding features between frames, while the 13 double delta coefficients represent changes of the delta coefficients between frames.

3.1.2. Acoustic model

The acoustic model is a set of HMMs each modeling one label l from the label set L, which can either represent a whole word or a phoneme-like subunit, depending on the system design. A single HMM is a sequence classifier [10] that assigns a sequence of observations $O = \boldsymbol{o}_1 \dots \boldsymbol{o}_T$ to the most likely sequence of hidden states $q_1 \dots q_K$. The emission probability or observation likelihood $B = b_k(\boldsymbol{o}_t)$ expresses the probability that a given observation vector $\boldsymbol{o}_t$ was generated by the hidden state q_k of an HMM. The transition matrix A models the transitions between the hidden states. In speech recognition usually left-to-right transition matrices are used, which means that no transitions to previous states are allowed. The most common method to model the emission probability $b_k(\boldsymbol{o}_t)$ of a state is to compute a probability density function (PDF) [10] over the feature space using a Gaussian mixture model (GMM).

The number of states in an HMM as well as the linguistic unit representing it in the acoustic model is depending on the task and system design. For small vocabulary tasks, e.g. digit recognition, a unit can be modeled to represent a whole word. For large vocabulary tasks it is usually more useful to model phone-like subunits that are concatenated to words using a pronunciation dictionary. This way shared sounds can be modeled across words and redundancy as well as model size is reduced, as the words of a language usually are formed from a relatively small, closed set of sounds. In addition the model is also capable of recognizing unseen words as long as they are composed of the same labels. For example the lexicon of the Speechdat(II)-AT (Speechdat) database contains more than 15.000 words, all of them consisting only of 46 phonemes. For acoustic models with phoneme-like subunits HMMs with 3-states are commonly used, while for whole word acoustic models the number of states per HMM is usually higher.

As the neighboring phones may alter the pronunciation of a certain phoneme, a common extension to modeling phonemes as HMMs is to use triphone HMM-models. For each phone and its possible left and right context a separate HMM-model is created. As this extension leads to an extensive increase of the number of training parameters the states of the different HMM-models are clustered using a decision tree based on prior knowledge about acoustic similarities of phones in different contexts of the given language. The states in each cluster are then 'tied' together, which means that they share the same GMM to model the emission probability.

3.1.3. Decoder

The decoder maximizes the joint probability of the language, pronunciation and acoustical model to find the most likely sequence of words

$$S = \arg\max_{W} P(O|W)P(W) \tag{3.1}$$

The Viterbi-algorithm [10] is a dynamic programming approach to recursively find S from all possible label combinations. The probability $v_t(k)$ of observation $\boldsymbol{o}_t$ being observed in state q_k is computed as

$$v_t(k) = \max_{j=1}^{K} v_{t-1}(j) a_{jk} b_k(\boldsymbol{o}_t) \tag{3.2}$$

by selecting the most likely transition from all K previous states by weighting $v_{t-1}(j)$ with the transition matrix entry a_{jk} and the emission probability $b_k(\boldsymbol{o}_t)$. The most likely sequence S is found via backtracking of the most likely path through the probability network.

In the HTK-toolkit which is used for recognition in this work the Viterbi-algorithm is implemented in the tool `HVite`, that can also be used for performing forced alignment of given reference transcription to a speech sample. `HResults` is a HTK-tool to evaluate an output of `HVite` against a reference transcription for error analysis.

3.2. Acoustic model training

The training of the emission probabilities and the transition matrices is done using the Baum-Welch or forward-backward algorithm [10], which is a dynamic programming approach that updates the model parameters iteratively using a set of labeled training data. The HTK-toolkit [39] provides two tools `HRest` and `HERest` that implement the Baum-Welch algorithm which was used for training of the acoustic models in this work.

For training of the emission probability each state is first modeled by a single Gaussian component, which is initialized with the mean and standard deviation of the acoustic feature set of the training data. The HTK-toolkit provides the tool `HCompV` for this initialization step. After several training passes, in which the emission probabilities are updated with the Baum-Welch algorithm, the single Gaussians are split into two identical components and retrained. The splitting and retraining is repeated until the desired number of mixture components is reached. The update of the HTK-model definitions to alter the number of Gaussian mixtures per state was done with `HHEd`, which can also be used to update the parameters for state-tying when triphone models are derived from monophone models.

In this work both monophone and triphone HMM acoustic models were trained and evaluated on two small vocabulary recognition tasks: a command word task and a connected digits task. In addition whole word models were trained for the connected digits task. The recognition performance of the acoustic models was evaluated on data from the dysarthric speakers and an unimpaired reference speaker.

3.2.1. Speech sources

The HMM-based speech recognizers were trained from different speech sources using speech data from both impaired and unimpaired speakers. For the SI-models a large database containing speech data from 1000 unimpaired speakers was used. SD-models were trained as well from

category	category type	samples/speaker
B	isolated digit	2
	10-digit sequence	1
C	connected digits (4-16 digits/sample)	4
N	natural number	1
M	money amount	2
Q	yes/no	2
D	dates	3
T	times	2
A	application words	6
E	word spotting phrase with one embedded application word	1
O	directory assistance names	7
L	spellings (words and names)	3
W	phonetically rich isolated words	4
S, Z	phonetically rich sentences	12
Y	speaker specific material	7

Table 3.1.: Speech data recorded for Speechdat(II)-AT ordered by categories[1].

data recorded from five male speakers suffering from dysarthria and an unimpaired reference speaker.

Speechdat(II)-AT-database

The Speechdat-database contains A-law coded speech data from 1000 speakers and is balanced in gender (544 male/456 female speakers). The recordings have a bandwidth of 8 kHz and were collected over fixed telephone network. A detailed documentation of the database can be found in [1].

Each participating speaker received a data-sheet with different reading tasks and questions. A free telephone number stated on the sheet could be called by the speaker to record the data. Speech samples of different categories where collected along with speaker specific demographic data, such as age and native language. Details about the categories and number of recordings per speaker can be found in table 3.1. Almost all speakers can be considered to have Austrian German as native language. The distribution of speakers among age groups is shown in Table 3.2. The documentation contains no information if any of the speakers suffer from a speech impairment.

A lexicon is also included in the database which contains all words that occur in the speech samples including one phonetic transcription per word. The lexicon is based on an 'existing hand-corrected lexicon by Philips Speech Processing'[1]. The phones were coded in SAMPA-symbols. The complete phone set and the corresponding IPA-symbol can be found in table B.1 and table B.2 in appendix B. The transcription of the recorded data was done orthographically. In addition special labels were used to mark truncated utterances, mispronunciations, background noise, hesitations [fil] and nonverbal speaker sounds [spk].

Dysarthric speaker data

The speech samples from five dysarthric speakers were collected in a living room environment using a laptop and a headset. The selection of the speakers as well as the recordings were done

Age Group	Count	Percent
0-15	15	1.5
16-30	444	44.4
31-45	328	32.8
46-60	184	18.4
61-99	29	2.9

Table 3.2.: Age distribution among speakers from the Speechdat(II)-AT[1].

by members of the Simon project [31]. The age of the speakers at the time of the recordings varied between 17 and 38 years. This age group is also well represented in the Speechdat-database (see table 3.2).

All recordings were stored as 16bit WAV-files with a sampling rate of 16 kHz. During the recording process each sample was first read to the impaired speaker who then repeated it. Unfortunately this method lead to minor distortions in parts of the speech data due to audible whispering. The recordings of the reference speaker were made under similar conditions.

Sessions of three different data sets were collected from all speakers:

- 69 command words (5-16 sessions/speaker)
- 100 connected digits (1 session/speaker)
- 100 rhyme-test words (1 session/speaker)

The command words contain 69 German words suitable for doing simple tasks on a computer such as listening to music and doing basic calculations. The first idea was to use the list of application words from the Speechdat-database (see table 3.1), but this words seemed to be mainly selected for control of an answering machine and therefore were not suitable for this task. Consequently the words were selected manually and include e.g. 'abspielen' (play), 'hinauf' (up) and 'Hilfe' (help). The complete list of words can be found in Appendix B.2.2. As most of the sessions from this set were used for training of SD HMM-models, the words were also selected with the intention to cover all German phonemes. However, it turned out that the Vowel /ɛː/ (as it appears e.g. in the German word 'spät' [ʃpɛːt] (late)) occurs in none of the words.

The total number of command words sessions recorded for each speaker is shown in table 3.3. From the collected command words sessions of each speaker one was defined as development set for acoustic adaptation of the SI-models and one as test set. The remaining data was used to train SD-models for speaker `M000`, `M002` and `M063`. For `M066`, `M067` and `M068` only 5 sessions were available in total and three sessions were considered as not enough to train an SD-model. For `M063` six complete sessions were available. In 10 other sessions two words were missing while other words were recorded more often (up to 19 samples). In total 1091 command words samples were available from `M063` (between 6 and 25 per word). All samples of the incomplete sessions were added to the training set while the development set and test set were each selected from the six complete sessions.

The connected digits set contains 100 randomly generated four-digit-sequences. The complete list that was recorded from all speakers can be found in appendix B.2.1. Of these sequences 80 were defined as training set for SD HMM-models while the remaining 20 were split into a development and test set containing 10 digit sequences each.

For the rhyme-test words one word from each of the 100 ensembles of the Sotschek rhyme-test which will be described detail in Section 4.1 was selected and recoded by all speakers. This data was used for evaluation of both the SI-models and SD command word models and

Speaker	# train-sessions	# dev-sessions	# test-sessions	# sessions	# samples
M000	4	1	1	6	414
M002	8	1	1	10	690
M063	4[a]	1	1	6	414 (1091)
M066	3	1	1	5	345
M067	3	1	1	5	345
M068	3	1	1	5	345

[a]plus additional data

Table 3.3.: Recorded command word sessions for individual speakers.

as development set for the lexical adaptation tasks.

Two lexica were developed for the command words and the rhyme-test respectively. As isolated digits are part of the command words no separate lexicon was needed for the connected digits. The phonetic transcriptions of the words in both lexica were mainly taken from the Speechdat-database lexicon and use the same set of SAMPA-symbols. For words that were not included in this lexicon, as it was the case for many of the rhyme-test one-syllable words, the transcription was derived from related words or parts of longer words from the lexicon. If no matching word could be found a proper Austrian-German transcription was selected from the larger ADABA-lexicon [21].

3.2.2. Monophone and triphone model training

Monophone and triphone SI and SD acoustic models were trained from the previously described speech sources based on the Speechdat reference recognizer (Refrec0.96) [16].

SI-models

The scripts from the Refrec0.96 framework were used to train SI-models with two different training sets: one model was trained on the full Speechdat-database and one using only the connected digits subsets B and C (see table 3.1).

The training procedure implemented in the framework follows the steps of the tutorial of the HTK-toolkit [39], but special preprocessing and bootstrap training was added as described in [16]. First the Speechdat-database data is loaded and split into predefined training, development and test-sessions. Speech samples with a transcription that includes a noise marker different from [fil] and [spk] are automatically excluded. Before the extraction of the MFCC components the speech samples are converted from 8bit A-law coding to 16bit linear coding with a short C-Program which is also part of the Refrec0.96 framework. Afterwards the Speechdat-database lexicon is loaded, the transcriptions are converted from SAMPA-symbols to HTK-SAMPA and a phone list is generated. The option of the framework to specify a phonetic mapping to avoid the modeling of rare phonemes was used to map two rare phonemes present in the Speechdat-database lexicon to other phonemes: /d͡ʒ/ to /d ʃ/ and /ʒ/ to /ʃ/.

The main difference of the training procedure of the Refrec0.96 compared to the HTK-tutorial is an initial bootstrap training step with reduced training data (only the phonetically rich sentences described in section 3.2.1, detailed information about the phone statistics can be found in [1]). The preprocessing aims to remove samples with wrong or empty transcriptions.

Monophone models are initialized with the global mean and average of the training data and a 3-state HMM silence model as well as a 1-state HMM for modeling of short pauses are created. The state of the short pause model is then tied to the emitting state of the silence model as described in the HTK-tutorial. After 2 re-estimation steps a realignment of the training data is performed using the Viterbi Algorithm and speech samples that could not be aligned are excluded, while for the remaining data another 2 re-estimation steps are performed. Then the number of Gaussian mixtures that model the emission probabilities of the HMM-states is stepwise increased to 2, 4, 8, 16 and 32 GMMs per state followed by two re-estimation steps after each change. The 32 GMM monophone model is used for realignment of the whole training data to remove outlier utterances. A new prototype HMM is initialized with the realigned data and training of the monophone and triphone models is started from scratch following the steps described in the HTK-tutorial. For the triphone models the clustering of the states is done using decision tree with a question set included in the Refrec0.96 framework that was designed to support multiple European languages [16]. For both monophone and triphone models the number of GMMs per state is stepwise increased to 32 GMMs per HMM-state.

SD-models

The SD monophone and triphone models were trained using the recorded command words and the connected digits from the dysarthric speakers and the reference speaker as training data.

The scripts of the Refrec0.96 [16] were adapted for training of the models. A new preprocessing step was implemented to load both the lexicon and speech data from the dysarthric speaker-database. The HTK configuration parameters had to be changed as well to support WAV files and the conversion of the speech data before the feature extraction step was left out, as the speech data was already coded linearly. The processing of the lexicon needed no changes, as all lexica use the same format. The phonetic mapping between the vowels /ɛː/ and /ɛ/ was added to the predefined list of mappings to avoid errors in the recognition tasks, as the vowel /ɛː/ is not included in the command words training set (see Section 3.2.1), but occurs in the rhyme-test test set.

Only small changes were applied to the training algorithm. The bootstrap training was used as well for the dysarthric speakers to exclude distorted samples from training. But due to the small amount of training data available for the SD-models the initial training step was also done with the whole data available. However, only in case of `M063` samples were excluded from training. The realignment step was done with the model using 2 GMMs to model the emission probability per state. Monophone and triphone HMM-models with up to 16 GMMs per HMM were trained and evaluated.

3.2.3. Whole word model training

For the connected digits task SI and SD whole word models were trained as well. The reference recognizer [23] used for the work-items 007 and 008 of the AURORA project from the ETSI (European Telecomunication Standard Institute) DSR (distributed speech recognition) working group was therefore adapted to work with the same speech sources as for the monophone and triphone models. The feature extraction step was also changed, as the reference recognizer uses a separate front-end for feature extraction. For this work the MFCC-vectors were extracted in the same way as for the SD monophone and triphone models.

The training procedure was left unchanged. For each word a left-to-right HMM-model with 16 states was used. The high number of states was used in the original setting because of

the noisy environment. Related works have shown that a high number of states also increases recognition performance on dysarthric speech, e.g. in [28] whole word models with 24 states were used on a 57-word task containing phonetically rich Spanish words.

The HMM-states are initialized using the global mean and average of the train data. In addition a 3-state silence model and a 1-state short pause model were defined and added to the HMM set. The center state of the silence model is tied to the state of the short pause model. After 3 re-estimation steps the number of GMMs per state for the word and silence model is stepwise increased using seven re-estimation steps after each change. First only the silence model is increased to 2 GMMs per state. In the second step the word models are increased to 2 GMMs per state while the silence model is also increased to 3 GMMs. In the third and last step the word models are increased to 6 GMMs per state and the silence model is increased to 3 GMMs.

3.3. Acoustic adaptation

The general goal of acoustic adaptation is to optimize a trained speech recognition system like shown figure 3.1 for a new speaker using the same language and pronunciation models. Speaker adaptation can be done in supervised fashion with a data set with known transcriptions or in an unsupervised fashion were the transcription of the adaptation data is estimated.

Acoustic adaptation methods can be divided into two groups [15]: speaker normalization where the observation vectors of the new speaker are normalized for a closer match to the acoustic model and model adaptation techniques in which the parameters of the acoustic model are optimized to match the observation vectors of a new speaker. Two model adaptation approaches which are commonly used and covered in more detail in the following sections are MLLR and MAP adaptation.

A popular speaker normalization technique is vocal tract length normalization [38] where the frequency axis is rescaled to compensate differences between the speakers using frequency warping methods to estimate the scaling factor. For an unimpaired speaker the improvement of the word error rate (WER) using this is method can be expected to be around 10%. This method has also been applied to dysarthric speech in [28] using HMM whole word models with 24 states per word and 1 Gaussian component per state to model the emission probability. The 57 word Spanish corpus used for this work was described previously in section 1.4. For an SI-model trained from unimpaired speakers the average reduction of the WER on dysarthric speech reached up to 11% using vocal tract length normalization, while for an SI-model trained from dysarthric speakers the average reduction was even higher reaching up to 17%. In this work adaptation using vocal tract length normalization is not covered in further detail, as no experiments were made in that direction. Instead this work focuses on adaptation of the acoustic model parameters and adaptation on lexical level.

3.3.1. MAP

MAP adaptation updates the mean vectors of the emission probability distributions of the HMM-states as the 'weighted sum of the SD observation vectors $\boldsymbol{o}_t$, regularized by the addition of a regularization constant λ'[6] times the prior mean vector $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}} = \frac{\lambda\boldsymbol{\mu} + \sum_t \gamma(t)\boldsymbol{o}_t}{\lambda + \sum_t \gamma(t)} \tag{3.3}$$

The value for the bias estimate λ between the mean vectors of the data and the prior is typically between two and twenty [38]. '$\gamma(t)$ is the probability of the Gaussian at time t'[38]. It has been shown that the MAP estimate converges to the maximum likelihood (ML) estimate when the amount of adaptation data increases towards infinity. However, standard MAP adaptation is a local approach and only updates the HMM-parameters that are observed in the adaptation data. In this work the data sets evaluated on the acoustic models do not always share the same set of triphones as present in the adaptation data, so MAP was not used.

MAP adaptation has been successfully applied to adapt acoustic models to dysarthric speech, e.g. in [29] where the average WER over 14 Spanish-children suffering from dysarthria was evaluated on a 57 word vocabulary (see section 1.4). Acoustic adaptation was applied in two steps on the 1-state SI-model with 16 GMMs. First, a task dependent adaptation was performed using speech samples of the 57-word vocabulary uttered by unimpaired children, then a second adaptation step was performed using three of the four recorded sessions of the dysarthric speakers. Evaluations were done using a leave-out crossvalidation. The average recognition rate over all speakers increased from 66.8% to 85.9% with 9 of 14 speakers achieving a recognition above 90%. In a related work on lexical adaptation an acoustic model with 3-states and 16 GMMs to model the acoustic likelihood was adapted with MAP which lead to a decrease of the average WER over all 14 speakers from 31.96% to 16.2%.

3.3.2. MLLR

MLLR-adaptation is done by reestimating the existing mean vectors $\boldsymbol{\mu}$ of the GMMs that model the emission probabilities of the HMM-states using a linear transformation matrix W that is estimated to maximize the adaptation data:

$$\hat{\boldsymbol{\mu}} = W\boldsymbol{\mu} \tag{3.4}$$

The transformation can either be applied globally or cluster-specific [6] where separate transformation matrices are applied to smaller subsets of HMM-states that are part of the acoustic model. In this case the states can be clustered using regression class trees. This method was also used in this work using the tools provided by HTK.

The MLLR-adaptation of the means of the Gaussian mixture components was done in supervised fashion using subsets of the speech data from the dysarthric speakers and the reference speaker for adaptation of the SI-models. Therefore it was necessary to downsample the recordings from 16 kHz to 8 kHz and apply a POTS-band filter. This was done with a Matlab^{TM} script using functions from the `VOICEBOX`-toolbox Version 1.6 [2].

The same preprocessing steps as for training of the SD-models were used for loading the speech data and for feature extraction. A forced Viterbi-alignment was performed using `HVite` to align transcriptions to the data provided for adaptation. The re-estimation of the means was done with `HERest`.

The whole word models were adapted to the new data set using one global transformation. For the monophone and triphone models the adaptation process is implemented as two-pass adaptation that follows the steps from the HTK-tutorial example [39] for MLLR-adaptation. The Gaussian mixture components were clustered with a binary regression class tree using the Euclidean distance between the means as measure for acoustic similarity. The number of terminal leaf nodes of the regression tree was varied depending on the model type and task between 10 and 40 nodes. The first split of each of the regression trees is used to separate

the speech and non-speech sounds. A global transformation is estimated and used as input transform providing better estimates for the second-pass transformation using the regression class tree. The transformation based on the classes of the regression tree is done automatically by `HERest` which uses a top down approach, generating transforms for all nodes with sufficient data that are either leaf nodes or have children without sufficient data [39].

