Graz University of Technology

Bachelorarbeit

Detektion von Bursts in Plosiven für die automatische Spracherkennung

erstellt im Rahmen des
Speech Processing Seminars

Institut für Signalverarbeitung
und Sprachkommunikation
Technische Universität Graz

von
Sebastian Grill und André Menrath

Betreuer & Begutachter:
Dr.Mag.rer.nat. Barbara Schuppler

Graz, 10. Juli 2013

Eidesstattliche Erklärung

Wir erklären an Eides statt, dass wir die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht haben.

Graz, am	(Unterschrift)
Graz, am	(Unterschrift)

Inhaltsverzeichnis

1 Einführung **7**
1.1 Charakterisierung von Plosiven . . . 7
1.2 Typen von Plosiven . . . 8
1.2.1 Stimmhafte und stimmlose Plosive . . . 8
1.2.2 Ort der Striktion . . . 10
1.2.3 Reduktion . . . 12
1.3 Rolle des Bursts für die menschliche Spracherkennung und Perzeption . . . 15
1.4 Rolle von Bursts in der automatischen Spracherkennung . . . 17
1.5 Ziel der vorliegenden Arbeit . . . 17

2 Datenbank **19**
2.1 Wahl des Sprachcorpus . . . 19
2.2 Das Kiel Corpus . . . 19
2.2.1 Struktur . . . 19
2.2.2 Segmentierungsdatei . . . 20

3 Detektor **21**
3.1 Überblick . . . 21
3.2 Funktionsweise . . . 21
3.2.1 Signalaufbereitung . . . 21
3.2.2 Entscheidungsstufe . . . 24
3.3 Parametrisierung . . . 25
3.3.1 Vorgehensweise bei der Parameteroptimierung . . . 26

4 Ergebnisse **29**
4.1 Datenset . . . 29
4.2 Auswertungsparameter . . . 29
4.2.1 Trefferquote . . . 29
4.2.2 Temporale Genauigkeit . . . 30
4.3 Auswertung . . . 30
4.4 Analyse anhand repräsentativer Beispiele . . . 32
4.4.1 Korrekte Detektion . . . 32
4.4.2 Fehlerhafte Detektion . . . 32
4.4.3 Grenzfälle und Fehler in der Datenbank . . . 33
4.4.4 Multiple Bursts . . . 34

5 Conclusio **35**
5.1 Diskussion der Ergebnisse . . . 35
5.2 Ausblick . . . 36

Literaturverzeichnis **36**

1 Einführung

Dieses Kapitel soll einen groben Überblick darüber geben, wie überhaupt ein Plosiv definiert ist, welche Mechanismen zu seiner Entstehung beitragen, welche Arten bzw. Realisierungen von Plosiven es gibt und was im folgenden unter einer Plosivlösungsphase (im Englischen: Burst) verstanden wird. In weiterer Folge soll dargelegt werden, welche Rolle Plosive und Bursts in der menschlichen Sprachwahrnehmung spielen, warum die Kenntnis über das Vorhandensein/Fehlen eines Bursts für automatisierte Spracherkennung wichtig ist und worin somit die Motivation für die vorliegende Arbeit liegt.

1.1 Charakterisierung von Plosiven

Unter einem Plosiv versteht man jene Laute, die folgendermaßen artikuliert werden: Es erfolgt ein vollständiger Verschluss (engl. Closure) des Vokaltraktes, wobei der Atemluftstrom dabei vollkommen blockiert wird. Durch die unmittelbar darauf folgende Wiederöffnung des Vokaltraktes ensteht durch das rasche sich „Lösen“ des angestauten Luftdrucks der für einen Plosiv charakteristische Klang, im folgenden als Burst bezeichnet. Physikalisch stellt dieser Burst einen raschen breitbandigen Anstieg der Energie dar [1] [2].

Die Grafik 1.1[1] zeigt den Signalverlauf und das Spektogram eines von einem männlichen Sprecher realisierten /d/. Es handelt sich dabei um das mittlere im Wort „Nordwind“. Im Spektrogram kann man sehen, dass bei einsetzendem Verschluss die Energie des Signals, beginnend von den höheren Frequenzen her abnimmt. Während des Verschlusses, welcher circa von 920 ms bis zum Einsetzen des Bursts reicht, ist im Frequenzbereich über einem kHz nahezu keine Energie vorhanden. Wieso im tieffrequenten Teil dennoch weiterhin sichtbar mehr Energie präsent ist, lässt sich dadurch erklären, dass dieser tiefe Frequenzbereich durch das Organ, welches den Verschluss bewirkt, nicht stärker gedämpft wird. Bei circa 956 ms wird der Verschluss wieder gelöst und es ensteht ein Burst. Ein wesentliches Kritierium für einen Burst stellt das etwa gleichmäßige Vorhandensein von Energie im gesamten Frequenzbereich dar. Bei Vokalen jedoch gibt es eine deutliche Tendenz, dass der Schwerpunkt der frequentiellen Energieverteilung bei den tiefen Frequenzen liegt.

[1] Alle nicht anderweitig gekennzeichneten Spektrogramme in dieser Arbeit wurden mit einer von den Autoren modifizierten Version der Open Source Toolbox „Voicebox“ für Matlab erstellt. [3]

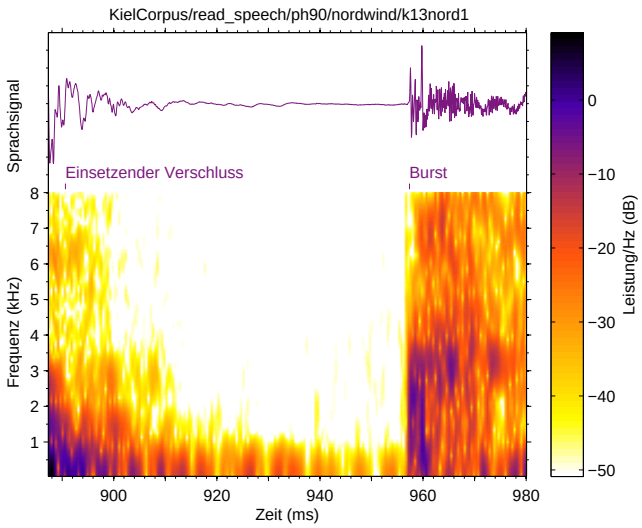


Abbildung 1.1: Spektogram und Signalverlauf eines kanonisch artikulierten Plosivs [4].

1.2 Typen von Plosiven

Aufgrund der Komplexität des Vokaltraktes und der Vielzahl von Kombinationen im Zusammenspiel seiner einzelnen Komponenten gibt es bei Plosiven eine Reihe von Möglichkeiten, wie diese artikuliert werden können. Ein Charakteristikum ist beispielsweise, ob bei der Aussprache eine Beteiligung des Stimmapparates gegeben ist. Je nachdem ob eine solche Beteiligung vorliegt, unterteilt man Plosive in stimmhaft (/b/, /d/, /g/) oder stimmlos (/p/, /t/, /k/) ([2], S.170f).

Eine andere Unterteilung basiert auf der Stelle im Vokaltrakt, an der sich der Verschluss formt. Die Sprachwissenschaft trifft hier sehr genaue Unterscheidungen zwischen einer Kombination von Abwandlungen der Verschlussarten und den damit verbundenen Bewegungsabläufen (siehe Abb. 1.5, 1.8, 1.10 bzw. [5], S.323ff). Dieses Kapitel beschränkt sich jedoch zum Zweck eines groben Überblicks auf die Betrachtung der drei grundlegenden Typen von Striktion, und zwar alveolar (zwischen Zunge und Alveolarkamm), bilablial (zwischen den Lippen) und velar (zwischen hinterem Teil der Zunge und der Decke der Mundhöhle).

Weiters werden Plosive gerade in spontaner Sprache häufig nicht vollständig kanonisch artikuliert bzw. teilweise sogar vollständig ausgelassen (Reduktion).

1.2.1 Stimmhafte und stimmlose Plosive

Stimmhafte Plosive

Wie bereits erwähnt liegt bei stimmhaften Plosiven eine Beteiligung des Stimmapparates vor. Die so hervorgerufenen Spektralanteile stellen den Großteil der während der Verschlussphase eines Plosivs vorkommenden Energie. Dies wird dadurch bedingt, dass die Filterwirkung des Verschlusses Energie in höheren Bändern „schluckt“, während tiefere Frequenzen weniger stark bedämpft werden (siehe Abb 1.2).

Wegen des Druckstaus im verschlossenen Vokaltrakt kann ein stimmhafter Verschluss im Grunde genommen nur sehr kurz dauern, ohne dass die Stimmbänder aufgrund fehlendem Luftflusses aufhören zu schwingen. Es kommt jedoch vor, dass die Form des Vokaltraktes angepasst wird, um zusätzliches Volumen für Luft zu schaffen und so die Schwingung der Glottis über einen längeren Zeitraum aufrechtzuerhalten. Es besteht sowohl die Möglichkeit, dass die Glottis während der gesamten Verschlussphase weiterschwingt, als auch dass es aufgrund des zuvor genannten

Druckstaus zu einer Reduktion der Amplitude oder einem kurzzeitigen kompletten Aussetzen der Schwingungen kommt ([2], S.172ff).

Spektral wird in [2] noch zwischen krächzender (creaky), modaler und gehauchter (breathy) Stimme differenziert. Der Unterschied liegt dabei primär in der Spannung der Stimmlippen und dem daraus resultierenden Verhältnis zwischen dem Zeitraum, in dem die Stimmlippen geschlossen und dem in welchem sie offen sind.

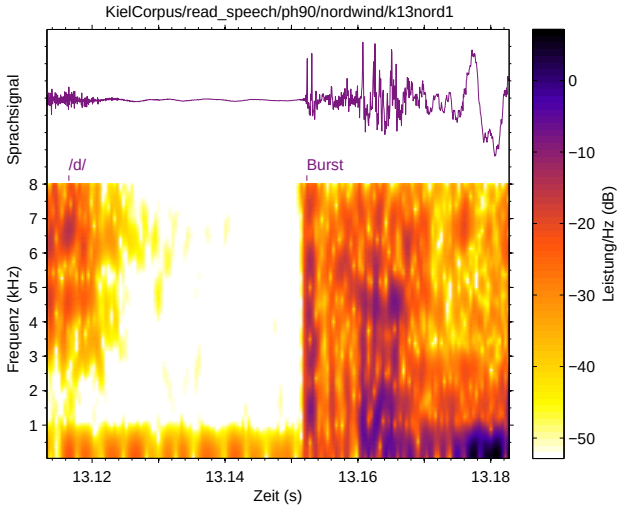


Abbildung 1.2: Spektrogramm von /d/ (aus [4]).

Stimmlose Plosive

Ein stimmloser Plosiv liegt dann vor, wenn die Stimmlippen während der Artikulation des Plosivs nicht schwingen. Die Unterschiede im Spektrum liegen auf der Hand, während der Verschlussphase ist deutlich weniger Energie in dem Frequenzbereich unter $1000Hz$ vorhanden. Siehe dazu Abb. 1.2 und 1.3, die geringere Energie unter $1000Hz$ sticht in Abb. 1.3 deutlich heraus.

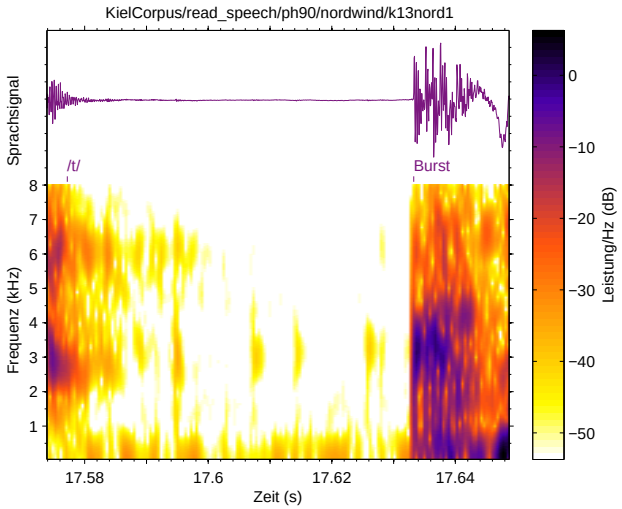


Abbildung 1.3: Spektrogramm von /t/ (aus [4]).