The Gaussian variances of the emission probabilities can also be transformed in a similar way as the means. However, as this approach was not used in this work it is therefore not covered in detail.

4. Lexical adaptation

For an unimpaired speaker the pronunciation of a word is influenced in many ways depending on e.g. on the environment and the accent of the speaker. Also, e.g. read speech is pronounced differently than continuous speech. Pronunciation variation also influences the recognition performance of a speech recognizer [10]. The development of pronunciation dictionaries that allow multiple pronunciations for words has improved recognition rates for several ASR applications.

This concept is also promising when applied to dysarthric speech, because as shown in chapter 2 many speakers produce characteristic phone errors depending on the speech organs affected by their impairment. If this information is extracted properly and corresponding alternatives are added to the lexicon the recognition rate should increase. However, pronunciation variation might not overcome all aspects of dysarthric speech such as phonatory-prosodic insufficiency.

In [35] an overview of several lexical adaptation approaches and examples of their implementations is given. The information about the mispronunciation can be derived from existing knowledge or extracted directly from speech data, e.g. from incorrect recognizer outputs or a confusion matrix [4, 20]. From this information new pronunciation variants can be generated and added to the pronunciation dictionary. Possible approaches for pronunciation alternative generation [35] are phonological rules, finite state transducers, neural networks, decision trees and confusion matrices.

4.1. Generation of pronunciation alternatives

There are two major approaches how the information about pronunciation variation is created: Knowledge-based approaches rely on existing sources such as pronunciation dictionaries and general linguistic rules for pronunciation variation. In case of a dysarthric speaker a possible source could e.g. be information provided by a speech therapist. Data-driven methods extract the information about pronunciation alternatives directly for a given speech recognition system and/or speech data. This is usually done by transcribing the utterances either manually or with the speech recognizer.

A common data-driven method described in [3] is to extract pronunciation variants from a given data set by phone recognition using phone bi-gram or tri-gram grammars as language model. The result of the phone recognizer can then either be added directly to the lexicon or it can be aligned to the baseform transcriptions from a lexicon for further formalization.

A phone recognizer was also used in [27] to find new pronunciations for words uttered by dysarthric speakers. The output of the phone recognizer was added directly to the lexicon as new pronunciation variant for the given word. With this method an average relative improvement of 17% on the WER on an SI-model trained from unimpaired speakers could be achieved. However, in the same work a significantly higher average improvement of 49% could be achieved using acoustic adaptation. A combination of both adaptation approaches lead to a decline of the relative improvement compared to the acoustic adaptation.

Using a phone recognizer has the drawback that the recognition system has no lexical information and is sensitive to distortions. Long pauses, hesitations, breathing sounds and stretched

word	SAMPA-transcription
Sicht	s I C t
dicht	d I C t
Gicht	g I C t
nicht	n I C t
richt	r I C t
Licht	l I C t

Table 4.1.: Example for a set of words from the Sotschek rhyme-test.

vowels that occur frequently in dysarthric speech can result in many insertions and therefore a kind of noise is introduced in the pronunciation variant. Also an alignment of the recognized phone sequence to the original pronunciation can be difficult when a lot of insertions occur. In this work a different data-driven method is proposed to extract the pronunciation information in a closed setting based on the recognizer results of the recorded speech-data of words taken from the Sotschek rhyme-test.

4.1.1. Sotschek rhyme-test

The Sotschek rhyme-test [33] is a German rhyme-test which was originally developed to evaluate the intelligibility of speech transmitted over radio channel. The test contains 99 sets of German one-syllable words. The six words of each ensemble are phonetically different in exactly one part of the syllable, 33 differ in the nucleus, 33 in the onset and 33 in the coda. For example the set in table 4.1 contains six words with a different onset. The selected words are also phonetically balanced. One more set was added manually to the rhyme-test containing the affricate /p͡f/ in the coda of the syllable. Most words contained in the test also exist in the German Duden, although some, e.g. 'Haff' are not very common. The complete word list can be found in table B.3 in Appendix B. The manually added ensemble is number 100.

With this test it is possible to extract potential substitutions without taking side noises too much into account with the drawback that the detection of insertions and deletions is very limited.

For the test set one word was selected of each ensemble, bound to three conditions:

- each word must not appear more than one time
- each onset/nucleus/coda should appear at least once
- the word should be common and easy to pronounce (which was not possible for all ensembles)

Two types of lexical adaptation algorithms were evaluated based on the information about phoneme confusions (substitutions, insertions and deletions) found for the dysarthric speakers on the rhyme-test evaluation: phonological rules and FSTs. The phonological rules were extracted from the phone output of the HMM-recognizer aligned to the reference transcription. The weights of the FSTs were derived from the confusion matrix of the HMM-recognizer output on the rhyme-test. Both approaches for pronunciation will be described in detail in the following sections.

4.2. Pruning

It turns out that having too many pronunciations in the dictionary reduces the recognition rate as the words in the dictionary are more likely to be confused [3]. This is due to the nature of the Viterbi algorithm that aims to return the best phone sequence not the best matching word and 'biases against words with many pronunciations, since the probability mass is split up among more pronunciations'[10].

What has to be taken into account as well, especially when working with dysarthric speech, is that a phone confusion can also be a result of some random irregularity in the individual speech sample analyzed, e.g. through distortions like breathing sounds. As described in chapter 2 these phenomena are known to be common in dysarthric speech.

This means it is eligible to add only the 'best' new pronunciations to the lexicon. There are several approaches to rank a set of candidate pronunciations. Most of them are based on the probability of the candidate phone sequence. From this ranking many simple pruning methods can be applied to the list of ranked pronunciations:

- select the n most likely variants - with n being a predefined value
- select all variants with a probability higher than a threshold t
- discard all variants with a probability 'less than some fraction f of the most likely pronunciation'[3]

In this work the pronunciation variants generated with phonological rules were ranked according to their probability. For pruning a hard as well as relative threshold was used.

Another possibility to rank a set of candidate pronunciations is based on the degree of confusability. Pronunciation variants are only allowed to be confusable with other words in the lexicon to a certain degree [4, 35]. The degree of confusability can e.g. be measured by computing the likelihood for a phone sequence for each of the confusable word. A candidate pronunciation is pruned e.g. if a predefined number of words is more likely for the phone sequence than the input word. This pruning method was used in this work in combination with FSTs.

Another method for pruning are confidence measures [4], were the pronunciation variants are selected to maximize the confidence that a word is actually pronounced like the candidate phone sequence according the acoustic model.

Also the similarity between the canonical transcription and a candidate phone sequence can be used for pruning as measure for the accuracy gain [24].

4.3. Phonological rules

Phonological rules [3] are a formalization in linguistics to describe how a phone changes in a certain context

$$A \rightarrow B \,|\, C \_ D \tag{4.1}$$

which simply means that phone A is replaced with phone B if A follows C and precedes D. These rules can be derived from both knowledge-based and data-driven approaches and have been used widely in research. In [35] a list of works using phonological rules can be found. Especially for data-driven approaches it is important to consider that 'a good phonological

representation describes phonological alternatives with as concise, general rules as possible'[4]. So when using aligned phone sequences a certain abstraction level has to be introduced to create a well matching rule set. In this work the vowels (see table B.2 for complete list) were divided into the following three groups for abstraction:

- free vowels and diphthongs
- checked vowels
- schwa (@)

The consonants were grouped according to the manner of articulation in SAMPA:

- plosives
- phonemic affricates
- fricatives
- sonorants

The grouping of the phones in the phone set presented in appendix B corresponds mainly to the grouping used for this approach, with the major difference that the diphthongs and free vowels group were merged.

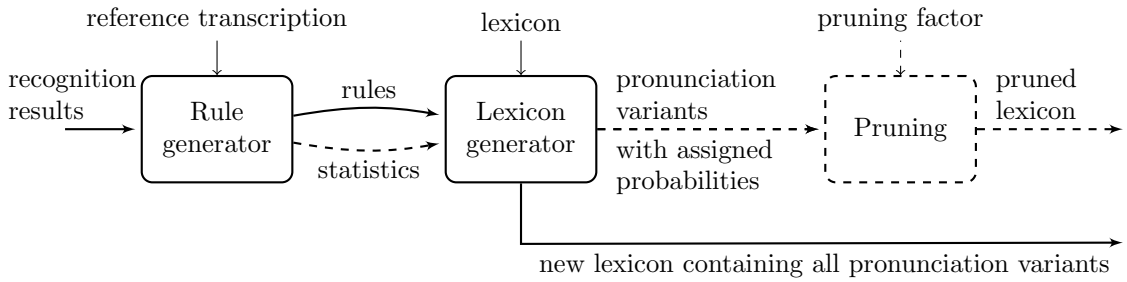

Figure 4.1.: Lexical adaptation implementation using phonological rules.

Figure 4.1 shows the main steps used to generate new pronunciation variants with phonological rules. First the recognition results of a certain HMM-model on the rhyme-test are evaluated to extract a set of rules. This was done for each dysarthric speaker for the SD, SI and the adapted SI-model with a data-driven approach to extract both the rule sets and corresponding phone statistics from a given recognition result and a reference transcription. For one dysarthric speaker (`M002`) a set of rules for each model was also designed manually. In this case phone statistics were not taken into account.

The found rule set was applied to the canonical transcriptions of a given lexicon to generate new pronunciation variants. In case of the data-driven approach the phone statistics are used to calculate the probabilities of the generated pronunciation variants in the new lexicon. Based on the probabilities two pruning methods are applied to discard unlikely pronunciation variants.

4.3.1. Manual rule generation

The rhyme-test evaluation results of the SD, SI as well as the adapted SI acoustic models along with the corresponding reference transcription were used to extract phone confusions with `HRESULTS` for speaker `M002`. The sound samples and the aligned transcriptions by the recognizers were also analyzed with Wavesurfer, an open-source tool that supports sound visualization [32] and Matlab^{TM}. This way the sound samples, along with the spectrogram and the aligned HTK-transcriptions could be visualized giving a more detailed insight why the particular phone confusions occur. The sound samples and aligned transcriptions of the ref-

T	A	B	/	C	D	Description
S	d	l	/	*	Vc	Replace /d/ with /l/ before a checked vowel
S	ts	f	/	*	Vc	Replace /t͡s/ with /f/ before a checked vowel
S	N	n	/	*	Cp	Replace /ŋ/ with /n/ before a plosive
S	i:	aI	/	*	Cp	Replace /iː/ with /a͡ɪ/ before a plosive
I	p	m p	/	*	*	Insert an /m/ before any /p/
I	p	p sp	/	*	C	Insert a short pause after /p/ before a consonant
I	t	sp t	/	+	*	Insert a short pause before /t/ except at the beginning of a word

Table 4.2.: Examples of the manual rule set derived from the analysis SI triphone model the format in which the rules are written.

erence speaker `M000` were analyzed in the same way to compare the manner the phones are pronounced differently by `M002`.

From this analysis the pronunciation of particular sounds by `M002` could be identified as reason for many confusions that occurred in the rhyme-test data. In other cases disfluencies in `M002`'s speech caused confusions or deletions, so rules allowing short pauses within words between certain phones were also introduced. Other confusions occurred due to distortions caused by the rise of loudness in the voice of `M002` at the beginning or end of a word which was already described in chapter 2. These confusions were ignored in the manual rule generation and showed the importance to implement pruning methods when rules generated by data-driven methods are in use for generating pronunciation variants.

The rules found for each acoustic model were stored in a file. Each rule is written in one line in the form 'T A B / C D'. T defines the type of the following rule (S for substitution, I for Insertion and D for Deletion). The context of the rule was generalized as far as possible, in some cases it was possible to find more general groups than described in the previous section. Examples for derived rules are shown in table 4.2. Rules of that form can then be read by the lexicon generator used for generation of the pronunciation variants. More details about findings during the analysis, as well as recognition results of the extended lexica can be found in section 5.4.1.

4.3.2. Data driven rule generation

The data-driven rule generation is based on a given recognizer output and a reference transcription. The aligned phone confusions generated by `HRESULTS` are parsed iteratively to find all substitutions, deletions and insertions along with their phonetic context. The number of occurrences of each phone from the label set in the test set $\#A_i$ is also determined based on the reference transcription and is used for calculation of phone statistics. The probability for each confusion between all phones A_i and B_j in a specific context $P(B_j|A_i, C, D)$ is calculated as

$$P(B_j|A_i, C, D) = \frac{\#(B_j|A_i, C, D)}{\#A_i}, A_i \neq B_j \tag{4.2}$$

where $\#(B_j|A_i, C_i, D_i)$ is the number of times phone A_i occurs in the phonetic context C_D in the reference transcription and is incorrectly recognized as phone B_j. For every confusion between the phones A_i and B_j in a different context a separate rule is generated. To avoid the generation of many rules for the same confusion with a low probability the context phones

T	A	B	/	C	D	$P(B_j\|A_i, C, D)$	Description
D	k		/	Cs	$	0.1111	k is deleted after a sonorant at the end of a word with a probability of 11.11%
I	g	d r	/	^	Vc	0.2000	/g/ is replaced with d r at the beginning of the word before any checked vowel with a probability of 20%
S	C	m	/	Vf	$	0.3333	/ɕ/ is replaced with /m/ after a free vowel at the end of a word with a probability of 33.3%
S	N	n	/	Vc	Cp	0.5000	/ŋ/ is replaced with /n/ between a checked Vowel and a plosive with a probability of 50%

Table 4.3.: Examples for automatically generated phonological rules.

were grouped according to phonetic similarities, as described at the beginning of section 4.3.

The generated rules and their probabilities are written to a file in the same form as described in the previous section, but in this case an additional column for the probability $P(B_j|A_i, C, D)$ is added. Table 4.3 shows some examples for automatically generated rules for speaker `M002` based on the SI-model recognition results on the rhyme-test.

For the phone statistics the script also calculates the probability for the phone A_i to be recognized correctly based on all confusions found for the phone

$$P(A_i) = 1 - \frac{\sum_{j=1}^{L} \#(B_j|A_i)}{\#A_i}, j \neq i \tag{4.3}$$

where $\#(B_j|A_i)$ is the number of times phone A_i is recognized as any other phone B_j from the label set containing L phones in total, with $B_j \neq A_i$. The list of phones and their probabilities are written to a second file. Based on these statistics the probability of the canonical transcription of a word $p_{canonical}$ can be calculated as product of the probabilities of each phone A_k contained in the transcription.

$$P(p_{canonical}) = \prod_{k=1}^{K} P(A_k) \tag{4.4}$$

The calculations also have to take into account the special case in which a phone is never recognized correctly. As the probability of the canonical transcription is the product of the probabilities for each phone $P(A_k)$ setting the probability of one phone in the label set to zero would also cause the probability of any canonical transcription containing this phone to be zero. Therefore in that case the calculation of the probability $P(B_j|A_i, C, D)$ is modified from eq. (4.2) to

$$P(B_j|A_i, C, D) = \frac{\#(B_j|A_i, C, D)}{\#A_i + 1} \tag{4.5}$$

which is equal to the assumption that there is one imaginary correct recognition of phone A_i. The probability for the phone A_i is then also modified to

$$P(A_i) = 1 - \frac{\sum_{j=1}^{N} \#(B_j|A_i)}{\#A_i + 1}, i \neq j \tag{4.6}$$

4.3.3. Lexicon generation

The generation of the lexicon was implemented independently from the rule generation. This way both types of generated rules could be handled in the same way.

A given file containing a rule set is read line-by-line and for each rule a structure is stored that contains the replacement phone B_j, as well as a Perl regular expression with the center phone A_i and its generalized left and right context (C and D), e.g. the regular expression for the second rule in table 4.3 look like this `(i:|e:|E:|a:|o:|u:|y:|2:|aI|aU|OY)C($)`. In case of the data-driven approach the probability of the rule is also added to the rule-structure. All rule-structures are stored in a rule set @*rules*.

For the data-driven approach the probability of all canonical transcriptions is calculated as in eq. (4.4) using the phone statistics computed together with the rule set. Both the canonical transcription and its probability are stored in a structure @*canonical_pron* to which the phonological rules are applied.

Algorithm 1 shows a pseudo code that illustrates the generation of new pronunciation alternatives based on a given set of rules @*rules* and a canonical transcription structure @*canonical_pron*. The function returns an array of alternative pronunciations @*alternatives* with each entry similar to the structure of @*canonical_pron* which is also added to the array as first entry *alternatives*[0].

The transcriptions stored in @*alternatives* are iteratively segmented into slices, starting from the first phone, each slice containing the current center phone with its left and right phonetic context $C\_\ D$. All rule structures that match the current center phone of the canonical transcription are then loaded and the regular expressions are matched against the current slice. If a rule matches the slice of the variant from @*alternatives* currently analyzed a new variant is generated by copying the entry from @*alternatives* and exchanging the current center phone with the new center phone stored in *current_rule.B*. The probability of the new pronunciation variant is calculated as

$$P(p_{new}) = \frac{P(p_{prev}) \cdot P(B|A_i, C, D)}{P(A_i)} \tag{4.7}$$

where $P(A_i)$ is the probability of the original center phone taken from the phone statistics and $P(w_{prev})$ is the probability of the transcription of the existing entry from @*alternatives* to which the rule has been applied. This can be either the canonical transcription (*alternatives*[0]) or any variant already generated from it.

The same slice of each pronunciation variant stored in @*alternatives* is matched against all regular expressions for the current center phone of the canonical transcription before the next segment is analyzed. Consequently the segmentation of a new variant always starts at the position of the phone after the one that matched the regular expression, as the center phone of the current slice of the new variant does not match the center phone in the canonical transcription after the application of the phonological rule.

In a final step the probabilities of the resulting pronunciation variants of the function *apply_rules* are normalized to sum up to 1 for each word in the lexicon. All transcriptions and their probabilities are written to a new lexicon file.

Algorithm 1 Application of a rule set to the canonical transcription of a word

```
1: function APPLY_RULES(@canonical_pron, @rules, %phone_rating)
2:     push (@alternatives, @canonical_pron)
3:     @original_phonemes = split_into_phones(canonical_pron.transcription)
4:     for i = 0;i <=length(@original_phonemes);i = i + 1 do
5:         for all @alternatives do
6:             @phonemes = split_into_phones(current_alternative.transcription);
7:             if i == 0 then
8:                 slice = @phonemes[i..i + 1]
9:             else if i == length(@phonemes) then
10:                slice = @phonemes[i − 1..i]
11:            else
12:                slice = @phonemes[$i − 1..$i + 1]
13:            end if
14:            @current_rules=get_rules_for_phone(@rules, original_phonemes[i])
15:            for all @current_rules do
16:                if match(slice, current_rule.A, current_rule.C, current_rule.D) then
17:                    new_pron.transcription = current_pron.transcription
18:                    new_pron.transcription[i] = current_rule.B
19:                    if defined(phone_rating− > @original_phonemes[i]) then
20:                        @new_pron.prob = current_pron.prob·current_rule.prob / %phone_rating→{@original_phonemes[i]}
21:                    else
22:                        @new_pron.prob = current_pron.prob · current_rule.prob
23:                    end if
24:                    push( @alternatives, @new_pron);
25:                end if
26:            end for
27:        end for
28:    end for
       return alternatives
29: end function
```

The manually generated rules do not have assigned probabilities. Instead a fixed internal value v_p is assigned to each rule when a rule set without assigned probabilities is loaded. In this case also the phone statistics are undefined and the value assigned to the canonical transcription is set to 1. The computation of the pronunciation variants is done in the same way as for the data-driven rules and the updates for the values stored along with the newly generated transcriptions are calculated as

$$P(w_{new}) = P(w_{prev}) \cdot v_p \tag{4.8}$$

So in this case instead of using probabilities the fixed internal value can be seen as penalty for any variation added to a given pronunciation. Due to the normalization of the pronunciation variants of each word to 1 the canonical transcription has by definition the highest score. The new variants to which one rule was applied score second best, in case that two rules were applied the score is the third highest and so on.

4.3.4. Pruning

Two different pruning mechanisms were implemented to reduce the number of generated pronunciation variants from the data-driven rules. The first is a hard pruning threshold. From the list of pronunciation variants all candidate transcriptions p_{cand} having a normalized probability below a certain threshold t

$$P(p_{cand}) \leq t \tag{4.9}$$

are discarded. The canonical transcriptions and the remaining variants are written to a new lexicon file.

The second method uses a relative threshold. From a set of candidate pronunciations p_{cand} for a word the one with the highest probability is selected and an individual pruning threshold t for this word is calculated as

$$t = P(w_{canonical}) \cdot f \tag{4.10}$$

by multiplying the probability with a given pruning factor f. For each word only the candidate transcriptions fulfilling eq. (4.9) are added to the new lexicon.

The same pruning methods could also be applied to the manually generated rule set. But in this case the pruning factor is not controlling the minimum probability, but the number of allowed changes applied to the canonical transcription relative to the total number of generated pronunciation variants. Because of the normalization the values assigned to each variant get lower with the number of total pronunciation variants generated for the word. However, no evaluations using that approach are presented in this work.

4.4. Weighted finite state transducers

In a simple way one can think of an FST-network as 'a machine that reads one string and generates another'[10]. Formally the FST is defined by the following parameters:

- a finite set of K states $Q = \{q_1 \cdots q_k\}$
- a start state $q_s \in Q$
- a set of final states $\in Q$
- a finite set of input symbols
- a finite set of output symbols
- a transition matrix between the states

There are two basic operations that can be applied an FST T that are useful in many applications:

- Inversion of an FST T^{-1}: The input and output symbols of the transitions are exchanged, so the inverted transducer T^{-1} maps the output symbols to the input symbols of the original FST T.
- Composition of two FSTs $T_1 \circ T_2$: The resulting FST maps input symbols of T_1 to the output symbols of T_2.

A weighted finite state transducer (WFST) is formally defined in the same way as an FST with the addition that each transition $a : b$ has an assigned cost $c(a : b)$.

In this work a WFST was used to generate new pronunciation variants based on the ideas from [4]. In this approach both the acoustic model and the language model are represented as FSTs. New variants are generated by composition of the language model FST with the acoustic

model WFST that holds the confusion information of the recognizer. The new variants are evaluated regarding their confusability with other words and their gain in accuracy.

For evaluation of this approach for the dysarthric speakers an existing framework implemented in [24] was used. The framework was left unchanged, only the configuration was adapted. New pronunciation variants were generated and added to the existing lexica for the SD, SI and adapted SI-models for the dysarthric speakers.

4.4.1. WFST representations

The acoustic model WFST C was built based on the confusion matrix generated from the recognizer output of the rhyme-test. The input and output symbols of C are the L phones from the label set of the acoustic model and two additional symbols to represent an insertion and deletion. C has one single (input and output) state with weighted transductions between all symbols in input and output set. The normalized confusion matrix Θ for a given speaker and acoustic model was used to assign a weight to each transition between an input symbol a_i and any output symbol b_j with $i, j = 1 \cdots L + 2$ and N being the number of phones. The costs for each transition $c(a_i : b_j)$ were calculated from the confusion matrix entry of the corresponding phones $\Theta[A_i, B_j]$ as follows:

$$c(a_i : b_j) = -\log \Theta[A_i, B_j] \tag{4.11}$$

To avoid numerical problems because of the computation of the *log*-function Θ was modified in two ways. To all entries of Θ a small offset (10^{-5}) was added to avoid zero entries in the confusion matrix. In addition if no sample of phone A_i was recognized correctly the corresponding entry in the main diagonal of Θ was changed to $\Theta[A_i, A_i] = \Theta[A_i, A_i] + 1$.

An example for a small subset of an acoustical model WFST C with the weights for each transduction $c(a_i, b_j)$ calculated from the confusion matrix generated from the SI-model rhyme-test recognition results of speaker `M002` is shown in figure 4.2.

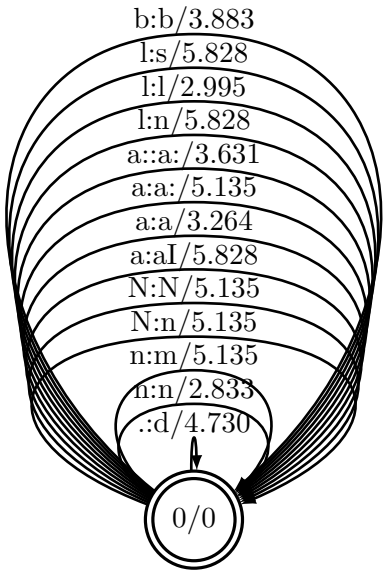

Figure 4.2.: Example for an acoustic model WFST C for a small subset of phones. The assigned weights for each transition are based on the rhyme-test results from `M002` on the SI-model.

The pronunciation lexicon FST L models the transitions of words into phones. The translator is built directly from a given lexicon with the words as input symbols and corresponding phone sequences as output symbols. By inverting the lexicon transducer L a pronunciation model FST P that translates phone sequences into words is generated. An example for a pronunciation model transducer P is shown in figure 4.3.

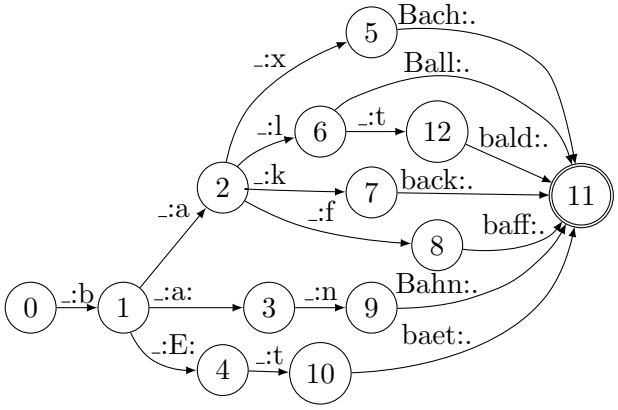

Figure 4.3.: Example for a pronunciation model FST P translating phone sequences into words from the rhyme-test vocabulary.

4.4.2. Pronunciation variant generation

For a given word FST w_i, that consists of a single state with one transition path containing the word, the canonical transcription can be found by composition with the lexicon transducer L. By composition of the result with the acoustic model WFST C a phone network p_{cand} is generated.

$$p_{cand} = w_i \circ L \circ C \tag{4.12}$$

This WFST contains the phones of the canonical transcription as well as the phone-level confusions with the corresponding weights. As C contains transitions between all phones in the label set the phone network p_{cand} is pruned to discard unlikely transitions. An example for a pruned phone network can be found in figure 4.4. It shows the result of the composition of the word 'Ball' [bal] from the rhyme-test word list with the lexicon transducer FST L and the acoustic model transducer C of the SI-model with the phone confusions detected for `M002` of which a subset is shown in figure 4.2.

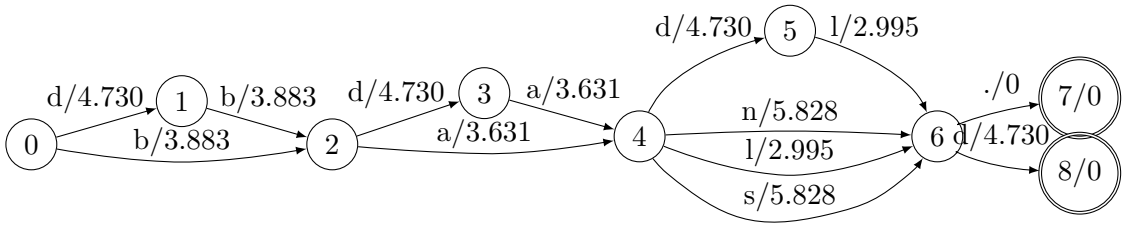

Figure 4.4.: Example for the pruned composition as described in eq. (4.12) for the word 'Ball' from the rhyme-test word list and the confusion matrix FST C from the Speechdat-database model evaluated on `M002`.

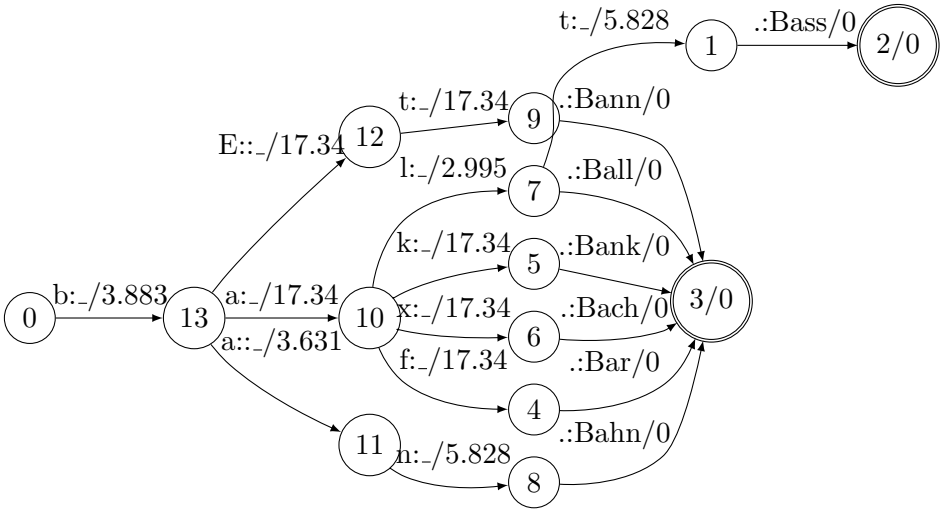

Figure 4.5.: Composition example for the composition of the possible pronunciation variants p_{cand} of the word 'Ball' with the lexicon FST P as described in eq. (4.13) and the confusion matrix FST C from the SI-model evaluated on `M002`.

By summing up the costs along the paths all possible phone sequences can be ranked. As the canonical transcription would always have the lowest cost and therefore the best ranking it is discarded.

4.4.3. Confusability score

The composition of the phone network p_{cand} with the pronunciation model FST P translates the phone sequences into words of the lexicon w_{conf}:

$$w_{conf} = w_i \circ L \circ C \circ P \tag{4.13}$$

The returned words different from the input word w_i can be considered as confusable words. Figure 4.5 shows an example for the confusable words w_{conf} resulting from the phone network p_{cand} of 'Ball' shown in figure 4.4.

By summation of the costs along each path the most likely words that are returned by the phone network p_{cand} can be determined. Again the canonical transcription has to be removed before a ranking of the words can be done starting from the word with the lowest costs like shown in figure 4.6. In this example the rank of the input word 'Ball' is 2, while the word 'Bahn' has the highest rank 1, which indicates that the most likely phone sequence generated from p_{cand} matches the word 'Bahn' better for the given acoustic model WFST.

The rank of the input word can be seen as confusability score. The higher the rank the more confusable is the phone sequence and the more likely additional recognition errors are introduced when the variant is added to the lexicon. In the used framework a minimum rank of the input word can be configured for pruning of unsuitable variants. If the input word has a rank lower than the confusability score no new variant is generated from the given phone network p_{cand}. The implementation of the calculation of the rank in the framework also handles the case that words have an equal ranking: All words with a higher ranking than the input word are assigned to a separate rank also when the costs are equal to an other word in the ranking. The input words and all words with equal costs are assigned to the same rank, so in

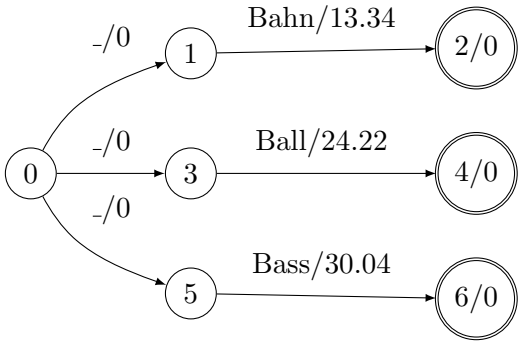

Figure 4.6.: Ranking example for the composition of the possible pronunciation variants p_{cand} of the word 'Ball' with the lexicon P as described in eq. (4.13) and the confusion matrix FST C from the SD-model.

this case there is no influence on the confusability score.

During this work it turned out that the internal pruning factors are a critical factor for a proper ranking of the candidate words. When the parameters are set too tight, it happens that all returned words of p_{cand} are pruned. In this case the script automatically sets the rank of the phone sequence to a high value as no analysis of the actual ranking can be done. To avoid this the internal factors were set to high values. In contrast for the graphs shown in this section the pruning factors were manually set to lower values to achieve a reasonable size of the graphs with representative contents.

4.4.4. Accuracy gain

The most likely phone sequence returned from the candidate phone network p_{cand} is analyzed regarding its accuracy gain. Using the values of the normalized confusion matrix from the rhyme-test evaluation of the acoustic model the string edit distance between the canonical transcription and the returned phone sequence is calculated, for the dysarthric speaker analyzed as well as the reference speaker, based on the confusion matrix generated from the same acoustic model evaluated on the rhyme-test.

The higher the difference between both string edit distances, the lower is the probability that the same phone sequence would be generated by the reference speaker for this word and the lower is the gain in general accuracy. If the difference exceeds a certain threshold the phone sequence is discarded. For this work the gain in accuracy had to be set carefully, as the phone sequences from dysarthric speakers do not necessarily have to be very likely for the reference speaker.

5. Evaluation

The performance of SI and SD acoustic models with monophone and triphone subword modeling were evaluated on data recorded from the five dysarthric speakers as well as an unimpaired reference speaker. Whole word models were evaluated on a small vocabulary connected digits task. In addition the proposed adaptation approaches on both, acoustical and lexical level described in chapter 3 and chapter 4, were evaluated and the changes of the recognition performance of the different speech recognizers was analyzed and compared.

Perl scripts were implemented to evaluate a given acoustic model on a list of speech samples. The loading of the speech data as well as the feature extraction was done in the same way as in the training scripts. The methods from Refrec0.96 were again used for lexicon preprocessing. The Viterbi-decoding of the speech samples was done using `HVite` from the HTK-toolkit [39]. The basic evaluations of the recognizer output were done using `HResults`, which automatically calculates the recognition rate on both, sentence and word level given the reference transcription of the test data.

To evaluate the SI acoustic models on the speech data recorded from the dysarthric speakers it was necessary to downsample the recordings to 8 kHz and apply a bandpass filter.

Different task specific grammars were used as language model for each task. For the command word task and rhyme-test isolated word grammars were used while the connected digit task was evaluated using a loop grammar. The Refrec0.96 framework provided functions to automatically generate a proper grammar representation that can be interpreted by HTK from a given word list.

5.1. Measures

For the command word and the rhyme-test evaluations the recognition rate on sentence level and its changes through different adaptation techniques were of main interest. The recognition rate on sentence level r_s is calculated as

$$\begin{aligned} r_s &= 100 - \mathrm{SER} \\ &= 100 - \frac{S}{N_s} \cdot 100 \end{aligned}$$

where S is the number of sentences incorrectly recognized and N_s is the total number of sentences in the corresponding test set.

For the connected digits task it was also useful to evaluate the results on word level, where two different values of interest are calculated by `HResults`. The recognition rate on word level r_w is

$$\begin{aligned} r_w &= 100 - \mathrm{WER} \\ &= 100 - \frac{H - S - D}{N_s} \cdot 100 \end{aligned}$$

where H is the number of sub-items (phones or words, depending on the configuration of `HResults`) that occur in the test samples, D is the number of deletions and S is the number of substitutions detected in the recognizer output. The recognition accuracy a also takes into account the number of insertions I in the recognizer output

$$a = 100 - \frac{N_s - D - S - I}{N_s} \cdot 100 \tag{5.1}$$

The values for r_s and r_w are always between 0% and 100%, as the number of substitutions and deletions cannot be higher than the number of words in the test set. In contrast the accuracy a can also be negative when a high number of insertions leads to $S + D + I > N_s$.

`HResults` also generates other useful outputs such as the alignment of the reference transcription to the recognizer output or a confusion matrix. This data could be used for further evaluations.

For the adapted HMM models the relative improvement of recognition rate in comparison to the baseline system was also of interest on both, word and sentence level. It is calculated as

$$\Delta r_s = r_{s_{adapt}} - r_{s_{base}} \tag{5.2}$$

in case of the improvement on sentence level and as

$$\Delta r_w = r_{w_{adapt}} - r_{w_{base}} \tag{5.3}$$

in case of the improvement on word level. The improvement is a suitable measure to show how well a particular adaptation method works, but the actual recognition rate achieved was of major interest for most evaluations.

5.2. Connected digits recognition

In this evaluation the performance of SI and SD acoustic models for dysarthric speakers is compared on a task containing only a very small vocabulary of 10 digits. For the impaired speakers as well as the reference speaker SD HMM-models with monophone, triphone and whole word modeling were trained using 80 recordings of 4-digit sequences (see section 3.2.1). The SI-model used for this tasks was trained from the B and C sub-corpora of the Speechdat-database which contain recordings of connected digit sequences of various length (see table 3.1). Details about the training procedures for all model types can be found in section 3.2.2. The trained acoustic models were evaluated using a digit loop grammar as language model.

The evaluation of the recognition rate of the different speech recognizers was done on word level. The accuracy had to be taken into account as well to measure the recognition performance because evaluations showed that all systems introduced a large number of insertion errors if the p-parameter controlling the 'word insertion log probability'[39] was not set to a properly low value for the recordings of the dysarthric speakers. In some cases more than 25 insertions errors occurred during the evaluation of the SI-recognizer on a test set containing only 10 utterances (40 digits).

Figure 5.1 shows how the setting of the p-parameter influences the number of insertion errors of both, the SI and SD triphone acoustic models for speaker `M066`. The utterance analyzed is the connected digit sequence 'zwo vier fünf eins' (two four five one). The aligned transcriptions shown in green are the recognizer outputs of the SI triphone model with two different settings for p. The first row shows the recognizer output with $p = 0$ while the second row shows the

result for $p = -100$. The transcriptions printed in blue are the recognizer outputs of the SD triphone model trained for `M066` with the same p-values used as for the SI-model.

The spectrogram clearly shows that the disfluences in the second and third word trigger insertions in the SI-recognizer. The number of insertion errors is reduced from three to one when a lower p-value is used. However, the disfluency in the second word still leads to an insertion in that case. The same insertion error occurs in the SD-recognizer with $p = 0$, but with $p = -100$ the sequence is recognized correctly.

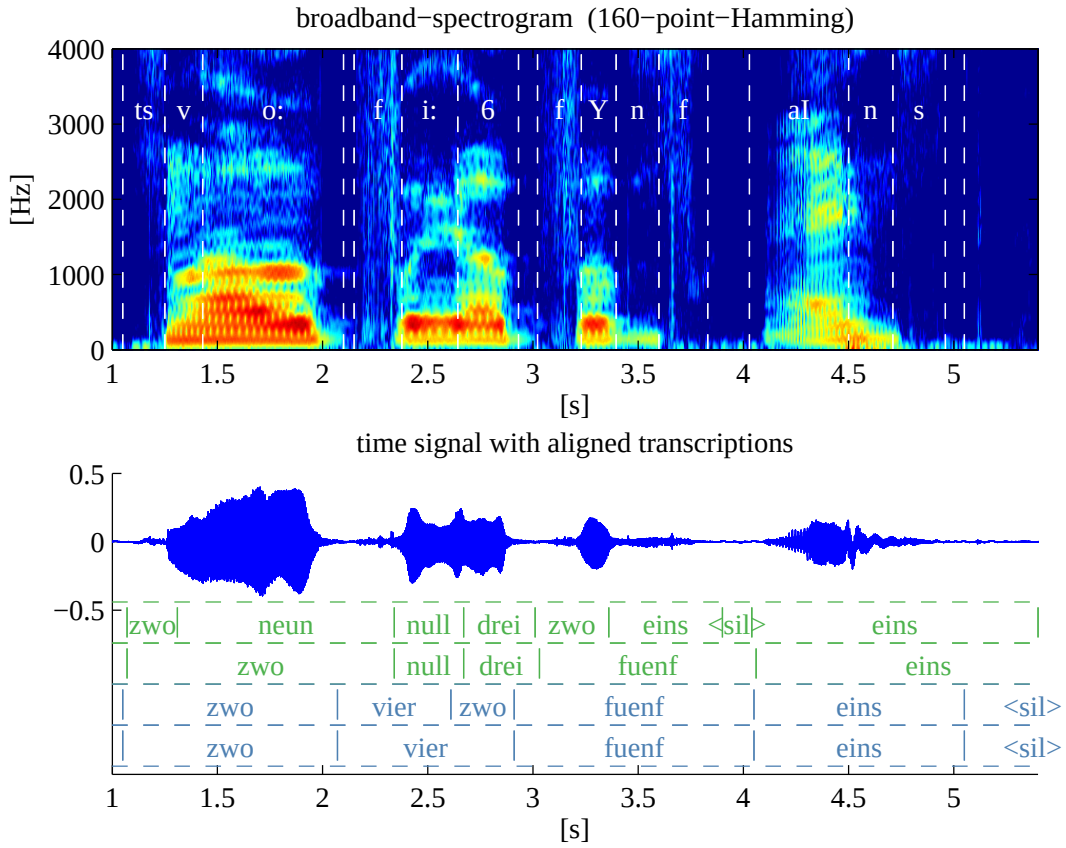

Figure 5.1.: The spectrogram of the connected digits sequence 'zwo vier fünf eins' uttered by `M066` and a comparison between the labels assigned on word level by the connected digit SI (green) and SD (blue) triphone recognizers using two different values for the word insertion log probability p. In the first and third alignment $p = 0$ and in the second and fourth alignment $p = -100$. The SAMPA-labels of the phones in the spectrogram were assigned manually.