1.2.2 Ort der Striktion

Bilabiale Plosive

Bei der bilabialen Striktion werden die Ober- und Unterlippe gegeneinander gepresst (Abb. 1.4). Der Ort des Verschlusses liegt somit am äußersten Ende des gesamten Vokaltrakts. Abb. 1.5 zeigt ein Modell der Möglichkeiten, wie der Vokaltrakt hinter dem Verschluss ausgeformt werden kann, um den Klang zu beeinflussen ([5] S.340 ff). Zu beachten ist, dass aufgrund der Lage des Verschlusses beim bilabialen Typ in der Burstphase kein Resonator vorhanden ist. Das Spektrum des Bursts weist daher keine Beeinflussung durch Formanten auf und ist relativ diffus ([2], S.176). Zu den Plosiven mit labialem Verschluss zählen `/b/` und `/p/`.

Abbildung 1.4: Bilabialer Verschluss (aus [5]).

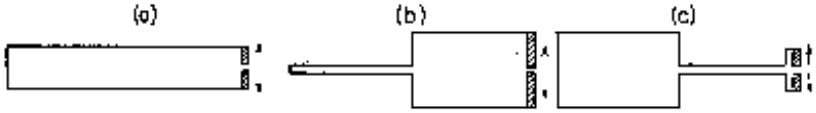


Abbildung 1.5: Akustomechanisches Modell des bilabialen Verschlusses mit unterschiedlicher Formantenausprägung (aus [5]).

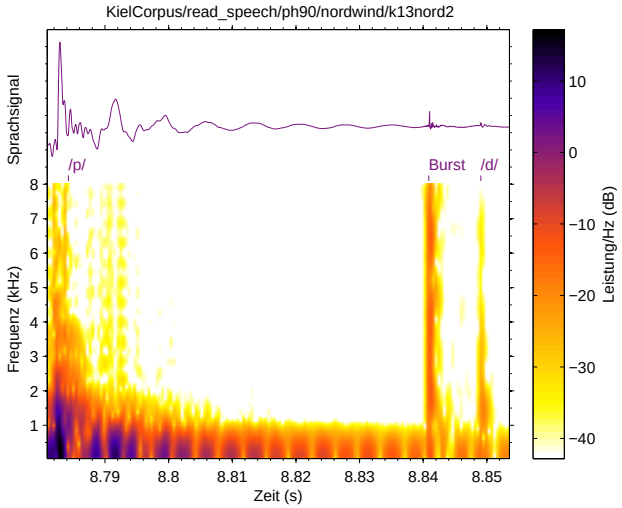


Abbildung 1.6: Spektrum von /p/ (aus [4]). Das Label /d/ gehört bereits zum nächsten Phonem.

Alveolare Plosive[2]

Wie in Abb. 1.7 ersichtlich, wölbt sich bei einem alveolaren Plosiv die Spitze der Zunge nach oben, um mit dem Alveolarkamm einen Verschluss zu formen, der typischerweise 1.5 bis 2.5cm hinter der Lippenöffnung liegt. Aufgrund der Form der Öffnung, die sich in der Lösungsphase bildet, die Geschwindigkeit mit der sich die Öffnung vergrößert sowie der Form der Mundhöhle, die von der Zunge beeinflusst wird, stellen sich bestimmte spektrale Verläufe ein. In den in [5], S.355 ff dargelegten Versuchen zur Modellierung sowie auch Versuchen mit einem Sprecher konnten im tieffrequenten Bereich kaum Unterschiede zwischen labialem und alveolarem Spektrum festgestellt werden, zu hohen Frequenzen hin enthielt das Spektrum von alveolaren Plosiven jedoch deutlich mehr Energie. In [2], S.176 wird dies dem kleinen Resonatorvolumen zwischen Verschluss und Mundöffnung mit dementsprechend hoher Eigenfrequenz zugerechnet. Zu den Plosiven mit alveolarem Verschluss zählen `/d/` und `/t/`.

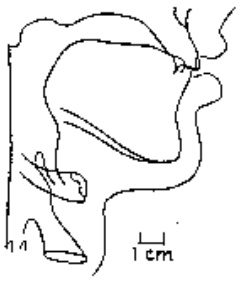


Abbildung 1.7: Alveolarer Verschluss (aus [5]).

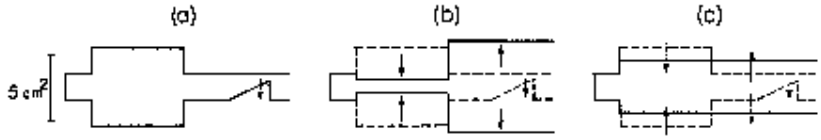


Abbildung 1.8: Akustomechanisches Modell des alveolaren Verschlusses mit unterschiedlicher Formantenausprägung (aus [5]).

Bezüglich eines exemplarischen Spektrums siehe Abb. 1.3.

Velare Plosive

Der velare Verschluss nimmt unter den Striktionsarten eine gewisse Sonderstellung ein. Einerseits liegt der Verschluss hier bedeutend weiter hinten im Vokaltrakt als bei den beiden zuvor angeführten Möglichkeiten, typischerweise $5 - 6cm$, was zur Folge hat, dass auch während der Verschlussphase der vordere Teil des Vokaltraktes offen bleibt. Andererseits ist der Verschluss selbst, wenn in der Sagittalebene betrachtet, merklich breiter. Auch die Größenordnung der zeitlichen Änderung des Querschnitts, sowohl bei der Verschluss- als auch der Öffnungsphase, ist deutlich niedriger als bei den beiden anderen Verschlussarten und trägt somit ebenfalls zur Sonderstellung bei ([5] S.365 ff). In ([2], S.176) wird angeführt, dass das Burstspektrum aufgrund des großen Resonators eine Überhöhung bei tiefen oder mittleren Frequenzen aufweist. Zu den Plosiven mit velarem Verschluss zählen `/g/` und `/k/`.

[2] Genauer müsste man hier noch zwischen alveolar und dental unterscheiden, wobei jedoch in deutscher Sprache dentale Plosive kaum eine Rolle spielen, weswegen aufgrund fehlender Relevanz für diese Arbeit auf eine Ausführung verzichtet wird.