The same sequence uttered by the reference speaker `M000` and the aligned transcriptions of the SI (green) and SD (blue) model with $p = 0$ is shown in figure 5.2. A comparison of the duration of the digit sequences uttered by `M000` and `M066` shows that the utterance of `M066` is longer and vowels are more stretched. Looking at the aligned recognizer outputs for the first digit 'zwo' [t͡svoː] an insertion error occurs in the SI-recognizer when $p = 0$ where most of the stretched vowel /oː/ is replaced by a new digit. A comparison shows that the duration of /oː/ is about 1.5 times longer (~ 0.75 s) than for `M000` (~ 0.5 s).

The final sound of the second word 'vier' [fiːɐ] is also stretched by `M066`. At the transition

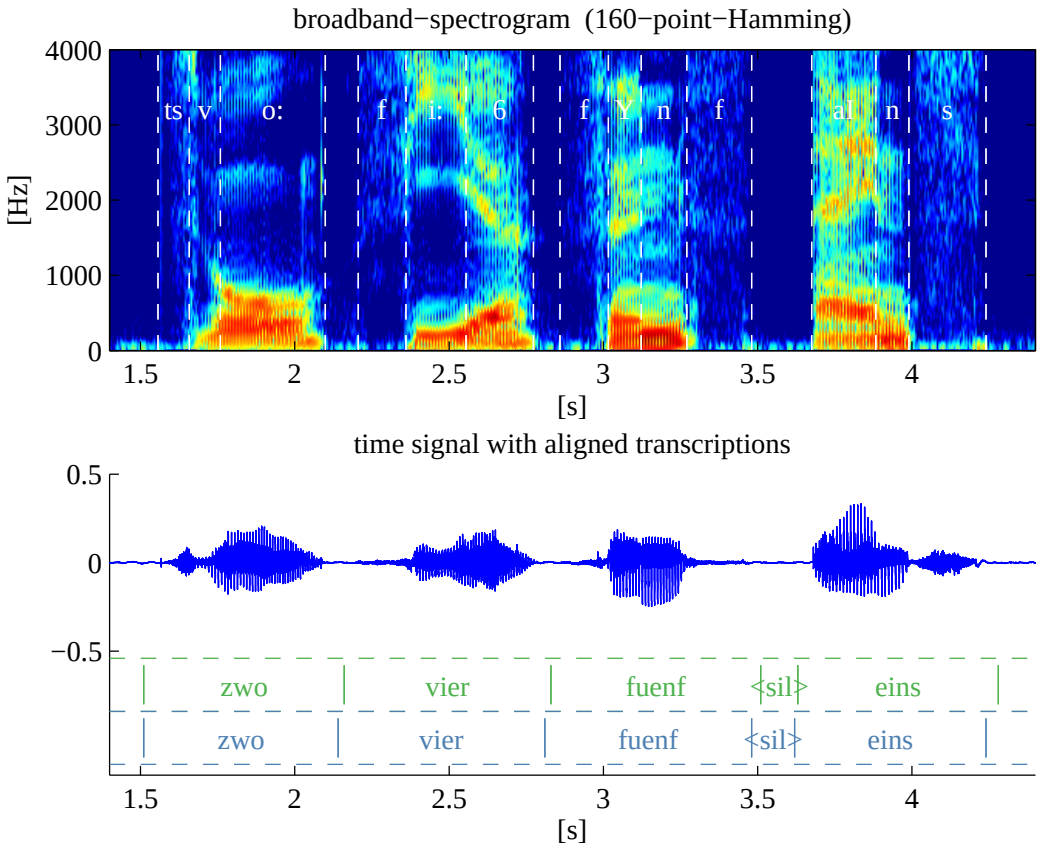

Figure 5.2.: The spectrogram of the connected digits sequence 'zwo vier fünf eins' uttered by `M000` and a comparison of the labels assigned on word level by the connected digit SI (green) and SD (blue) triphone recognizers with the word insertion log probability set to $p = 0$. The SAMPA-labels of the phones in the spectrogram were assigned manually.

between the front vowel /iː/ and central vowel /ɐ/ there is a notable jump in the formant structure. In contrast to `M000` the formant structure of the final sound /ɐ/ uttered by `M066` is more constant. This indicates problems with continuous movements of the tongue during the phone transition. In addition the tongue position during the final sound /ɐ/ is higher and more in the back, as the utterance sounds more like /ʌ/.

The pronunciation of the third digit 'fünf' [fʏnf] also leads to insertions in case of the SI-model for $p = 0$. The utterance of the nasal /n/ is quite mute and sounds more like an /m/. In the spectrogram the concentration to frequencies lower than 500 Hz typical for a nasal [17] is visible, but the intensity is clearly weaker compared to the neighboring sounds or the utterance of `M000`. There is also less measurable activity in the spectrogram for frequencies over 500 Hz compared to `M000`. The last digit 'eins' [a͡ins] is recognized correctly in all recognizer settings. One can see from the formant structure that speaker `M066` has less problems with the pronunciation of /n/ after the vowel /a͡ɪ/, as the intensity of the sound is higher and more activity could be measured in the spectrogram for frequencies over 500 Hz.

5.2.1. Parameter selection

For selection of the recognizer parameters possible combinations of acoustic models and p-values were evaluated for each speaker on a development set containing 10 recordings. For the monophone and triphone acoustic models the number of GMMs used to calculate the acoustic likelihoods was varied between 2 and 16 GMMs during this evaluation. The acoustic models were evaluated using p-values $0 \geq p \geq -150$ for recognition. The selection of the acoustic model parameters in combination with a good value for p was done manually. By observation of the average values for the recognition rate r_w and the accuracy a over the evaluated p-parameters the best acoustic model for each speaker was determined. The model that performed best for the majority of the dysarthric speakers was then selected along with a proper p-value.

In figure 5.3 the average number of insertion errors over the different p-values used for the evaluated acoustic models with a different number of GMMs is shown for the dysarthric speakers for both the SI and SD triphone models. The number of insertion errors was strongly dependent on the individual speaker, but in general the number of errors decreased when acoustic models with a larger number of GMMs per state were used. The increasing model complexity leads to a reduction of the number of insertion errors. The number of errors that occurred in the SD-model for all dysarthric speakers was lower than in the SI-model result for the same speaker. Comparing the ranking of the speakers from the best to the worst result between the SI and SD-model one can see that the best result is achieved by `M067` and the worst by `M063` in both models. The relative ranking of `M068` and `M002` is also the same in both models. Only `M066` is ranked second worst in the SI and second best in the SD-model. This is an evidence that the individual speaker characteristics have a general influence on the recognition performance on HMM-recognizers.

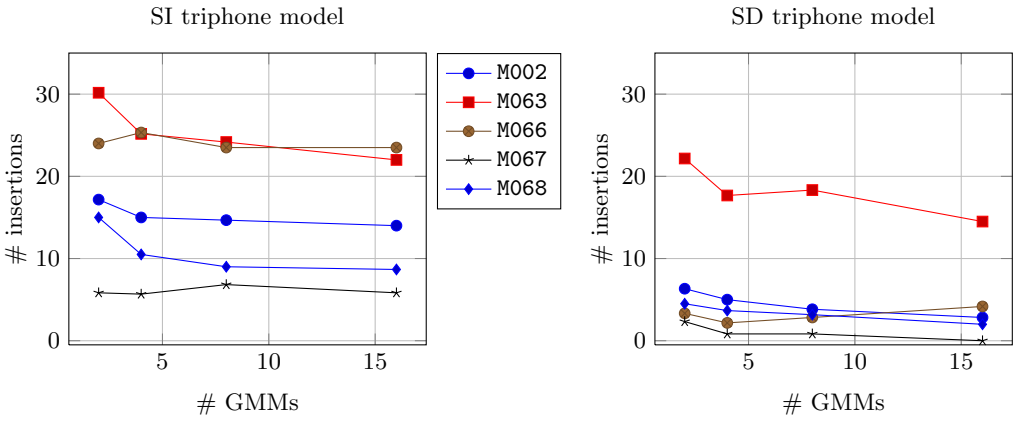

Figure 5.3.: The average number of insertion errors over the evaluated values for the word insertion log probability p on the development set for the SI-models with a different number of GMMs to calculate the acoustic likelihoods for the five dysarthric speakers.

In general using a smaller p-value reduces the number of insertion errors. However, a trade-off between fewer insertions and an increasing number of substitutions and deletions of correct digit occurrences could be observed in some of the evaluated acoustic models. The effect was present in almost all SI-models. An example for the number of errors over the different p-values is shown in figure 5.4 for the SI-model with 16 GMMs evaluated on speaker `M068`. The main reason for the strong trade-off in this case is that the large number of insertions that occur

model	#GMMs	p_{ref}	p_{dys}
SI triphone	16	-5.0	-150.0
SD triphone	8	-5.0	-100.0
SI monophone	16	-5.0	-150.0
SD monophone	8	-5.0	-100.0
SI whole word	6	-100.0	-150.0
SD whole word	6	-5.0	-50.0

Table 5.1.: The selected parameters used for connected digits evaluation. For all speakers an acoustic model with the same number of GMMs was used. For the word insertion log probability p two different values were chosen: one for the reference speaker and one for all dysarthric speakers.

with low p-values conceal errors in the alignment of the recognizer output. This effect was observed less often in the evaluation of the SD acoustic models. On average the SD-models with 16 GMMs introduced fewer insertions for some speakers, while the number of substitution and deletion errors lead to a lower average recognition rate r_w than for the 8 GMMs models. To decide which model is more suitable in this case the highest and lowest p-values were excluded from the average calculations. The new result favored the acoustic model with 8 GMMs which was chosen for further evaluation.

In case of the SI triphone recognizer the best values for r_w and a were achieved by the same acoustic model in most evaluations, which was either the one trained with 8 or 16 GMMs. When the lower p-values were ignored in the calculation of the average r_w and a the acoustic model with 16 GMMs had the best average performance and was selected.

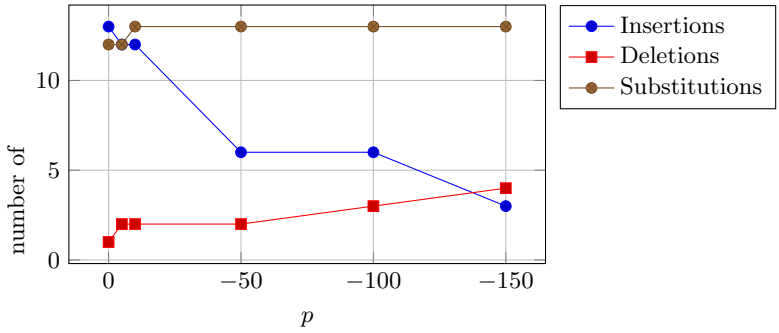

Figure 5.4.: Number of insertion, deletion and substitution errors over different number of GMMs for SI-models on development set for speaker M068.

Evaluation of the monophone models lead to similar observations. Both the SD-model with 8 GMMs and the SI-model with 16 GMMs had again the best performance and were selected for further evaluation.

A summary of the acoustic models chosen and the setting of the corresponding p-value for both the reference speaker and the dysarthric speakers is given in table 5.1. For the reference speaker the parameter p was set to -5.0 for most models although for the SI whole word recognizer it was necessary to choose a lower value of $p = -100.0$. For the dysarthric speakers the best results were achieved with values of $p \leq -50$ for all recognizers. The value for the SD-recognizers could be set to a higher value than for the SI-recognizers.

	M000		M002		M063	
model	r_w	a	r_w	a	r_w	a
SI triphone	99.17	97.50	55.28	47.50	50.28	37.78
SI monophone	100.00	100.00	51.11	46.11	51.67	40.56
SI whole word	100.00	100.00	59.44	52.50	51.67	43.89
	M066		M067		M068	
model	r_w	a	r_w	a	r_w	a
SI triphone	46.67	28.33	55.83	55.56	60.83	58.06
SI monophone	43.89	34.17	61.11	60.00	58.89	55.28
SI whole word	56.11	44.17	69.17	69.17	65.28	63.06

Table 5.2.: The recognition rate on word level r_w and the accuracy a (both in %) for SI-models evaluated on the connected digits data set.

5.2.2. Evaluation results

The selected acoustic models and parameters were evaluated on the connected digits for both the SI and SD-models individually on a data set containing 90 recordings of each speaker as the remaining 10 of the 100 recordings were used as development set for finding the parameter settings as well as for adaptation of the SI-models. The SI-model could be tested directly on the data set and results are shown in table 5.2. For the SD-model a 9-fold cross-validation was used. The results of the cross-validation can be found in table 5.3.

For the reference speaker r_w is above 95% for all models while there is great variation between the results of the dysarthric speakers. In general the SD-models perform significantly better than the SI-models for the dysarthric speakers which was expected, as the SI-models were trained from unimpaired speakers. The recognition rates of the SD whole word models are worse than for both the SD monophone and triphone models for all speakers evaluated including the reference speaker. In general the SD monophone models have an equal or slightly lower r_w than the SD triphone models in the evaluations of all dysarthric speakers.

Comparing the SI-models with different subword modeling for `M000` both the monophone and the whole word model reached $r_w = 100\%$. The SI whole word model achieved the best results for all dysarthric speakers. If the SI monophone or triphone model had the second best performance varied from speaker to speaker and it could not be clearly determined which model type is better in general from this evaluation.

The recognition rate of the SD-models was over 95% for four of the five dysarthric speakers, while r_w was below 70% for the SI-model evaluated on all dysarthric speakers, which is not a very satisfying result.

5.2.3. Acoustic adaptation

To improve the recognition rate of the SI-models for the dysarthric speakers MLLR-adaptation of the means of the GMM components modeling the observation probabilities of the HMM-states was performed on a 10-digit development set as described in section 3.3.2. The triphone models where adapted using 20 regression classes to cluster the states. For the monophone models 10 regression classes were used. The whole word models were adapted using one linear transformation for all states.

The adapted SI-models were again evaluated on the same 90 recordings. The results and improvements achieved are shown in table 5.4. For `M000` all models achieve $r_w = 100\%$ after adaptation and there are significant improvements for all dysarthric speakers. Comparing the

	M000		M002		M063	
model	avg. r_w	avg. a	avg. r_w	avg. a	avg. r_w	avg. a
SD triphone	99.72	99.44	97.78	94.72	71.39	57.22
SD monophone	100.00	100.00	96.67	91.94	70.83	41.39
SD whole word	97.22	97.22	87.78	76.94	61.11	28.06
	M066		M067		M068	
model	avg. r_w	avg. a	avg. r_w	avg. a	avg. r_w	avg. a
SD triphone	95.83	94.72	99.17	97.50	95.83	93.89
SD monophone	93.61	88.89	99.17	94.17	94.44	90.56
SD whole word	82.50	61.94	95.28	86.67	81.67	76.94

Table 5.3.: The average recognition rate on word level r_w and the average accuracy a (both in [%]) over the 9-fold crossvalidation of SD-models with different types of subword modeling on connected digits data set.

	M000			M002			M063		
model	r_w	a	Δr_w	r_w	a	Δr_w	r_w	a	Δr_w
SI triphone	100.00	100.00	0.83	79.72	78.89	24.44	75.00	73.89	24.72
SI monophone	100.00	100.00	0.00	81.11	79.72	30.00	72.22	71.39	20.55
SI whole word	100.00	100.00	0.00	79.44	78.33	20.00	61.67	61.39	10.00
	M066			M067			M068		
	r_w	a	Δr_w	r_w	a	Δr_w	r_w	a	Δr_w
SI triphone	83.89	82.78	37.22	94.17	94.17	38.34	91.11	91.11	30.28
SI monophone	76.39	74.44	32.50	92.22	91.94	31.11	91.11	90.56	32.22
SI whole word	71.39	70.56	15.28	93.06	92.78	23.89	87.50	87.22	22.22

Table 5.4.: The recognition results r_w in [%] for the MLLR adapted SI-models on the connected digits data set and the improvement compared to the baseline results presented in table 5.2 and table 5.3.

adapted monophone and the triphone models the results of the triphone models are equal or better than the results for the monophone models for all speakers except `M002` after adaptation. In case of the unadapted SI-model it was not clear which of the subword modeling types was the better choice.

For the dysarthric speakers the adaptation of the SI whole word models does not lead to improvements as high as for the monophone and triphone models. However, the improvements are still significant in that case.

The measured accuracy shows that almost no insertions occur in the evaluations of the adapted SI-models, as the difference between a and r_w is always less than 3%. In contrast in the evaluation the SI-model the difference between both values was only this low for `M067` and `M068` while the difference was greater than 4.5 % for the other dysarthric speakers (see table 5.2). This shows that the slower speech and the stretching of the vowels in dysarthric speech, which lead to insertions, could be learned well by adaptation of the acoustic likelihood distribution of the SI-model.

In figure 5.5 the results of the adapted SI-models are compared to the SD-model results. One can see that for most speakers the adapted SI triphone model is not able to achieve equal or even better results than the SD-model. However, in case of `M063` who had the worst recognition rates among all speakers the adapted SI-models with all three types of subword modeling achieve better results compared to the corresponding SD-models having seen only 1/8 of speech data from the speaker. The results for the monophone models are generally

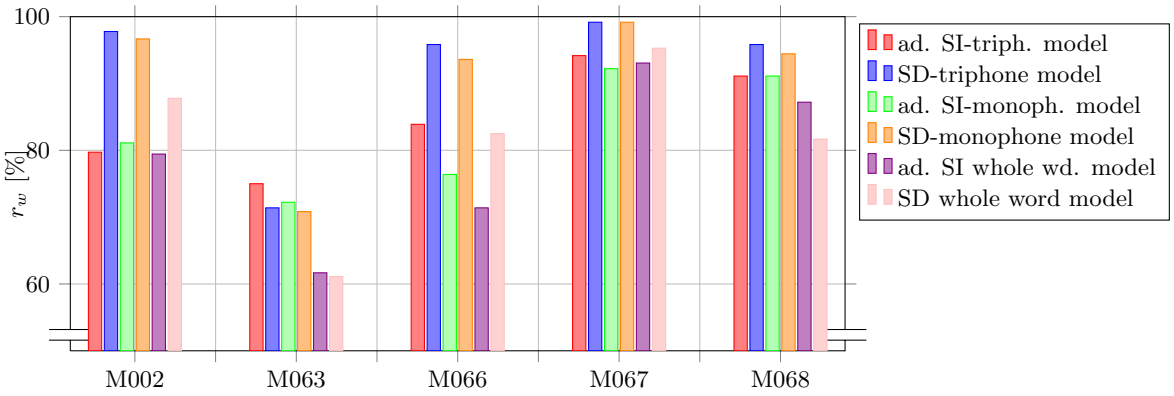

Figure 5.5.: A comparison of the recognition rate on word level r_w between the MLLR adapted SI-models (see table 5.4) on the connected digits data set and the crossvalidation results of the SD-models on the same data (see table 5.3).

slightly lower than for the triphone models. One can also see that the whole word models did not perform as well as the other models evaluated for most speakers.

5.3. Command word recognition

The recognition performance of SI, SD and adapted SI acoustic models was evaluated and compared on a second small vocabulary task containing 69 command words. Evaluations of the same acoustic models were done as well on the Sotschek rhyme-test.

The total number of command word sessions available for training and evaluation varied between five and ten complete sessions per speaker. In addition ten incomplete sessions from `M063` were available which were also used for evaluation. Details about the number of recordings of the individual speakers can be found in section 3.2.1. The complete word lists for both sets are available in Appendix B.2.

The SI-model was trained on the full training set of the Speechdat-database. Evaluation was done for all speakers on all available recordings except one session per speaker, which was excluded for acoustic adaptation. The total number of sessions evaluated on the SI-model varied between nine and four complete sessions.

From the command words recordings SD-models were trained for the reference speaker and two dysarthric speakers. For the remaining three speakers the total number of command word sessions recorded was too small for training of an SD-model. During evaluation of the SD acoustic model trained for `M063` it turned out that at least parts of the incomplete sessions of the recorded command words are distorted with background noise. Evaluation results also showed that the order of the samples in the training-data had a significant impact on the recognition rate. The recognition rate dropped when all complete sessions appeared before the distorted sessions in the training file. Therefore the training samples were shuffled once before training of the SD-model of `M063`, which helped to improve the recognition results.

For evaluation of the SD-models two approaches were used. As a large part of the command words data set used to evaluate the SI-models had been used for training the SD-model a leave-out crossvalidation was performed for `M000` and `M002`. Each session from the extended test set evaluated on the SI-model was used once for testing of an SD-model trained from the remaining sessions. It was not possible to carry out a cross-validation in the same way for `M063`

because the incomplete sessions could not be split into equal pieces containing a full command set each. Therefore both the SI and SD-models were also evaluated on one single session of the command words arbitrarily selected from each speaker, that was excluded from the test set of the SD-models.

All acoustic models were evaluated using an isolated word grammar as language model. Therefore the recognition rate was measured on sentence level in contrast to the connected digit evaluation.

5.3.1. Parameter selection

Like for the connected digit evaluation the acoustic models best suitable for the task were selected in a first step. The HMM-models trained with a different number of GMMs for estimation of the acoustic likelihoods were therefore compared. Figure 5.6 and figure 5.7 show the individual recognition results for the SI and SD-models, respectively. The results for the reference speaker `M000` on the SI-model were around 90 % in all evaluations, while all results for the dysarthric speakers were below 40%. To visualize the results achieved for the dysarthric speakers in a proper scale the results for `M000` were excluded from figure 5.6.

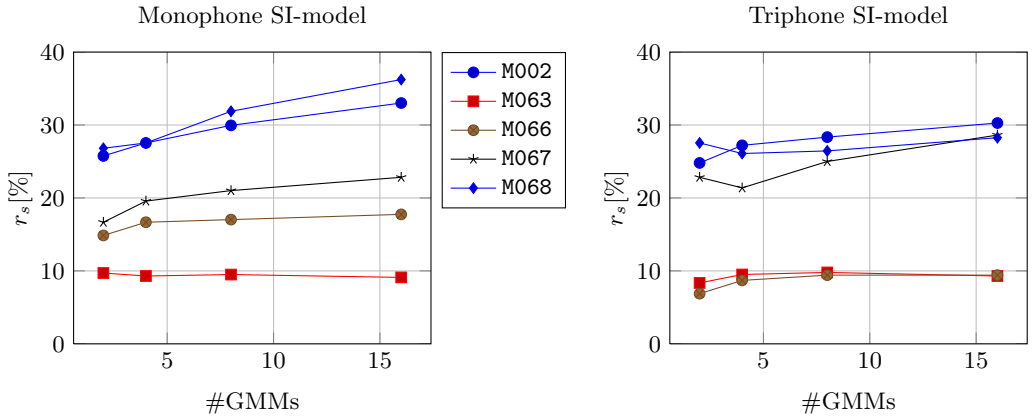

Figure 5.6.: The recognition rate r_s over the SI-models with a different number of GMMs to estimate the observation likelihood evaluated on the command words data set for both monophone and triphone subword modeling.

For the SI monophone models the recognition rate r_s for `M000` was between 86% and 89% with the best result achieved by the model with 16 GMMs. For the SI triphone models r_s was between 95% and 98%. The best result for the triphone models was achieved with 4 GMMs followed by the model with 16 GMMs. However, the difference between both results was less than 0.3%. Like in the connected digits recognition task the SI triphone models are performing very well for the reference speaker without any adaptation.

Looking at the overall results for the dysarthric speakers on the SI-models with different number of GMMs r_s tends to increase with the model complexity for both monophone and triphone subword modeling. Both, monophone and triphone models with 16 GMMs were used for further evaluation for all speakers, as they achieved the best results for four of the five dysarthric speakers evaluated. Only for `M063` the triphone model with 8 GMMs achieved better results than the model with 16 GMMs. The evaluation of the monophone model on the speech data of `M063` showed that the model with 2 GMMs performed best and the model with 16 GMMs had the worst recognition rate, but the difference between the two results was less

than 1%.

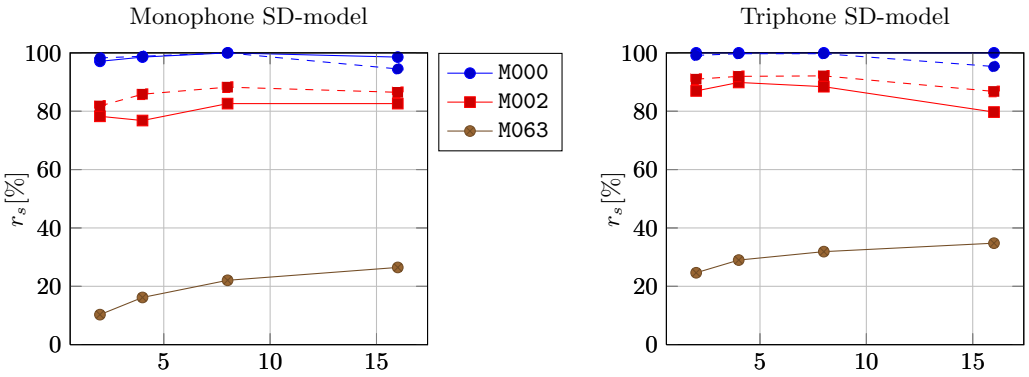

Figure 5.7.: The recognition rate r_s of the SD-model on the command words data set for monophone and triphone subword modeling with a different number of GMMs to estimate the observation likelihood. The solid lines show the result r_s on the predefined test session (see section 3.2.1) while the dashed lines show the average crossvalidation results $avg.r_s$ on the command words data set also used for evaluation of the SI-models.

The SD monophone models trained for M000 achieved the best results on the command words test set as well as for the cross-validation using 8 GMMs to model the emission probability. All SD triphone models trained for M000 achieved 100% recognition rate on the test set. The best crossvalidation results for the triphone models are $r_s = 99.71\%$ for the acoustic models with $4 \leq \#\text{GMM}s \leq 8$. For M002 the SD monophone models with 8 GMMs achieved the best results on the test set and for the cross-validation. The SD triphone models with $\#\text{GMM}s \leq 4$ performed best on the test set while the best crossvalidation result was achieved by the model with 8 GMMs. In case of M063 the SD monophone and triphone model with 16 GMMs achieved the best results.

Based on the results both, the SD monophone and triphone models with 8 GMMs were chosen. The lower recognition rate observed for M000 and M002 for $\#\text{GMM}s \geq 8$ indicates that the amount of training data available for the SD-models is not enough for proper training of the emission probabilities of each state in the acoustic model.

5.3.2. Baseline results and interpretation

Table 5.5 summarizes the selected acoustic models and the corresponding recognition rates. The values are used as baseline for comparison in all further adaptation experiments.

A comparison between the monophone and triphone model shows that the SI monophone model achieved better results for three of the dysarthric speakers (M063, M066 and M068) on the command words data set. For the reference speaker M000, as well as the dysarthric speakers M002 and M067 the results of the triphone SI-model are better. In case of the SD-models the triphone models achieved better or equal results than the monophone models for all speakers.

From the recognition rates r_s achieved by the SI and SD-model one can see that r_s is in the same range in case for M000 with both types of the subword modeling used. Both monophone models achieve lower results than the triphone models. This shows how well the SI acoustic model generalizes to unimpaired speakers. The word that caused most recognition errors in the evaluations of the speech-data of M000 on the SI-model was the word ‘simon’ [sa͡imən]. The words most often aligned were ‘vier’ [viːɐ] or ‘sieben’ [siːbən], probably because of the

	M000		M002		M063	
	test set	full set	test set	full set	test set	full set
model	r_s	r_s	r_s	r_s	r_s	r_s
SI monophone model	91.30	90.43	26.09	33.01	5.80	9.11
SI triphone model	98.55	97.68	30.43	30.27	2.90	9.30
	r_s	$avg.r_s$	r_s	$avg.r_s$	r_s	-
SD monophone model	100.00	94.49	82.61	88.25	22.06	-
SD triphone model	100.00	99.71	88.41	92.11	31.88	-
	M066		M067		M068	
	test set	full set	test set	full set	test set	full set
	r_s	r_s	r_s	r_s	r_s	r_s
SI monophone model	21.74	17.75	15.94	22.83	37.68	36.23
SI triphone model	15.94	9.42	18.84	28.62	28.99	28.26

Table 5.5.: The recognition rate r_s in [%] for the selected SI and SD-models evaluated for the five dysarthric speakers and the reference speaker on a predefined arbitrary command word test-session and on the command words data. To evaluate the latter on the SD-models a cross-validation was used in this case and the average r_s is given. These results are used as baseline for further evaluation.

schwa-sounds that occur in these words. In both, the SD monophone and triphone model crossvalidation two complex words 'Empfänger' and 'öffnen' were not recognized correctly in most folds among other words.

As in the connected digits task the unadapted SI-model is not able to achieve satisfying results on dysarthric speech. A wide range of errors occurred in the speech data of the dysarthric speakers. Compared to the connected digits evaluation the results on the command word recognition tasks are generally lower. One reason for this is the increased vocabulary size. In addition the command words contain more complex phone combinations and multi-syllable words of various length, while in case of the connected digits task the pronunciation of the words is relatively easy and all digits contain only 3-5 phonemes per word.

The phone alignments of the recognizer outputs of the triphone acoustic models of two command words 'Prozent' [prɔt͡sɛnt] and 'grün' [grʏn] are shown in table 5.6. The words were selected because the utterances of all dysarthric speakers of the two words in the corresponding command words test sets were incorrectly recognized. A further analysis of the recognition results on the command words data set for both words proved that the word 'Prozent' is problematic for all speakers, as only one utterance of M002 and one of M068 were recognized correctly in the evaluation of the SI-model. Four of the nine utterances of the word 'grün' uttered by M002 were recognized correctly and only one of the four utterances of M067 was incorrectly recognized. For the other dysarthric speakers all utterances of the word were not recognized correctly. Both words start with combinations of plosives and trills which seem to cause problems. For example M063 and M068 do not utter the /r/ after /p/, which leads to a confusion with the word 'Pause' [pa͡ʊsə]. In case of M067 the recognition error in the utterance of 'grün' is caused by a hesitation at the beginning of the word. Also the plosive /t/ at the end of the word 'Prozent' is a potential source of error.

Like in the connected digit task the results of M063 are the worst in this evaluation. Also the SD-model performs significantly worse for M063 than for M002 although the amount of data used to train the acoustic model of M002 was smaller. One reason for the unsatisfying results are the truncations of the words made by the speaker that were already mentioned in chapter 2.

speaker	model	transcription
M002	SI-model	LAB: p r o: ts E n t REC: l a N z a: m
		LAB: g r y: n REC: d r aI
M002	SD-model	LAB: p r o: ts E n t REC: f E r g r ox: s ah n
		LAB: g r y: n REC: d r aI
M063	SI-model	LAB: p r o: ts E n t REC: p aU z eh
		LAB: g r y: n REC: l oe S eh n
M063	SD-model	LAB: p r o: ts E n t REC: l oe S eh n
		LAB: g r y: n REC: oe f n eh n
M066	SI-model	LAB: p r o: ts E n t REC: n U l
		LAB: g r y: n REC: z E n d eh n
M067	SI-model	LAB: p r o: ts E n t REC: l oe S eh n
		LAB: g r y: n REC: z E n d eh n
M068	SI-model	LAB: p r o: ts E n t REC: p aU z eh
		LAB: g r y: n REC: f Y n f

Table 5.6.: Some alignments of the recognizer output to the reference transcriptions done by HResults of utterances of the words ‘Prozent’ and ‘grün’ of five dysarthric speakers.

A closer analysis of the data also showed that the phone alignments assigned by the SI-model are mainly based on vowels and nasals that occur in the word. Substitutions by the recognizers of the latter are often a better match to the actual utterance of the speaker than the phone from the orthographic transcription of the word. Figure 5.8 shows the spectrogram and time aligned transcriptions of the utterance 'Prozent' by `M063` from the test set from which also the recognizer outputs shown in table 5.6 were generated. From the manual phone alignment one can see that the actual utterance is more or less a concatenation of several vowels. The SI-model aligns the word 'Pause' mainly because of the presence of the vowels /a͡ʊ/ and /ə/. The SD-model aligns the word 'löschen', but only the last part of the alignment is visible as the beginning of the word was aligned to a breathing sound that is not shown in the time window. In addition the low quality of parts of the training-data appeared to have a certain impact on recognition rate of the SD-model. In total it has to be doubted that any of the acoustic models trained for the command word task are suitable for practical use in case of this speaker.

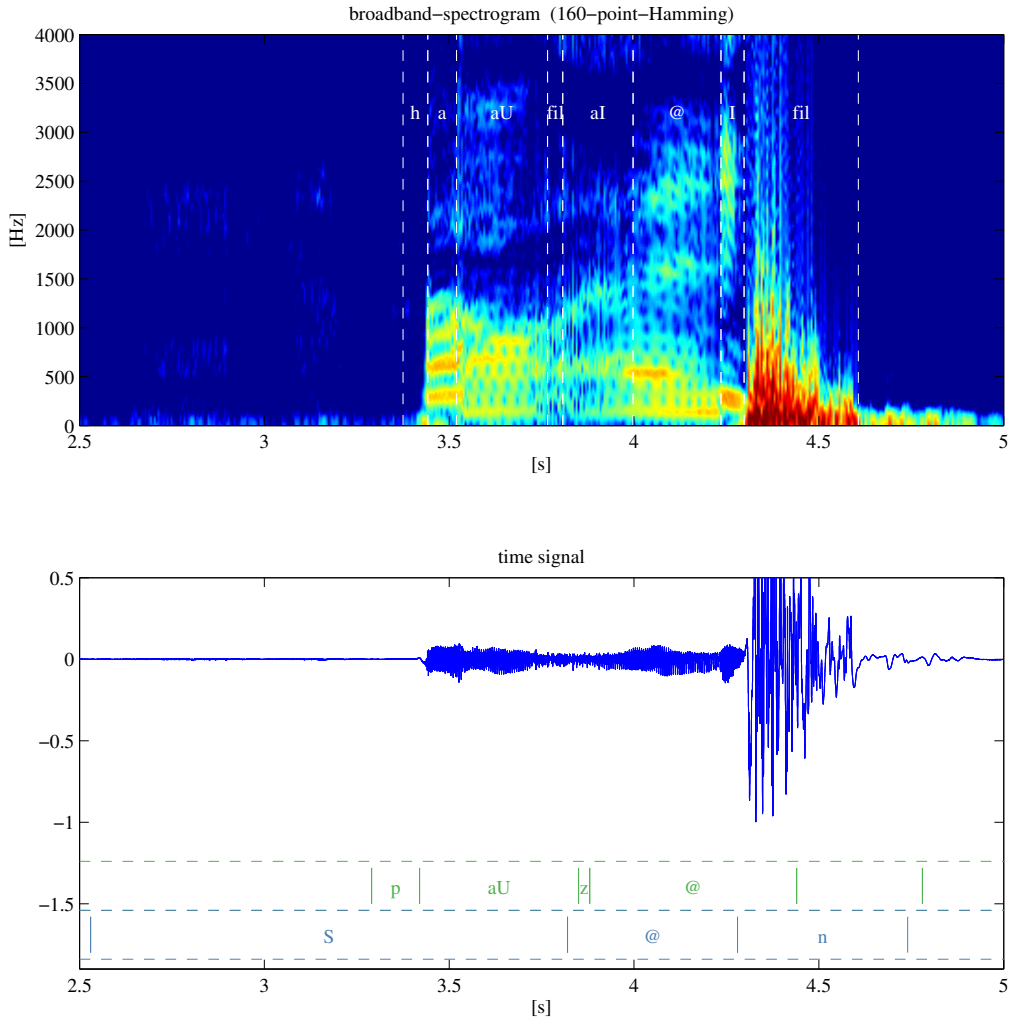

Figure 5.8.: The spectrogram of the word 'Prozent' uttered by `M063` and a comparison between the SAMPA-labels assigned by the SI (green), and SD (blue) triphone recognizers. The SAMPA-labels of the phones in the spectrogram were assigned manually.

5.3.3. Sotschek rhyme-test

The command word acoustic models described in the previous section were also evaluated on the Sotschek rhyme-test. Like before an isolated word grammar was used for this evaluation, but a separate grammar was generated for every ensemble, each containing only the 6 words from the analyzed rhyme. The lexicon contained all words from all ensembles.

The baseline results for the Sotschek rhyme-test are shown in table 5.7. Compared to the results on the command words r_s is significantly lower for the reference speaker `M000` for both the SI and SD-models. In contrast to the previous evaluation the results of the SI-model were also notably lower than for the SD-model. This shows that the test is suitable to point out the weaknesses in the acoustic models.

model	`M000`	`M002`	`M063`	`M066`	`M067`	`M068`
	r_s	r_s	r_s	r_s	r_s	r_s
SI monophone	74.00	47.00	21.00	39.00	52.00	51.00
SI triphone	65.00	46.00	27.00	31.00	53.00	47.00
SD monophone	82.00	64.00	26.00	-	-	-
SD triphone	80.00	72.00	38.00	-	-	-

Table 5.7.: The recognition rate r_s in [%] on the Sotschek rhyme-test for both the SI and SD-models on the five dysarthric speakers and the reference speaker. These results are used as baseline results for further evaluations.

For the dysarthric speakers the results achieved by the SD-models trained from speaker `M002` are lower on the rhyme-test evaluation than for the command word task where results of over 80% could be achieved. In contrast the recognition rates on the rhyme-test are higher in the SI-model evaluations of all dysarthric speakers, as well as in the evaluation of the SD-model trained from `M063`. These models achieved a very low recognition rate $r_s < 40\%$ on the command word task. The better results in this evaluation are very likely due to the smaller vocabulary used in the rhyme-test evaluation.

A detailed error analysis of the recognizer output of the SI triphone model was done in which the relative errors that occurred in samples with a different onset e_{on}, coda e_c and nucleus e_n of the syllable compared to the total number incorrect sentences S in the rhyme-test evaluation were analyzed.

$$e_{on} = \frac{\#errors\ in\ onset}{S} \tag{5.4}$$

$$e_c = \frac{\#errors\ in\ coda}{S} \tag{5.5}$$

$$e_n = \frac{\#errors\ in\ nucleus}{S} \tag{5.6}$$

This evaluation showed that for four out of five dysarthric speakers the relative error was the highest in rhymes with a different onset, followed by the relative error in ensembles with a different coda $e_{on} > e_c \geq e_n$. For speakers `M067` and `M068` the relative number of errors in the coda and nucleus was the same. `M063` was the only speaker were the distribution of the relative error was different during evaluation of the SI-model with $e_c > e_{on} > e_n$. The relative number of errors e_{on} was in the same range for the reference speaker and the dysarthric speakers. In the coda of the syllable `M000` even had the highest relative error rate e_c. But in contrast to the dysarthric speakers the relative number of errors in ensembles with a different nucleus e_n was

considerably lower for M000.