Abbildung 1.9: Velarer Verschluss (aus [5]).

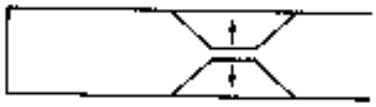


Abbildung 1.10: Akustomechanisches Modell des velaren Verschlusses (aus [5]).

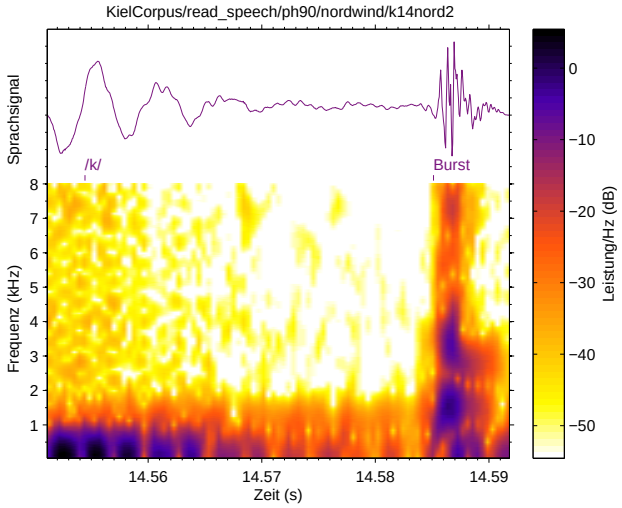


Abbildung 1.11: Spektrum von /k/ (aus [4]).

1.2.3 Reduktion

Von Reduktion spricht man im sprachwissenschaftlichen Kontext, wenn ein Plosiv nicht vollständig kanonisch artikuliert wird. Von kanonischer Artikulation zu vollständiger Reduktion ist der Übergang dabei fließend, generell nimmt die Häufigkeit der Auslassung mit der Spontanität der Sprache zu. Abbildung 1.12 zeigt einige Beispiele, wie Reduktion sich im Sprachsignal äussern kann (aus [6]).

Die Grafiken (A) und (B) zeigen zwei verschiedene Realisierungen von „`kast gegooid`“. Während bei (A) die Verschlussphase mit dem darauffolgenden Burst sehr deutlich auftritt, ist bei Beispiel (B) überhaupt kein Verschluss erkennbar, der Frikativ `/s/` geht nahtlos in den Plosiv `/t/` über, dessen Position auf Basis des Signalverlaufs und Spektrogrammes nur erahnt werden kann. Das `/t/` wird in diesem Beispiel nur durch eine Friktion angedeutet aber nicht wirklich ausgesprochen. In Beispiel (C) wollte der Sprecher „`kast veer`“ sagen, statt einem `/t/` findet sich in dem Signal jedoch nur ein Friktions-Geräusch mit geringer Lautstärke. Beispiel (D) enthält

das Wortpaar „`mest besteld`“. Das Beispiel enthält eine lange Phase mit Stille oder zumindest sehr leiser Friktion von $50ms$. Auf diese Stille folgt der Beginn von `/b/`. Da die Stille und der Beginn von `/b/` mit $100ms$ außerordentlich lang für Fließsprache sind, ist anzunehmen, dass die Verschlussphase sowohl `/t/` als auch `/b/` enthält [6].

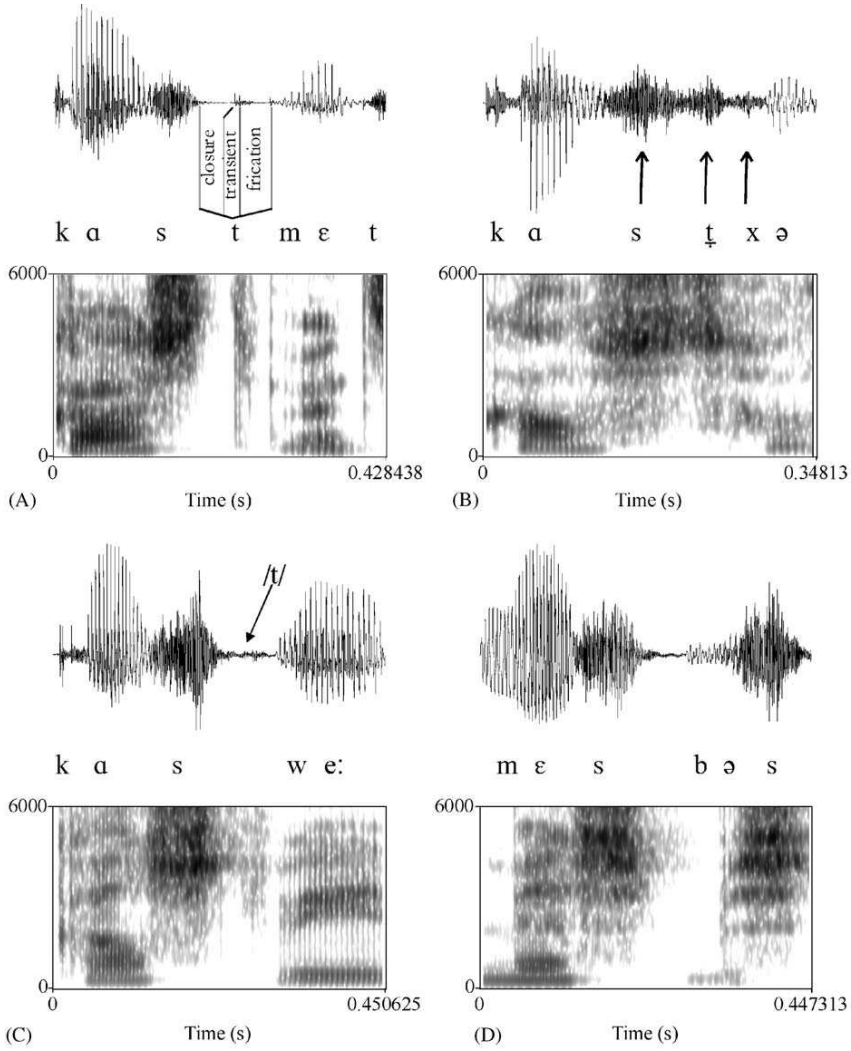


Abbildung 1.12: Unterschiedlich deutliche Aussprache von /t/ (aus [6]):
(A) kanonische Aussprache von „kast gegooid“;
(B) wie (A), jedoch kein Verschluss zwischen /s/ und /t/;
(C) „kast veer“, ohne Verschluss und Burst, stattdessen Friktion mit geringer Amplitude;
(D) „mest besteld“, vollständige Auslassung des /t/ in „mest“

Mitterer und Ernestus stellten in [6] Untersuchungen zur Reduktion von `/t/` in holländischer Sprache und deren Auswirkungen auf menschliche Hörer an. Dazu wurden 2 Untersuchungen, eine mit gelesener, eine mit spontaner Sprache unternommen. Für gelesene Sprache ergab sich, dass im Mittel in 8% der Fälle `/t/` ausgelassen wird, bei spontaner Sprache erhöhte sich der Mittelwert allerdings auf 46%. Es zeigte sich, dass die Auslassung von `/t/` nach `/s/`, unabhängig vom darauf folgenden Kontext, sehr häufig auftrat. Außerdem kam es oft vor bilabialen Lauten zur Auslassung. [3] Gleichzeitig zeigten die ebenfalls durchgeführten Hörversuche, dass menschliche

[3] Die Wahrscheinlichkeit der Auslassung ist stark vom Kontext abhängig, sprich welche Laute vor dem /t/ kommen und welche folgen. Für eine genaue Auflistung der von den einzelnen Parametern abhängigen Ergebnisse

Hörer mit Holländisch als Muttersprache die Reduktion kompensieren. Mitterer und Ernestus stellten dabei eine Korrelation fest, nämlich „hörten“ die Versuchspersonen intuitiv nach /s/ öfter ein /t/ als nach beispielsweise /n/.

[7] enthält ebenfalls eine Studie bezüglich Reduktion von wortfinalen /t/s in holländischer Sprache, hier wurde im Vergleich zu [6] auch Wert auf die Untersuchung der morphologischen Eigenschaften der Tokens gelegt. Es wurde gezeigt, dass /t/ bedeutend öfter reduziert wird, wenn es nur Teil des Wortstammes ist und ihm keine Bedeutung als grammatisches Morphem zukommt. Im Gegensatz zu [6] wurde außerdem bei einer Untersuchung die Reduktion von /t/ anhand von subphonemischen Features untersucht, anstatt mithilfe einer Datenreduktion nur zwischen Vorhandensein und Fehlen des /t/ zu unterscheiden. Die Ergebnisse dieser Untersuchung zeigen, dass dabei eine sehr große Variabilität vorliegt. So wurden etwa nur 11.7% aller Tokens kanonisch realisiert. Etwas mehr als 50% wiesen mindestens einen Burst auf, von diesen 50% wiesen etwa 60% mehrere Bursts auf.

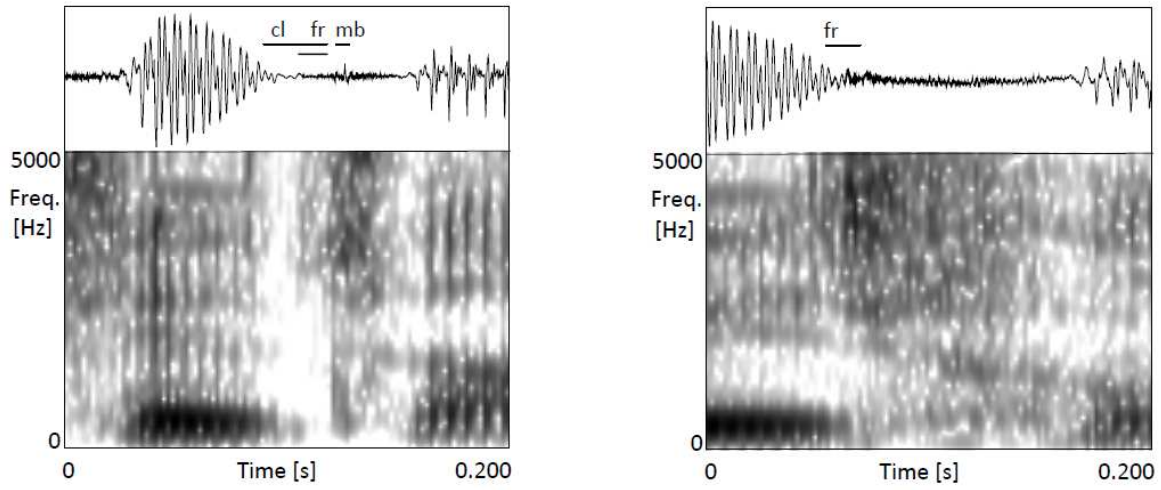


Abbildung 1.13: Linkes Bild: /zita/ cl = Verschluss, fr = Friktion, mb = Multipler Burst; Rechtes Bild: /itva/ fr = Beginn der /t/-Friktion (aus [7])

Abbildungen 1.13 und 1.14 zeigen einige der in [7] untersuchten Realisationen von /t/. Das linke Bild in Abb. 1.13 eine Artikulation mit einem Verschluss, einer Friktion während des Verschlusses und darauffolgenden multiplen Bursts. Das Sprachsignal im rechten Bild enthält eine Realisation, bei der Verschluss und Burst vollständig fehlen, stattdessen findet ein kontinuierlicher Übergang zu einer Friktion statt. Das linke Bild in Abb. 1.14 enthält ebenfalls einen Übergang in eine Friktion ohne Closure, der allerdings abrupt beginnt. Das rechte Bild zeigt einen stimmhaften Verschluss mit anschließendem Burst und Friktion.