The same error analysis was done for the SD triphone models as well. In contrast to the SI-models the results varied much more from speaker to speaker. For the reference speaker M000 the error distribution was $e_n > e_{on} > e_c$ with $e_c \approx 10\%$. For M063 also $e_c \approx 25\%$ was the lowest relative error, while most errors were detected in rhymes with a different onset. In case of M002 the distribution of errors $e_c > e_n > e_{on} = 25\%$ was the complete opposite compared to M063.

5.3.4. Acoustic adaptation

The means of the emission probabilities of the SI acoustic models were adapted to match the feature vectors of the individual dysarthric speakers using MLLR. The adaptation set for each speaker contained one session of the command words (see section 3.2.1). The HMM-states of both, the monophone and triphone acoustic models were clustered before adaptation using a regression tree with 40 leaves. Details about the adaptation can be found in section 3.3.2.

Table 5.8 shows the recognition rate r_s on the command word task and the rhyme-test of the MLLR adapted acoustic models as well as the absolute improvement Δr_s compared to the baseline results presented in table 5.5 and table 5.7.

For the reference speaker the adaptation of the triphone models does not lead to a high improvement on the command word task, which is not surprising as recognition rates $r_s > 97\%$ were achieved in the baseline experiments. However, on the monophone model significant improvements $\Delta r_s > 7\%$ could be achieved. Looking at the errors that occurred, the word that was incorrectly recognized in most cases was the word 'simon' [sa͡imən], which was again confused with 'sieben' [siːbən].

For all dysarthric speakers acoustic adaptation of the SI-model lead to significant improvements for both types of subword modeling on the command word evaluations. However, the recognition rates of speakers M002 and M063 are still well below the recognition rate of the SD-models trained for these speakers. A comparison between the recognition rates of the adapted SI and the SD-models evaluated on the command words test set of both speakers is shown in figure 5.9.

An analysis of the recognition results of the adapted triphone models showed that the utterances of the word 'Prozent' in the test sets of M066, M067 and M068 are recognized correctly after acoustic adaptation. In case of M067 all utterances of the word in the command words data set are recognized correctly after adaptation. More than half of the utterances of M066 and M067 are also recognized correctly, while two thirds of the utterances of M002 and all utterances of M063 are still incorrectly recognized. For M002 and M063 the lowest improvement was measured on the command words data set. Possible reason for the confusions are in case of M002 the variation in the loudness of the utterances and in case of M063 the truncations which were discussed and illustrated in both chapter 2 and section 5.3.2

The utterance of 'grün' remains problematic after adaptation and is only recognized correctly in the evaluation of the test set in case of M067. However, when looking on the command words data set evaluation for M067 it turns out that one other utterance of the same word is incorrectly recognized after adaptation.

On the rhyme-test evaluation the improvement of the acoustic models was significant for both, the dysarthric speakers and the reference speaker, although compared to the command word evaluation the improvement on the rhyme-test was lower for the dysarthric speakers. The rhyme-test evaluation shows that acoustic adaptation leads to an improvement of the SI triphone model for the reference speaker M000 which cannot be measured in the command

Command words test set												
	M000		M002		M063		M066		M067		M068	
	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s
triphone	100.00	1.45	68.12	37.69	13.04	10.14	50.72	34.78	62.32	43.48	68.12	39.13
monophone	98.55	7.25	66.67	40.58	5.80	0.00	44.93	23.19	46.38	30.44	71.01	33.33
Command words data set												
	M000		M002		M063		M066		M067		M068	
	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s
triphone	99.42	1.74	66.67	36.40	19.20	9.90	49.64	40.22	71.38	42.76	69.57	41.31
monophone	97.97	7.54	67.15	34.14	16.94	7.83	42.75	25.00	59.42	36.59	69.93	33.70
Sotschek rhyme-test												
	M000		M002		M063		M066		M067		M068	
	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s	r_s	Δr_s
triphone	77.00	12.00	56.00	10.00	33.00	6.00	39.00	8.00	66.00	13.00	49.00	2.00
monophone	87.00	13.00	67.00	20.00	25.00	4.00	43.00	4.00	70.00	18.00	53.00	2.00

Table 5.8.: The recognition rate r_s in [%] after MLLR adaptation of the SI-models evaluated on three different data sets and the improvement of the results compared to the baseline results stated in the previous sections.

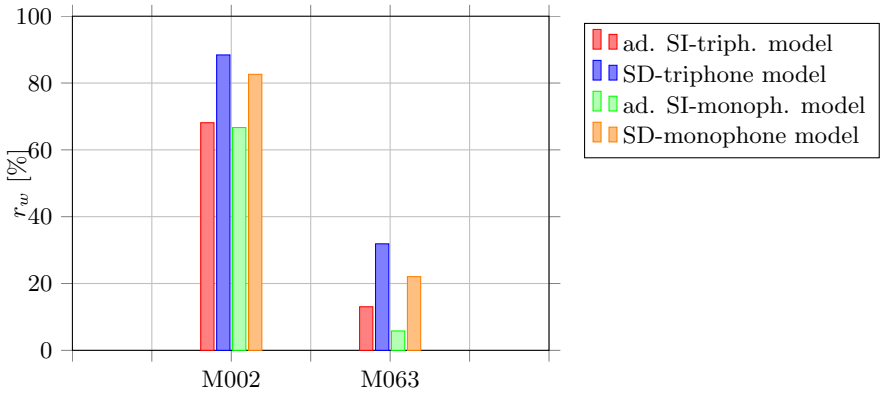

Figure 5.9.: A comparison of r_w between the MLLR adapted SI-models on the command words data set and the crossvalidation results of the SD-models the same data set.

words test.

The error analysis of the rhyme-test recognition results shows that acoustic adaptation of the SI triphone model influences the relative error distribution in different ways depending on the speaker. Only in case of M067 and M068 the total number of errors in rhymes with a different onset, nucleus and coda is reduced. For the reference speaker M000 the total number of sentence errors decreases, while the total of errors in ensembles with a different nucleus stays the same, which leads to a higher relative error e_n. For M002 and M063 the total number of errors decreases for samples with a different nucleus and coda, while more confusions occur with samples with a different onset. In case of M068 the opposite is the case. For the relative error the same ranking as for the unadapted SI-model $e_{on} > e_c \geq e_n$ applies to M000, M002, M063 and M066. For M067 and M068 the relative error e_n is highest and e_{on} the lowest. In both cases MLLR-adaptation leads to a very high gain in recognition performance on syllables with a varied onset, while the effect on the other rhymes is low in case of M067 and even negative in case of M068.

5.4. Lexical adaptation evaluation

For each of the five dysarthric speakers the pronunciation dictionaries used for both, the rhyme-test and command word task were adapted for a closer match between the transcriptions of the lexicon entries with the utterances of the speakers. Two adaptation approaches, that were proposed in section chapter 4 were used: phonological rules and FSTs. Adaptation was done for the SI and SD baseline acoustic models (see section 5.3.2) from the command word evaluation as well as the MLLR-adapted SI-models presented in section 5.3.4.

The phonological rules were derived from the rhyme-test evaluation of the acoustic models using the approach described in detail in section 4.3.2. The new pronunciation variants were added to the existing lexica and two types of pruning mechanisms were applied: hard and relative pruning. For speaker M002 an additional set of phonological rules was derived manually from the same rhyme-test evaluations and two lexica were generated based on these rules.

For lexical adaptation with FSTs the framework described in section 4.4 was used. New variants were generated based on the confusion matrix of the rhyme-test evaluation as representation of the acoustic model. The resulting lexica for each speaker contain the canonical transcriptions as well as the new variants that achieved both the desired confusability ranking and accuracy gain.

All newly generated lexica were stored in the same format as the base lexica. Evaluation of the SI, adapted SI and SD acoustic models was done on both, the rhyme-test and command word task in the same way as described in section 5.3 with the newly developed lexica as pronunciation model. Isolated word grammars were again used as language model. All results were evaluated on sentence level and compared to the baseline results from section 5.3.

5.4.1. Manually derived phonological rules

The general analysis of the speech data from the dysarthric speakers presented in chapter 2 showed that phone confusions could be found in the recordings of all speakers. Based on the rhyme-test-evaluation of the acoustic models presented in section 5.3.3 the speech data of speaker M002 was analyzed in more detail to derive phonological rules for the phone confusions that occur in the speech of M002. Based on the derived rule set speaker specific pronunciation variants were added to the existing rhyme-test and command words lexica for a closer match between the pronunciation model with the utterances of M002.

The following methods were combined to derive the phonological rule set:

- error analysis of the rhyme-test recognition results
- listening analysis of the speech samples
- evaluation of the spectrogram
- comparison of the findings to the reference speaker

In a fist step the recognition errors that occurred during the rhyme-test evaluation were analyzed regarding their reproducibility. Therefore potential pronunciation variants were manually added to the lexicon for words in ensembles with the same phonetic context. The acoustic models were then evaluated on the ensemble with the new lexicon to find out if the expected pronunciation variant appeared in the recognizer output. For example in the SD acoustic model a confusion between the vowel /a/ and the diphthong /ɔ͡ʏ/ occurred in ensemble 2 in the utterance of 'Ball' [bal]. This error could be reproduced in ensemble 1, in the utterance of 'bald' [balt] by adding variants with the vowel /a/ replaced with /ɔ͡ʏ/ for all six words of ensemble 1 to the lexicon. Using this approach more possible phone combinations could be evaluated in the ensembles than in the original setting and consequently also phone errors not represented in the original ensemble could be identified. For example for three ensembles (81, 82 and 83) with a varied onset the evaluation of the SI-model resulted in three different phone confusions for the plosive /t/. Further evaluation with the method previously described could identify the phone confusion of /t/ with /n/ as the most likely in all three ensembles.

By combining the listening analysis with further analysis of the spectrogram and aligned transcriptions using wavesurfer and Matlab^{TM} additional variations could be found in the speech data that were not obvious from the recognizer outputs. For example the insertion of the nasal /m/ before the plosive /p/, that was already mentioned in chapter 2, is clearly audible and also visible in the spectrogram, e.g. in the word 'paus' uttered by `M002` which is shown in figure 5.10. This insertion could also be observed in some of `M002`'s utterances of command words, e.g. in the word 'Prozent', which was also identified as being one of the words most often incorrectly recognized in the command word task (see section 5.3.2). In ensemble 56 of the rhyme-test, in which the word 'paus' occurs, possible alignments for the onset of the syllable are six consonants (/p/, /l/, /z/, /r/, /h/ and /g/) due to the closed setting of the rhyme-test. The phone alignments of the acoustic models are therefore different from the actual utterance of `M002` and are shown in figure 5.10. The SI and adapted SI-model both do not assign a label to the nasal before /p/, while the SD-model assigns the label /g/ to the first part of the utterance.

The better the models were adapted to `M002`'s speech the fewer rules could be clearly identified to be reproduceable, partly also because no comparable tests were available in the rhyme-test data. In addition it has to be considered, that not all confusions in the recognizer output are directly connected to a phone confusion made by the speaker. It has already been shown in chapter 2 that speaker `M002` tends to have certain variations in the loudness and the intensity of his voice, which lead to recognition errors as well. The recording for ensemble 16 of `M002` e.g. was completely unusable as the onset of the utterance is overdriven.

An additional source of error was the insertion of pauses between phones by `M002` which is typical for dysarthric speech [18], but not present in that form in unimpaired speech. Most non-linguistically motivated pauses can not be detected in the rhyme-test-evaluation as the pronunciation model only contains isolated words. In most cases these pauses lead to phone confusions and deletions due to incorrect segmentations of the speech signals. In the example utterance shown in figure 5.11 the SI and adapted SI-model do not segment the utterance of 'bunt' [bʊnt] correctly as the /t/ is uttered after a long occlusion phase. The output of the

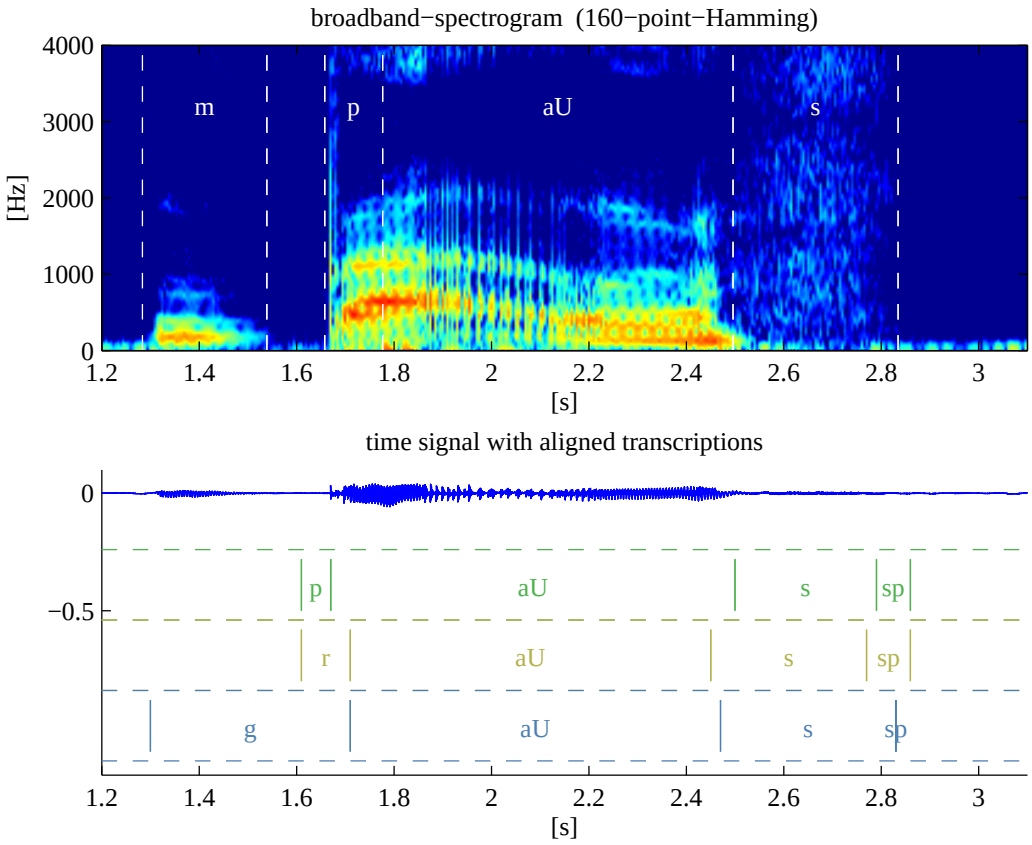

Figure 5.10.: The spectrogram of the word 'paus' uttered by M002 and a comparison between the SAMPA-labels assigned by the SI (green), adapted SI (yellow) and SD (blue) triphone recognizers. The SAMPA-labels of the phones in the spectrogram were assigned manually.

recognizers, however, indicate a deletion of /t/ and a substitution of /n/ with /m/, although /n/ is correctly pronounced by M002. Another example for a wrong segmentation was the utterance of 'flink' [flŋk] where M002 inserted a pause of about 0.4 s between /f/ and /l/. As the ensemble is only varied in the coda of the syllable the phone label /f/ is aligned at an incorrect position by all acoustic models and no phonetic variation is detected automatically in this case. Segmentation errors occurred more often in the phone alignments of the SI-model than in the adapted SI and SD-model. This shows that characteristics of the speech of M002, like the long occlusion phase of a plosive, are learned to a certain extend not only by the SD-model, but also by acoustic adaptation of the HMM-model parameters. Consequently, the number of rules found during the evaluation of the speech data depended strongly on how well the models were adapted to the dysarthric speaker. For the SI-model 20 rules could be identified, while only 9 could be found for the adapted SI-model and just 6 for the SD-model.

The rule sets were applied to the rhyme-test and command words lexica. The resulting pronunciation models were evaluated on the same triphone models as used in the command word task. Table 5.9 shows the number of variants generated by the defined rule sets for the different acoustic models, as well as r_s for the different test sets. For the SI-model the improvement on the rhyme-test was 6% which is around 2/3 of the improvement achieved by acoustic adaptation of the same model. The results on the acoustically adapted SI-model

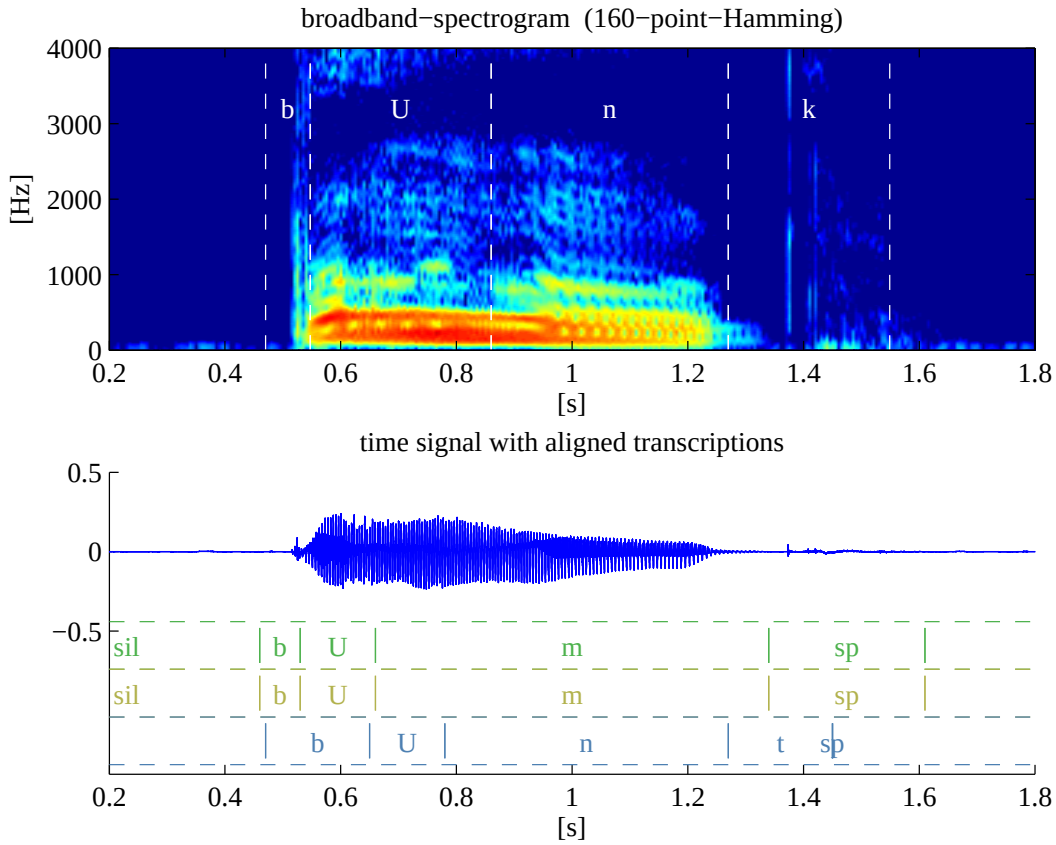

Figure 5.11.: The spectrogram of the word 'bunt' uttered by `M002` and a comparison between the SAMPA-labels assigned by the SI (green), adapted SI (yellow) and SD (blue) triphone recognizers. The SAMPA-labels of the phones in the spectrogram were assigned manually.

could be further improved by 4%. In contrast on the command word recognition task an improvement of almost 3% could only be measured for the full command words set in case of the SI-model. Compared to the improvement achieved with acoustic adaptation this is a poor result. For the adapted SI-model r_s was even lower than the baseline results of both command word evaluations. Lexical adaptation of the SD-model lead to an improvement of 5% on the rhyme-test data, but on the command word task only a small improvement $< 2\%$ could be measured on the test set.