[7] enthält außerdem eine Untersuchung zur Gegenüberstellung von menschlicher und automatischer Spracherkennung bei spontaner Sprache. Dabei wurde zuerst das Datenmaterial von zwei Phonetikern transkribiert, um Aufschlüsse zu erhalten, anhand welcher Eigenschaften Hörer ein /t/ als vorhanden klassifizieren. Anschließend wurde der selbe Versuch mit einem automatischen Spracherkennungssystem (ASR) durchgeführt und die Ergebnisse verglichen. Die Ergebnisse zeigen, dass das ASR-System aufgrund seines Trainings auf kanonische Aussprache bei der Detektion von /t/s in spontaner Sprache viele von den Phonetikern als vorhanden gelabelte /t/s nicht erkennt, da es beispielsweise auf schwache Bursts nicht so sensibel reagiert wie menschliche Hörer. Auch stimmhafte Realisierungen bereiten dem System größere Schwierigkeiten.

wird auf [6] verwiesen.

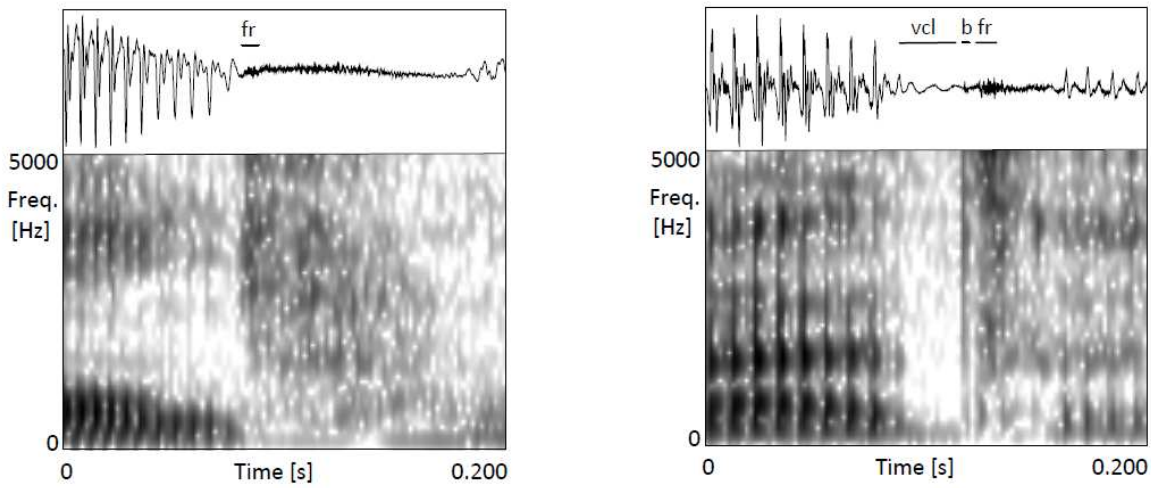


Abbildung 1.14: Linkes Bild: /antv/ fr = Abrupter Beginn der /t/-Friktion;
Rechtes Bild: /ayte/ vcl = Stimmhafter Verschluss, b = Burst, fr = Beginn der Friktion (aus [7])

Generell bereitet Reduktion, wie sie in spontaner Sprache sehr häufig vorkommt, statistisch arbeitenden Spracherkennungssystemen große Probleme. Abhilfe verspricht sich die Forschung derzeit davon, zuerst eine grobe Unterteilung des Sprachsignals vorzunehmen und anschließend in Segmenten wo dies sinnvoll ist mit anderen, auf menschlicher Wahrnehmung basierenden Verfahren detailiertere Informationen zu extrahieren [7][8].

1.3 Rolle des Bursts für die menschliche Spracherkennung und Perzeption

Welche Bedeutung der Burst für die korrekte Erkennung eines Plosivs für den Menschen hat wurde von Abhinauv Kapoor und Jont B. Allen untersucht [9]. Sie stellten fest, dass die relative Energie des Plosiv–Bursts in Bezug zum Vokal in direktem Zusammenhang mit der Verständlichkeit des Plosives steht, d.h. ein gewisser Plosiv korrekt als dieser erkannt wird. Weiters deklarieren sie den Burst als primäre, für die Plosividentifikation entscheidende, akustische Eigenschaft.

Zu diesem Ergebnissen kamen sie, indem sie Plosive auf verschiedene Art und Weise modifizierten und diese Testpersonen vorspielten. Die Modifizierungen sahen zum einen so aus, dass der Burstanteil im Signal einmal nicht verändert 1.15 a), ein anderes mal um 6 dB angehoben 1.15 b), um 6 dB abgesenkt 1.15 c), und zuletzt auch komplett entfernt wurde 1.15 d). Zusätzlich wurde anschließend noch der Signal-Rauschabstand dieser modifizierten Sprachsignale variiert.

Das wichtigste Ergebnis ihrer Arbeit ist, dass die Wahrscheinlichkeit der korrekten Identifikation eines Plosives bei allen Signal-Rausch-Verhältnissen in etwa proportional mit der Stärke der Burst Modifikation zu oder abnimmt. Für den Fall, dass das Hauptmerkmal, i.a.W. der Burst, vollständig entfernt wurde, waren die Testpersonen quasi nicht mehr in der Lage den Plosiv zu erkennen — auch bei höchsten Störabständen.

Die Abbildung 1.16 zeigt, wie sich bei einem durch den männlichen Sprecher „115“ gesprochenes `/ka/` die Warscheinlichkeit der korrekten Plosividentifikation durch eine Testperson, aufgrund der vorgenommenen Modifizierungen ändert. Für alle vier Modifizierungen wurde in Abhängigkeit des Signal-Rausch-Verhältnisses die Wahrscheinlichkeit der korrekten Erkennung des Plosivs