An analysis of the errors that occurred before and after lexical adaptation showed that the adaptation of the pronunciation model changes the recognizer output on the command word evaluation more than suggested by the recognition results. For example on the command words test set evaluated on the SI-model eight utterances were recognized correctly after lexical adaptation while eight other utterances were incorrectly recognized, due to confusions with newly generated resulting in no detectable changes in the overall result. This shows the importance of a pruning strategy to control the confusability of the newly generated lexica.

5.4.2. Automatically generated phonological-rules

Phonological rules to model the pronunciation of the dysarthric speakers where generated automatically based on the rhyme-test evaluation presented in section 5.3.3. The `HRESULTS`

M002	rhyme-test			Command words				
				test set			full set	
	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
SI-model	1788	52.00	6.00	367	30.43	0.00	33.17	2.90
adapted SI-model	220	60.00	4.00	17	65.22	-2.90	64.73	-1.94
SD-model	133	72.00	5.00	16	89.86	1.45	92.11	0.00

Table 5.9.: The number of variants generated with lexical adaptation based on manually derived rewrite rules for M002 for the SI, adapted SI and SD-model along with the achieved recognition rates r_s in [%] evaluated on both the rhyme-test and command words as well as the absolute improvements compared the baseline results presented in section 5.3.

# vars.	M002	M063	M066	M067	M068
SI-model	55	71	70	50	55
adapted SI-model	46	69	60	35	53
SD-model	31	73	-	-	-

Table 5.10.: The number of automatically generated rules for each of the dysarthric speakers for the SI, adapted SI and SD-models based on the Sotschek rhyme-test recognition results.

alignments of the phonemes of the recognizer outputs to the reference transcriptions were used to extract the rule sets. The probability for each rule and phone statistics of the rhyme-test were computed as well to assign each pronunciation variant a probability, which could then be used for pruning. Details about the algorithms used for generation of the rules and the extension of the lexica can be found in section 4.3.

The phonological rules were applied to two different lexica containing the rhyme-test words and the command words respectively. The lexica were adapted to each of the five dysarthric speakers for the triphone acoustic models that were previously trained for the command word evaluation in section 5.3. The number of rules generated for each speaker and model can be found in table 5.10. The evaluation of the new pronunciation dictionaries was done in the same way as for the manually derived phonological rules on the rhyme-test and command word recognition tasks using isolated word grammars as language model.

In table 5.11 the number of generated pronunciation variants before and after pruning and the recognition rate r_s for the different acoustic models are presented for the different tasks.

For all speakers lexical adaptation lead to improved results on the rhyme-test evaluation. The improvement varied strongly between 4% and 21% depending on the speaker and model. This variation is caused by the different sizes of the rule sets generated, which depend on the number of errors on the baseline results. For models with a high number of recognition errors more confusions were detected from which more rules are generated. Applying more rules to the development set leads to a higher improvement after adaptation. This can also be observed in the mean and standard deviation of the recognition results compared to the baseline results. Besides the average r_s being higher after lexical adaptation also the standard deviation of the results is lower. Comparing the rhyme-test lexical adaptation results without pruning to the baseline results the standard deviation of the average recognition rate dropped from 12.7 to 7.5 for the SI-model and from 13.2 to 9.7 in case of the adapted SI-model results.

A comparison of the recognition errors that occurred before and after lexical adaptation shows that in the rhyme-test evaluation many errors from the baseline results are corrected, but also new confusions occur. To remove unlikely variants from the lexicon and reduce the

			rhyme-test			Command words				
SI-model							test set		full set	
speaker	pruning	factor	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
M002	no pruning		1196	60.00	14.00	65	27.54	-2.89	27.21	-3.06
M002	hard pruning	0.01	891	59.00	13.00	56	27.54	-2.90	28.02	-2.25
M002	relative pruning	0.001	562	66.00	20.00	25	27.54	-2.89	27.21	-3.06
M063	no pruning		1630	47.00	20.00	89	5.80	2.90	10.09	0.79
M063	hard pruning	0.007	1290	51.00	24.00	79	5.80	2.90	10.19	0.89
M063	relative pruning	0.01	1299	52.00	25.00	77	5.80	2.90	10.19	0.89
M066	no pruning		1904	50.00	19.00	101	18.84	2.90	11.96	2.54
M066	hard pruning	0.001	1814	51.00	20.00	101	18.84	2.90	11.96	2.54
M066	relative pruning	0.07	596	52.00	21.00	57	17.39	1.45	10.87	1.45
M067	no pruning		1140	65.00	12.00	66	15.94	-2.90	26.81	-1.81
M067	hard pruning	0.001	1105	65.00	12.00	66	15.94	-2.90	26.81	-1.81
M067	relative pruning	0.003	1048	65.00	12.00	62	15.94	-2.90	26.81	-1.81
M068	no pruning		1131	59.00	12.00	61	27.54	-1.45	27.54	-0.72
M068	hard pruning	0.003	1061	60.00	13.00	58	27.54	-1.45	27.54	-0.72
M068	relative pruning	0.005	1029	61.00	14.00	57	27.54	-1.45	27.54	-0.72
adapted SI-model										
speaker	pruning	factor	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
M002	no pruning		907	67.00	11.00	41	66.67	-1.45	63.93	-2.74
M002	hard pruning	0.003	860	66.00	10.00	41	66.67	-1.45	63.93	-2.74
M002	relative pruning	0.005	807	66.00	10.00	40	66.67	-1.45	63.93	-2.74
M063	no pruning		1630	49.00	16.00	82	15.94	2.90	20.18	0.98
M063	hard pruning	0.007	1230	51.00	18.00	71	15.94	2.90	20.18	0.98
M063	relative pruning	0.01	1221	52.00	18.00	71	15.94	2.90	20.18	0.98
M066	no pruning		1252	56.00	17.00	66	56.52	5.30	52.90	3.26
M066	hard pruning	0.001	1221	56.00	17.00	66	56.52	5.80	52.90	3.26
M066	relative pruning	0.09	428	56.00	17.00	35	53.62	2.90	52.54	2.90
M067	no pruning		445	74.00	8.00	30	62.32	0.00	70.65	-0.73
M067	hard pruning	0.003	426	74.00	8.00	30	62.32	0.00	70.65	-0.73
M067	relative pruning	0.003	428	74.00	8.00	30	62.32	0.00	70.65	-0.73
M068	no pruning		1056	63.00	14.00	45	71.01	2.89	69.57	0.00
M068	hard pruning	0.001	981	63.00	14.00	45	71.01	2.89	69.57	0.00
M068	relative pruning	0.15	153	64.00	15.00	14	69.57	1.45	68.48	-1.09
SD-model										
speaker	pruning	factor	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
M002	no pruning		563	80.00	4.00	25	89.86	1.45	91.95	-0.16
M002	hard pruning	0.05	273	78.00	6.00	20	89.86	1.45	91.95	-0.16
M002	relative pruning	0.08	221	78.00	6.00	14	89.86	1.45	91.79	-0.32
M063	no pruning		1332	44.00	6.00	64	31.88	0.00	-	-
M063	hard pruning	0.08	427	46.00	8.00	32	31.88	0.00	-	-
M063	relative pruning	0.09	483	48.00	10.00	37	31.88	0.00	-	-

Table 5.11.: The number of variants generated for the rewrite rules with and without pruning as well as the achieved recognition rates r_s in [%] on both the rhyme-test and command words evaluations along with the absolute improvement Δr_s compared to the baseline results presented in section 5.3.

confusability two types of pruning were applied to the pronunciation variants based on their probability: hard pruning and relative pruning. Both methods are described in detail in section 4.3.4. The selection of a proper pruning factor was done on the rhyme-test for all models and both pruning variants. The pruning factor was varied between 0.001 and 0.2 and the value that achieved the best result on the rhyme-test evaluation of the individual speaker was selected. If two factors achieved the same recognition rate the higher was selected for both pruning methods, favoring lexica with more pruned samples. The selected pruning variant and factor as well as the size of the lexica along with the achieved results on both, the rhyme-test and command word task are shown in table 5.11 for all speakers. One can see that for the majority of the speakers the results on the rhyme-test improved or stayed equal after pruning was applied. Again after pruning some errors are eliminated in the rhyme-test evaluation by the changes in the lexicon, while some new errors occurred.

Looking at the command words results the rules found do not seem to generalize well for most speakers and acoustic models. Only the results of the SI and adapted SI-model evaluations on the command words data set for speakers `M063` and `M066` show a small improvement. The SD-model trained for `M002` showed a marginal improvement on the command words test set while the average cross-validation result was worse than the corresponding baseline result. An analysis of the errors that occurred in the baseline and lexical adaptation results showed that like in the evaluation of the manually generated rules lexical adaptation does lead to an improvement and incorrect baseline results are recognized correctly after adaptation. But the result is again deteriorated, as the confusability of the new lexicon was high. In total about the same number of errors occurred because of confusions with newly generated variants.

Unfortunately, also the chosen pruning methods did not have a strong influence on the recognition results on the command word evaluation. In most cases no change in the recognizer output was detected after pruning although the number of pronunciation variants was lower in the pruned lexicon. For example an analysis of the errors that occurred in the cross validation evaluation of the SD-model of `M002` showed that the recognition results after pruning only change in one fold. Beside the fact the rhyme-test evaluations were the development set for which the rules are optimized, another reason for the low results on the command words might be the different number of syllables of the words in both sets. As the rhyme-test contains only one-syllable words rules can only be derived automatically for three types of phone settings:

- consonant-vowel-consonant
- beginning of word - consonant - vowel
- vowel - consonant - end of word

Many of the command words consist of two or more syllables and mispronunciations or pauses at the intersections are not present in the rhyme-test and cannot be reflected in the rule set. In addition it could be shown in the manual evaluation of the phone alignments of speaker `M002` (see section 5.4.1) that not all utterances are segmented correctly during recognition, which in case of automatic rule generation results in rules that model substitutions that are not entirely correct.

A comparison between the number of pronunciation variants generated for speaker `M002` based on the manual analysis and the automatic approach showed that for the SI-model in the automatic approach a considerably lower number of variants was generated for the command words lexicon. This is a result of the limited phone contexts possible in the rules. On the other hand the number of generated variants is higher for the adapted SI and SD-model, as for these models only very few rules could be developed during manual rule generation as confusions could not be reproduced on other utterances of the speaker. In contrast this limitation did

not apply to the automatic rule generation, but might lead to rules being generated based on outlier utterances.

5.4.3. Finite State Transducers

The framework described in section 4.4 was used to adapt the pronunciation dictionaries of the rhyme-test and command words to dysarthric speech. For the five speakers pronunciation variants were generated based on the orthographic transcriptions in the lexica as language model FST and an acoustic model WFST representing the phone errors of the confusion matrix of the rhyme-test evaluation of the corresponding acoustic model. Evaluations of the lexica were done in the same way as for the phonological rules on the command word and rhyme-test recognition tasks.

The number of pronunciations added to the lexica and the achieved recognition results are shown in table 5.12. Compared to the previous adaptation approach with phonological rules the number of pronunciations added to the rhyme-test lexicon is significantly lower in the FST-approach. This is caused by the general differences in both approaches. In the FST-approach only one new pronunciation variant is generated per word. In addition in case of rhyme-test words many generated variants are discarded because of the high confusability with other words. The rhyme-test lexicon is per definition confusable, as every word differs from at least five others only in one phone. In contrast the number of phonological rules created and indirectly also the number of resulting variants is directly related to the number of recognition errors that occurred on the rhyme-test. Therefore, the number of rules also decreased when an acoustic model achieved better results on the rhyme-test, which resulted also in a lower number of generated variants.

Another difference is that the FST-approach does not take into account the phonetic context of a phone confusion, which overcomes the limitations of variant generation for the command words described for the phonological rule approach in the previous section. Consequently one pronunciation variant was added for almost all words in the command words lexicon for all models in the FST-approach, as the orthographic transcriptions and generated variants of most words are not likely to be confused. However, this does not change the fact that only information about the three types of confusions described in the previous section are modeled in the confusion matrix. To reduce the number of variants generated in the FST-approach the FST framework allows the setting of the confusability score and the maximum difference of the similarity measures between the canonical transcription and the generated pronunciation variant for the dysarthric speaker and the reference speaker.

During the evaluation it turned out that the high improvements of the recognition rate on the rhyme-test that were achieved with the phonological rules could not be measured in the evaluation of the WFSTs. In most cases the evaluations of the new lexica showed that more errors occurred than in the baseline results. However, it has to be considered that in the previous approach the rules were built to match exactly the confusions that occurred in the rhyme-test, in contrast to the WFST where the context of the confusion is not taken into account. An analysis of the error outputs showed that in this evaluation like in the command word evaluation of the phonological rules a high number of new confusions is the main reason for the reduced recognition rate. This result was not expected as the confusability score was designed to vote against variants that are easily confused with other words in the lexicon.

Although the rhyme-test evaluation did not lead to improvements on the evaluation of the SI-model of the dysarthric speakers, the evaluation on the command words data set showed marginal improvements for the speakers `M066`, `M067` and `M068`. Compared to the improvement

	rhyme-test			Command words				
SI-model					test set		full set	
speaker	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
M002	259	39.00	-7.00	59	27.54	-2.89	28.99	-1.28
M063	197	24.00	-3.00	61	2.90	0.00	9.21	-0.09
M066	230	26.00	-5.00	61	15.94	0.00	10.87	1.45
M067	317	44.00	-9.00	65	20.29	1.45	29.71	1.09
M068	273	51.00	4.00	61	33.33	4.34	31.16	2.90
adapted SI-model					test set		full set	
speaker	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
M002	285	53.00	-3.00	60	66.67	-1.45	66.67	0.00
M063	165	33.00	0.00	60	13.04	0.00	20.47	1.27
M066	177	38.00	-1.00	58	50.72	0.00	49.28	-0.39
M067	402	61.00	-5.00	67	62.32	0.00	70.29	-1.09
M068	273	51.00	2.00	63	69.57	1.45	69.57	0.00
SD-model					test set		full set	
speaker	# vars.	r_s	Δr_s	# vars.	r_s	Δr_s	r_s	Δr_s
M002	347	72.00	0.00	61	92.75	4.34	91.95	-0.16
M063	428	38.00	0.00	63	31.88	4.34		

Table 5.12.: The number of variants generated using WFSTs as well as the achieved recognition rates r_s in [%] on both the rhyme-test and command words evaluations along with the absolute improvement Δr_s compared to the baseline results presented in section 5.3.

resulting from acoustic adaptation these results are negligible. Again the reason for the small improvement was the number of confusions that occurred after lexical adaptation. The SD-models trained for M002 and M063 both achieved an improvement of 4.34% on the test set. However, this result could not be approved in the cross-validation on the data of M002.

The main reason for the high confusability after lexical adaptation with the FST-approach is that the entries of the confusion matrix from the rhyme-test evaluation are sparse. As the confusion matrix is generated from only 100 test samples the differences between the entries are too small to model a proper representation of the acoustic model in the corresponding WFST. This is a general problem of the rhyme-test evaluation as used in this work. As described in section 4.4 the confusion matrix contained a lot of entries that were zero, some even in the main diagonal. Therefore, the confusion matrix had to be modified to be used for the task. The introduction of a smoothing strategy for the confusion matrix as described in [20] could lead to better results. It has been shown in section 5.3.3 that there is a correlation between the number of errors that occurred in ensembles with a different onset, nucleus and coda for the reference speaker and the dysarthric speakers. A possible smoothing strategy for the SI-model could be based on merging the confusion matrix of a dysarthric speaker with a normalized matrix generated from several unimpaired speakers. In this work the phonetic distance of the new variant to the canonical transcription was evaluated against the phonetic distance of the variants for the reference speaker M000, based on the corresponding confusion matrix of M000's rhyme-test evaluation which can also be considered as sparse. The normalized matrix from several unimpaired speakers would also provide a more accurate measure for the pronunciation accuracy. For the SD-model a possible strategy to increase the amount of data could also be a leave-out training strategy as used in crossvalidation with evaluation passes on the rhyme-test to generate more data for the confusion matrix.

6. Conclusion

6.1. Summary

The evaluations in this work have shown that great improvements could be achieved in the recognition performance of dysarthric speech on SI-models using acoustic adaptation. The major advantage of adaptation methods like MLLR is that only very small sets of data of the dysarthric speakers are needed. The recognition rate achieved on a small vocabulary connected digits task exceeded 90% for two of the five dysarthric speakers evaluated. For one speaker suffering from severe dysarthria the MLLR adapted SI-model even outperformed the SD-model trained for this task. However, it could also be shown that acoustic adaptation cannot overcome the differences of dysarthric and unimpaired speech for all speakers. In a command word task using a 69 word vocabulary only a maximum recognition rate of 70% could be achieved with acoustic adaptation. It also turned out that for one speaker the vocabulary from this task was too demanding and neither an SD nor an acoustically adapted SI-model could achieve satisfying results.

For all speakers mispronunciations due to their speech impairment were detected in the recorded data and two data-driven approaches to model this information in the pronunciation dictionary were evaluated: phonological rules and FSTs. It turned out that both approaches lead to improvements as samples are recognized correctly after adaptation but also a lot of new confusions occur that conceal the improvements or even lead to a negative result compared to the baseline recognition system with the pronunciation dictionary containing only the orthographic transcriptions of the words. Several pruning strategies were applied, but the expected improvement of the results could not be achieved. The most likely reason for this is insufficient data, as both approaches derive the confusion information from a rhyme-test evaluation containing only 100 single observations. In addition there is evidence that the confusions found in the one-syllable rhyme-test words do not generalize well to the multi-syllable command words.

6.2. Outlook

The performance on the command word task could be improved significantly with MLLR acoustic adaptation. However, the improvements are limited even if the whole training set is used for adaptation. A further improvement of the results could be achieved by combining MLLR and MAP-adaptation using the data from the training set of the SD-models, for which the data of the three speakers `M066`, `M067` and `M068` also would be sufficient.

For the lexical adaptation it would be important to gain more information about the phone confusions. This could either be done by recording more samples from the dysarthric speakers, or by introducing a three pass rhyme-test where for each recorded word 3 ensembles are designed in which the onset, nucleus and coda of the syllable is varied respectively. This way two times more information could be extracted using the same number of recordings. In case of the SD-model also a leave-out strategy like for crossvalidation could be used to generate more data from the rhyme-test evaluation. Evaluations have also shown that the number of words in one

ensemble might be too small to cover all possible phone confusions.

Further improvements for the lexical adaptation with FSTs could also be achieved, like proposed in [20], by smoothing of confusion matrix of rhyme-test. This could be done with the average confusion matrix of different unimpaired speakers.

Another idea would be to use a different rhyme-test setting for lexical adaptation, e.g. the rhyme-test developed by Ziegler et. al., which 'uses polysyllabics in addition to monosyllabics and dispensed with infrequent words and inflected forms' [40] to avoid pronunciation problems resulting from the unfamiliarity of dysarthric speakers with the word. This rhyme-test might also be a better match to the multi-syllable command words.

Bibliography

[1] Baum, M. and Erbach, G. (2000). SpeechDat(AT) Austrian Database for the fixed telephone network.

[2] Brookes, M. (2009). VOICEBOX: Speech Processing Toolbox for MATLAB, Version 1.6. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[3] Fosler-Lussier, E. (2003). A tutorial on pronunciation modeling for large vocabulary speech recognition. In Renals, S. and Grefenstette, G., editors, *Text- and Speech-Triggered Information Access*, volume 2705 of *Lecture Notes in Computer Science*, pages 38–77. Springer-Verlag, Berlin / Heidelberg, DE.

[4] Fosler-Lussier, E., Amdal, I., and Kuo, H.-K. J. (2005). A framework for predicting speech recognition errors. *Speech Communication*, 46:153–170.

[5] Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M. S., and Parker, M. (2003). Automatic speech recognition with sparse training data for dysarthric speakers. In *Proceedings 8th European Conference on Speech Communication and Technology*, Geneva, CH.

[6] Hasegawa-Johnson, M. (2009). Speech tools minicourse 2009, lecture 6. `http://www.isle.illinois.edu/sst/courses/minicourses/2009`.

[7] Hawley, M. S., Enderby, P., Green, P., Brownsell, S., Hatzis, A., Parker, M., Carmichael, J., .Cunningham, S., O'Neill, P., and Palmer, R. (2003). STARDUST - Speech Training and Recognition for Dysarthric Users of Assistive Technology. `http://www.fastuk.org/research/projview.php?id=216`.