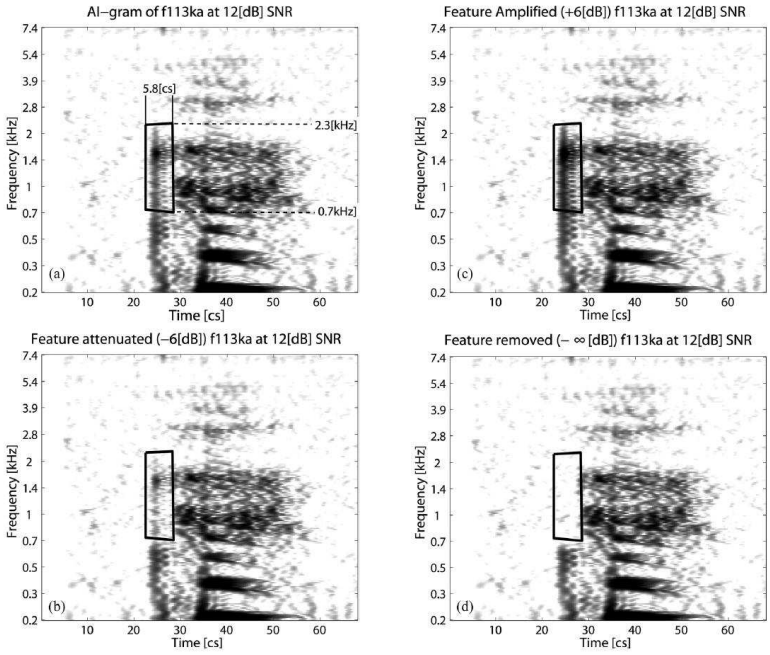


Abbildung 1.15: Spektrogramme der modifizierten Plosive. Der jeweils eingerahmte Bereich in der Zeit-Frequenz-Ebene zeigt den Bereich der Modifikation. Burst-Anteil a) unverändert, b) um 6 dB angehoben, c) um 6 dB abgesenkt und d) vollständig entfernt (aus [9]).

aufgezeichnet. Hebt man den Burst-Bereich um 6 *dB* an, so wird der modifizierte Plosiv im Vergleich zum unveränderten bereits bei einem circa 8 *dB* schlechterem Signal-Rausch-Verhältnis mit über 90 % Wahrscheinlichtkeit richtig klassifiziert. Senkt man das Burst-Feature jedoch um 6 *dB* ab, so muss man bezogen auf das Ursprungssignal den Störabstand um ungefähr 10 *dB* verbessern um ebenfalls eine richtige Plosiverkennung mit über 90 % Wahrscheinlichkeit zu erzielen. Für die Modifikation, wo der Burst im Wesentlichen vollständig entfernt wurde, wird egal bei welchen SNR keine Erkennungswahrscheinlichkeit größer als 40 % erreicht — dabei ist keine signifikante Steigung der Kurve (Verbesserung der Erkennungswahrscheinlichkeit) hin zu höheren Störabständen zu erkennen.

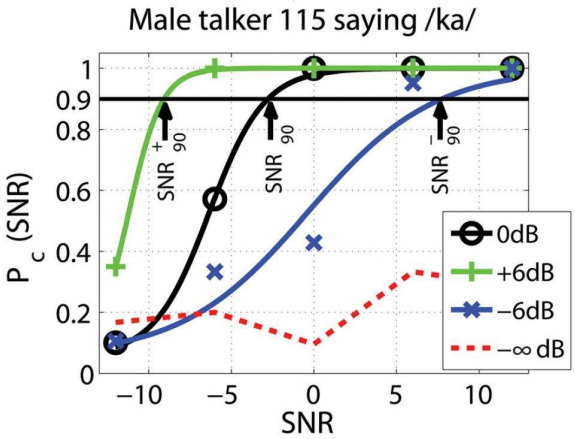


Abbildung 1.16: Grafische Auswertung der Plosivverständlichkeit (aus [9]).

Die Ergebnisse der Untersuchungen an diesem Beispiel decken sich grob mit den meisten anderen, weshalb dieses hier auch exemplarisch gezeigt wird.

1.4 Rolle von Bursts in der automatischen Spracherkennung

Automatische Spracherkennungsverfahren (Automatic Speech Recognition, ASR) basieren derzeit größtenteils auf dem in Abb. 1.17 dargestellten Prinzip. Das Audiosignal wird vorverarbeitet, fouriertransformiert, entsprechend der Mel-Skala gewichtet und anschließend in den Cepstralbereich rücktransformiert. Die so gewonnenen Mel Frequency Cepstral Coefficients (MFCCs) haben

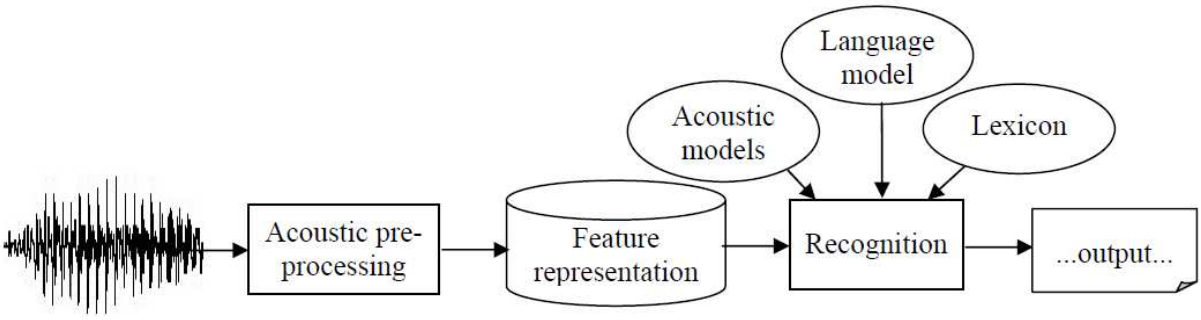


Abbildung 1.17: Schematische Darstellung eines ASR Algorithmus (aus [10]).

aufgrund der Eigenschaften des Cepstrums den Vorteil, dass eine klare Trennung von anregendem Signal (Glottis) und Filter (Mund- und Rachenraum) erreicht wird. Die MFCCs werden schließlich an die statistisch arbeitende Spracherkennungsstufe weitergegeben. In jedem der vorhandenen Modelle werden schließlich die Best Fits ermittelt, die am besten auf ein Wort aus dem Lexikon passende Kombination an Fits aus den Modellen landet anschließend in einer Hypothese. Für jeden Pfad (Kombination von hypothetisierten Wörtern) wird letztlich ein Score erstellt, wobei der Pfad mit dem höchsten Score als Resultat ausgegeben wird [10].

Diese Verfahren erreichen bei ausreichendem Training auf einen Sprecher und für den Fall, dass alle zu erkennenden Wörter im Lexikon vorhanden sind bereits sehr gute Ergebnisse. Falls jedoch vom Ideal abweichende Bedingungen vorherrschen, beispielsweise aufgrund von Störanteilen im Signal, fehlendem Training oder nicht-kanonischer Aussprache (siehe z.B. Kapitel 1.2.3), lässt die Performance dieser Verfahren stark nach.