[8] Hawley, M. S., Green, P., Enderby, P., Cunningham, S., and Moore, R. K. (2005). Speech technology for e-inclusion of people with physical disabilities and disordered speech. In *Proceedings 9th European Conference on Speech Communication and Technology*, pages 445–448, Lisbon, PT.

[9] Hosom, J.-P., Jakobs, T., Baker, A., and Fager, S. (2010). Automatic speech recognition for assistive writing in speech supplemented word prediction. In *Proceedings 11th Annual Conference of the International Speech Communication Association*, pages 2674–2677, Makuhari, Chiba, JP.

[10] Jurafsky, D. and Martin, J. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall.

[11] Kain, A. B., Hosom, J.-P., Niu, X., van Santen, J. P. H., Fried-Oken, M., and Staehely, J. (2007). Improving the intelligibility of dysarthric speech. *Speech Communication*, 49:743–759.

[12] Kain, E., Niu, X., Hosom, J.-P., Miao, Q., and Santen, J. V. (2004). Formant resynthesis of dysarthric speech. In *Proceedings 5th ISCA Workshop on Speech Synthesis*, pages 25–30, Pittsburgh PA.

[13] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008). Dysarthric Speech Database for Universal Access Research. In *Proceedings 9th Annual Conference of the International Speech Communication Association*, pages 22–26, Brisbane, AU.

[14] Leggetter, C. and Woodland, P. (1994). Speaker adaptation of continuous density HMMs using multivariate linear regression. In *Proceedings 3rd International Conference on Spoken Language Processing*, pages 451–454, Yokohama, JP.

[15] Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185.

[16] Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgank, A., C, Z. K. C., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recogniser based on Speechdat(II). In *Proceedings 6th International Conference on Spoken Language Processing*, pages 370–373, Beijing, CN.

[17] Machelett, K. (1996). Das lesen von sonagrammen v1.0 - begleitendes hypertext-dokument zur vorlesung. `http://www.phonetik.uni-muenchen.de/studium/skripten/SGL/SGLHome.html`.

[18] Magnuson, T. and Blomberg, M. (2000). Acoustic analysis of dysarthric speech and some implifications for automatic speech recognition. *THM-Quaterly Progress and Status Report*, 41(1):19–30.

[19] Menendez-Pidal, X., Polikoff, J., Peters, S., Leonzio, J., and Bunnell, H. (1996). The Nemours database of dysarthric speech. In *Proceedings 4th International Conference on Spoken Language Processing*, pages 1962–1965, Philadelphia, PA.

[20] Morales, S. O. C. and Cox, S. J. (2009). Modelling errors in automatic speech recognition for dysarthric speakers. *Journal on Advances in Signal Processing*, 2009:2:1–2:14.

[21] Muhr, R. (2008). The Pronouncing Dictionary of Austrian German (AGPD) and the Austrian Phonetic Database (ADABA): Report on a large Phonetic Resources Database of the three Major Varieties of German. In *Proceedings of the 6th International Language Resources and Evaluation Conference*, Marrakech, MA. European Language Resources Association (ELRA).

[22] Oscar Saz, William RodrÃguez, E. L. and Vaquero, C. (2008). A novel corpus of children's disordered speech. In *The 1st Workshop on Child, Computer, and Interaction*.

[23] Pearce, D. and Hirsch, H.-G. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings International Speech Communication Association Tutorial and Research Workshop Automatic Speech Recognition*, pages 29–32, Paris, FR.

[24] Petrik, S. (2010). *Phonetic Similarity Matching of Non-Literal Transcripts in Automatic Speech Recognition*. PhD thesis, Graz University of Technology, Graz, Austria.

[25] Rosengren, E. (2000). Perceptual analysis of dysarthric speech in the ENABL project. *THM-Quaterly Progress and Status Report*, 41(1):13–18.

[26] Sanders, E., Ruiter, M., Beijer, L., and Strik, H. (2002). Automatic recognition of dutch dysarthric speech: a pilot study. In *Proceedings 7th International Conference on Spoken Language Processing*, pages 661–664, Denver, CO.

[27] Saz, O., Lleida, E., and Miguel, A. (2009a). Combination of acoustic and lexical speaker adaptation for disordered speech recognition. In *Proceedings 10th Annual Conference of the International Speech Communication Association*, pages 544–547, Brighton, UK.

[28] Saz, O., Miguel, A., Lleida, E., Ortega, A., and Buera, L. (2006). Study of time and frequency variability in pathological speech and error reduction methods for automatic speech recognition. In *Proceedings 9th International Conference on Spoken Language Processing*, Pittsburgh, PA.

[29] Saz, O., Yin, S.-C., Lleida, E., Rose, R., Vaquero, C., and Rodríguez, W. R. (2009b). Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, 51:948–967.

[30] Sharma, H. V. and Hasegawa-Johnson, M. (2009). Universal access: Speech recognition for talkers with spastic dysarthria. In *Proceedings 10th Annual Conference of the International Speech Communication Association*, pages 1451–1454, Brighton, UK.

[31] simon listens - non profit organization for research and apprenticeship (founded 2007). simon SpeechInterfacedaeMON- Spracherkennungssoftware für körperlich behinderte Menschen. `http://www.simon-listens.org/`.

[32] Sjölander, K. and Beskow, J. (2005). Wavesurfer - An open source speech tool. `http://www.speech.kth.se/wavesurfer/`.

[33] Sotschek, J. (1982). Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte. *Der Fernmeldeingenieur*, 36:345–353.

[34] Statistical Speech Technology Group, University of Illinois at Urbana-Champaign (2009). UA-Speech Database. `http://www.isle.illinois.edu/sst/data/UASpeech/`.

[35] Strik, H. (2001). Pronunciation adaptation at the lexical level. In *Proceedings International Speech Communication Association Tutorial and Research Workshop Adaptation Methods for Speech Recognition*, pages 123–130, Sophia Antipolis, FR.

[36] Susan K. Fager, David R. Beukelman, T. J. and Hosom, J.-P. (2010). Evaluation of a speech recognition prototype for speakers with moderate and severe dysarthria: A preliminary report. *Augmentative and Alternate Communications*, 26(4):267–277.

[37] Wells, J. C. (1997). SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R., and Winski, R., editors, *Handbook of Standards and Resources for Spoken Language Systems*, chapter Part IV, Section B. Mouton de Gruyter, Berlin, DE and New York, NY. `http://www.phon.ucl.ac.uk/home/sampa`.

[38] Woodland, P. (2001). Speaker Adaptation for Continuous Density HMMs: A Review. In *Proceedings International Speech Communication Association Tutorial and Research Workshop Adaptation Methods for Speech Recognition*, pages 11–19, Sophia Antipolis, FR.

[39] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book Version 3.4*. Cambridge University Press, Cambridge, UK.

[40] Ziegler, W., Hartmann, E., and von Cramon, D. (1988). Word identification testing in the diagnostic evaluation of dysarthric speech. *Clinical Linguistics & Phonetics*, 2(4):291–308.

A. Acronyms and Symbols

List of Acronyms

ASR	automatic speech recognition
FFT	fast Fourier transform
FST	finite state transducer
GMM	Gaussian mixture model
HMM	hidden Markov model
HTK	Hidden Markov Model Toolkit
IFFT	inverse fast Fourier transform
IPA	international phonetic alphabet
MAP	maximum a posteriori
MFCC	Mel-frequency cepstral coefficients
ML	maximum likelihood
MLLR	maximum likelihood linear regression
NN	neural network
PCM	pulse code modulation
PDF	probability density function
POTS	plain old telephone service
SAMPA	Speech Assessment Methods Phonetic Alphabet
SD	speaker-dependent
SER	sentence error rate
SI	speaker-independent
WER	word error rate
WFST	weighted finite state transducer

List of Symbols

#	number of operator
/.../	marker for the symbol of a single phone
$\gamma(t)$	probability of the Gaussian at time t
$\circ$	FST composition operator
T^{-1}	inverse of an FST T
$P(A)$	probability of A
$P(B\|A)$	conditional probability of B given A
μ	mean value of a Gaussian PDF

B. Lexicon

B.1. Phone label set

The label set of the acoustic models is based on the lexicon of the Speechdat(II)-ATdatabase. Transcriptions are in SAMPA-coding using the German phone set published in [37] without the rare phonemic affricate /d͡ʒ/, which was mapped to /d ʃ/ instead. In addition the phone /ɾ/ is transcribed as as lowercase /r/ instead of the uppercase /R/ used in [37]. For model training the SAMPA symbols were converted to HTK-SAMPA. The complete phone set can be found in table B.1 and table B.2.

IPA	SAMPA	HTKSAMPA
Plosives		
p	p	p
b	b	b
t	t	t
d	d	d
k	k	k
g	g	g
Phonemic affricates		
p͡f	pf	pf
t͡s	ts	ts
t͡ʃ	tS	tS
Fricatives		
f	f	f
v	v	v
s	s	s
z	z	z
ʃ	S	S
ʒ	Z	Z
ç	C	C
j	j	j
x	x	x
h	h	h
Sonorants		
m	m	m
n	n	n
ŋ	N	N
l	l	l
ɾ	r (R)	r

Table B.1.: German consonants in IPA, SAMPA and HTK-SAMPA [37].

IPA	SAMPA	HTKSAMPA
Checked vowels		
ɪ	I	I
ɛ	E	E
a	a	a
ɔ	O	O
ʊ	U	U
ʏ	Y	Y
ɶ	9	oe
Free vowels		
iː	i:	i:
eː	e:	e:
ɛː	E:	E:
aː	a:	a:
oː	o:	o:
uː	u:	u:
yː	y:	y:
ʌː	2:	ox:
Diphthonges		
a͡ɪ	aI	aI
a͡ʊ	aU	aU
ɔ͡ʏ	OY	OY
Schwa		
ə	@	eh
ɐ	6	ah

Table B.2.: German vowels in IPA, SAMPA and HTK-SAMPA [37].

B.2. Word lists

B.2.1. Connected Digits

null drei acht zwo
null zwo neun sieben
zwo fuenf eins null
eins fuenf zwo sieben
fuenf zwo null fuenf
zwo acht fuenf drei
sechs zwo neun null
acht eins sieben zwo
drei vier zwo drei
sechs sieben sieben zwo
sechs eins acht sechs
fuenf vier zwo neun
vier sechs fuenf vier
null zwo sechs sieben
fuenf drei vier fuenf
neun zwo fuenf null
zwo neun zwo sieben
fuenf neun sechs acht
zwo vier neun zwo
zwo neun fuenf neun
zwo vier neun sechs
neun zwo neun sechs
sechs null eins zwo
neun acht fuenf null
zwo sechs zwo neun
zwo acht sechs acht
zwo drei fuenf vier
sechs sieben null zwo
vier eins vier acht
sechs zwo drei drei
sechs zwo sechs fuenf
null zwo zwo neun
zwo sechs eins eins
vier fuenf zwo drei
zwo sieben vier sechs
sieben drei sieben zwo
fuenf neun null zwo
acht drei sieben zwo
vier zwo fuenf vier
zwo sieben zwo zwo
sieben sechs fuenf sieben
drei null zwo eins
neun fuenf neun vier
fuenf fuenf neun drei
sechs zwo neun acht
sechs null vier null
null acht eins acht
drei fuenf null fuenf
drei drei zwo null
null neun eins acht
zwo drei drei vier
zwo neun neun zwo
sieben neun acht acht
null neun fuenf sechs
zwo vier fuenf eins
null null sechs fuenf
sechs sieben drei eins
sieben eins zwo null
null sechs zwo neun
vier drei sieben sieben
acht zwo vier eins
zwo drei zwo fuenf
vier null acht vier
fuenf zwo eins zwo
vier zwo zwo fuenf
null zwo acht sieben
sechs fuenf zwo eins
eins fuenf acht zwo
vier drei vier vier
sieben neun eins zwo
acht null zwo null
drei sieben eins neun
drei acht sechs neun
drei neun acht sechs
sieben null eins null
sechs eins sechs eins
drei eins sechs eins
drei acht zwo zwo
eins drei null neun
zwo neun fuenf zwo
vier neun sieben sechs
sechs acht zwo fuenf
null sechs eins null
vier eins fuenf eins
sieben neun drei sechs
drei null drei eins
eins drei acht acht
zwo zwo vier zwo
drei sechs fuenf sechs
neun fuenf zwo sieben
fuenf fuenf eins eins
sechs sieben drei vier
sieben vier null neun
drei null null zwo
acht eins eins zwo
drei zwo null neun
neun fuenf zwo eins
neun drei eins vier
sechs null zwo acht
fuenf eins fuenf acht

B.2.2. Command Words

eins
zwei
drei
vier
fuenf
sechs
sieben
acht
neun
null
zwo
Zahl
Rechner
mal
plus
minus
durch
Prozent
Komma
Seite
vor
lauter
leiser
stopp
Pause
abspielen
Texte
senden
Kontakte
simon
Liste
schliessen
ausschalten
schlafen
zuhoeren
Hilfe
Ja
Nein
aus
ein
links
rechts
hinauf
hinunter
rauf
runter
zurueck
weiter
ok
abbrechen
loeschen
wiederholen
suchen
Eingabe
Farbe
rot
gruen
blau
gelb
Empfaenger
Training
langsam
schnell
Computer
Englisch
Deutsch
oeffnen
vergroessern
verkleinern

B.2.3. Rhyme-test Words

Test 1	Bach	Bann	bang	Bank	Ball	bald
Test 2	Ball	buhl	bell	Boell	Beil	beul
Test 3	biet	bitt	Beet	Bett	boet	baut
Test 4	back	Bock	bueck	boeg	Boeck	beug
Test 5	Boss	Bosch	Bock	Bonn	Bord	Born
Test 6	bis	Bass	Bus	buess	boes	beiss
Test 7	Buff	Bus	Busch	Bucht	bum	bunt
Test 8	des	daecht	Depp	Deck	daemm	denn
Test 9	wen	sehn	den	Gen	zehn	lehn
Test 10	Sicht	dicht	Gicht	nicht	richt	Licht
Test 11	dir	der	dar	Dur	duerr	doerr
Test 12	doch	Docht	Dock	dort	Dorn	doll
Test 13	saeng	haeng	draeng	peng	meng	laeng
Test 14	wumm	stumm	dumm	drum	Mumm	Rum
Test 15	viel	fehl	Fall	Fell	fuell	feil
Test 16	Siel	fiel	schiel	Ziel	Kiel	Nil
Test 17	Flip	flitz	flick	Flint	flink	flirr
Test 18	focht	Fock	vom	von	fort	vorn
Test 19	Fund	Schund	Hund	bund	Mund	rund
Test 20	Geck	Gent	Gerd	gern	gell	Geld
Test 21	wies	hiess	dies	giess	nies	liess
Test 22	fing	hing	dring	ging	Thing	Ring
Test 23	Graf	grab	Grat	Graz	Gram	Gral
Test 24	Schuss	Bus	Guss	Kuss	muss	Nuss
Test 25	hin	Hahn	Hohn	Huhn	hoehn	Hain
Test 26	weiss	heiss	beiss	Geiss	Reis	leis
Test 27	Hall	hell	hoehl	Hoell	Heil	heul
Test 28	Hieb	heb	hob	hopp	Hub	hupp
Test 29	hier	Heer	Haar	harr	Herr	hoer
Test 30	hiss	hin	hing	hink	Hirt	Hirn
Test 31	Sinn	hin	bin	drin	Zinn	Kinn
Test 32	Hof	Hos	hoch	hob	Hohn	hohl
Test 33	hat	hott	Hut	haett	heut	Haut
Test 34	Wacht	sacht	dacht	Macht	Nacht	Jacht
Test 35	war	Bar	gar	Kar	Jahr	rar
Test 36	doch	poch	noch	Joch	roch	Loch
Test 37	Haff	baff	gaff	paff	Kaff	raff
Test 38	Kien	Kinn	Kahn	kenn	kuehn	kein
Test 39	Kuss	kusch	kuck	kund	Kurt	Kult
Test 40	leis	Laich	leicht	Leib	Leid	Leim
Test 41	lieg	leg	lag	Leck	lueg	laug
Test 42	Lied	litt	lad	luett	Leid	Leut
Test 43	Los	Lob	Lot	lotz	log	Lohn
Test 44	Mast	mach	Macht	matt	Matz	Mann
Test 45	Mehl	Mal	Moll	Mull	Muell	Maul
Test 46	mies	miss	Mass	muss	Mais	Maus
Test 47	Wein	Schein	dein	mein	nein	rein
Test 48	miet	mit	matt	Mett	Maid	Maut
Test 49	Mief	mies	miet	Miez	mim	mir
Test 50	wisch	Fisch	drisch	Tisch	zisch	misch
Test 51	Mist	Most	musst	messt	muesst	meist

Test 52	Muff	muss	Mumm	Mund	murr	Mull
Test 53	Nas	nach	Naab	Naht	nag	nahm
Test 54	wenn	saenn	Fenn	denn	nenn	renn
Test 55	Nest	Nepp	nett	Netz	neck	nenn
Test 56	Saus	Haus	Gauss	paus	raus	Laus
Test 57	Pest	Pech	petz	penn	peng	pell
Test 58	sag	Tag	mag	nag	jag	rag
Test 59	wann	dann	dran	Tann	Mann	ran
Test 60	rauf	raus	Rausch	Rauch	Raub	Raum
Test 61	reif	reich	reib	reit	Reim	Rhein
Test 62	rief	Riff	raff	Ruf	Reif	rauf
Test 63	Riff	Riss	Rist	richt	rinn	Ring
Test 64	Ritt	Reet	rot	raet	rett	reit
Test 65	Rist	Rost	Rest	ruest	roest	reist
Test 66	Ruf	Russ	Ruch	Ruth	Ruhm	Ruhr
Test 67	sass	Saat	sag	sahn	Saar	Saal
Test 68	sind	sehnt	Sand	Sund	send	suehnt
Test 69	sacht	Sack	sann	Sand	sang	sank
Test 70	seht	Saat	satt	saet	Sued	seit
Test 71	Saum	Schaum	Baum	Zaun	kaum	Raum
Test 72	schief	Schiff	Schaf	schaff	schuf	schuef
Test 73	schiel	Schill	Scheel	Schall	scholl	schael
Test 74	schier	schirr	scher	Schar	scharr	schurr
Test 75	schief	schiess	schieb	schied	schien	schiel
Test 76	schwitz	schwimm	schwind	schwing	schwirr	schwill
Test 77	sind	find	schind	bind	Kind	Rind
Test 78	Sinn	sehn	sann	Sohn	Senn	sein
Test 79	Stiel	still	Stahl	Stall	stell	steil
Test 80	Wut	Sud	gut	Nut	Ruth	lud
Test 81	fad	bat	Tat	Maat	Naht	Rat
Test 82	wer	sehr	Teer	zehr	Meer	leer
Test 83	weil	Seil	heil	peil	Teil	Zeil
Test 84	West	Fest	best	Test	Nest	Rest
Test 85	wir	vier	Bier	Pier	Tier	mir
Test 86	vor	bohr	gor	Tor	Moor	Rohr
Test 87	trief	triff	traf	troff	Treff	troeff
Test 88	Trieb	Trip	trapp	Trupp	trueb	treib
Test 89	triff	trist	Tritt	Trick	trimm	trink
Test 90	saet	baet	taet	naeht	jaet	raet
Test 91	Wahn	sahn	Bahn	Zahn	mahn	Lahn
Test 92	Wild	Wald	waehlt	Welt	wuehlt	weilt
Test 93	was	Hass	das	Pass	nass	lass
Test 94	wieg	Weg	wag	wog	waeg	weck
Test 95	web	Weg	wem	wen	Wehr	Wert
Test 96	weiss	weich	Weib	weit	Wein	weil
Test 97	Wind	Wand	wohnt	wund	waehnt	weint
Test 98	wisch	wich	Wicht	wink	will	wild
Test 99	vorn	Horn	Born	Dorn	Zorn	Korn
Test 100	Zupf	Zapf	Zipf	Zoepf	Zaepf	Zopf

Table B.3.: Full list of the 100 sets of Sotschek rhyme-test.