Ein Ansatz, um dem entgegenzuwirken ist die sogenannte Landmark-based ASR [8][11][12][13]. Im Gegensatz zum klassischen Ansatz wird hier bei der Verarbeitung nicht jeder Frame als gleichwertig betrachtet, stattdessen versucht man, angelehnt an die Funktionsweise der humanen Sprachperzeption, spezielle Anhaltspunkte im Sprachsignal (=Landmarks) zu detektieren, zu denen auch Bursts zählen. Auf diesen Anhaltspunkten fußend kann mit sehr viel spezialisierterer Signalverarbeitung zeitlich gezielt angesetzt werden, um zusätzliche Informationen zu gewinnen, die die Hypothesen unterstützen und präzisieren können.

1.5 Ziel der vorliegenden Arbeit

Ganz allgemein bedarf sowohl die Forschung im Bereich der Linguistik, als auch im Bereich der ASR, entsprechend annotiertem Sprachmaterials. Um die Rolle von Plosiven in diesen Bereichen untersuchen zu können, müssen diese im Sprachmaterial annotiert sein (siehe Kapitel 1.3 und 1.4).

Nach herkömmlicher Methode wird das Sprachmaterial in aufwändiger Handarbeit analysiert und phonetisch segmentiert und Bursts müssen in der Transkription manuell annotiert werden. Daher ist die Erstellung der Transkription kostenintensiv und zeitaufwändig und erschwert der Forschung somit den Zugang zu umfassenden Sprachcorpora. Hinzu kommt noch, dass die

Ergebnisse aus händischer Transkription, insbesondere wenn durch verschiedene Transkribienten erstellt, naturgemäß einer gewissen Streuung unterworfen sind, die mit der Spontanität des gesprochenen Materials zunimmt ([7] und [14]). Aus den genannten Gründen ist es nicht verwunderlich, dass die Entwicklung von Werkzeugen zur automatischen Transkription ein Gebiet laufender Forschung und Entwicklung darstellt.

Die vorliegende Arbeit präsentiert eine automatische Methode, welche eine Komponente eines Transkriptionstools für einen Sprachcorpus darstellt. Bei diesem Tool wird zuerst mit Forced Recognition [14] eine grobe Segmentierung erstellt. In einem zweiten Schritt wird diese Segmentierung schließlich einem Algorithmus zugeführt, der eine möglichst exakte Aussage über das Vorhandensein eines Bursts sowie im Falle des Vorhandenseins über den Zeitpunkt des Auftretens treffen soll. Ziel dieser Bachelorarbeit war die Entwicklung eines eben solchen Algorithmus unter Zuhilfenahme einer vollständigen manuellen Transkription, die es ermöglicht, die Genauigkeit des Detektors zu evaluieren. In weiterer Folge soll das Ergebnis der Arbeit dazu dienen, die Transkriptionsarbeit an dem aktuell am Institut für Signalverarbeitung und Sprachkommunikation der TU Graz in der Entwicklung befindlichen Graz Corpus of Austrian German zu unterstützen.

2 Datenbank

2.1 Wahl des Sprachcorpus

Die Hauptaufgabe eines Sprachcorpus in diesem Projekt war es uns zu ermöglichen, die Funktionsweise des Burst-Detectors zu testen. Dazu wurde eine Sprachcorpus namens Kiel Corpus verwendet, welche am Institut für Phonetik und digitale Sprachsignalverabeitung an der Universität Kiel seit Anfang der 90er Jahre entwickelt wird. Für diese Arbeit entscheidend war, dass hier neben der üblichen phonetischen Transkription und Segmentierung auch der Beginn der Plosivlösungsphase/Aspiration (manuell) etikettiert und zeitlich bestimmt wurde. Wenn in einem Plosiv, so wie er im jeweiligen Falle ausgesprochen wurde, kein Burst mit vorausgehenden Verschluss vorhanden ist, dann kann man auch nur schwer auf einen Beginnzeitpunkt der Plosivlösungsphase schließen. In einen solchen Fall wurde das enstrechende Segmentierungs-Label dann ausgelassen. Diese genaue Segmentierung erlaubte es die Ergebnisse des Burst-Detektors mit den manuellen Werten zu vergleichen und daraus Rückschlüsse für dessen Optimierung zu ziehen.

2.2 Das Kiel Corpus

2.2.1 Struktur

Derzeit umfasst die Datenbank mehr als vier Stunden gelesene, sowie auch knapp vier Stunden spontane deutsche Sprache. Im für in dieser Bachelorarbeit relevanten Teil, dem der Lesesprache, existieren mehrere Textvorlagen, welche von einzelnen Sätzen bishin zu kleineren Geschichten reichen. Diese wurden jeweils von mehreren männlichen und weiblichen Sprechern vorgetragen und aufgenommen. Die Audiodateien liegen mit 16 *bit* Wortbreite und einer Samplefrequenz von 16 kHz als headerlose `.16`-Dateien beginnend mit dem LSB im gelesenen und als `.wav`-Datein im spontanen Teil vor. Zu jeder resultierenden Audiodatei existiert eine dazugehörige Datei mit Segment- und Ettikettierungsinformationen.

2.2.2 Segmentierungsdatei

Jede Segmentierungsdatei beinhaltet folgende Informationsteile, welche durch Schlüsselwörter voneinander abgegrenzt werden:

- Dateinamen
- 7-bit breite orthographische Repräsentation des Gelesenen
- kanonische phonemische Transkription des Textes in SAMPA
- phonemische Transkription der Äußerung (i.a.W.: das tatsächlich realisierte Gesprochene) inkl. „Burst-Label“
- Samplenummern und Etiketten

Für uns von Interesse ist hier nur der letzte Teil: Sample-Nummern und Etiketten. Dieser stellt eine mit dem Audiofile synchronisierte und erweiterte Transkription dar, basierend auf der phonemischen Transkription der Äußerung. Die Samplenummer steht für den Zeitpunkt des Beginns des Signalabschnittes, der für das Label verantwortlich war. Außerdem erhält jedes Element einen Präfix, der Auskunft über die Position des Labels im Wort gibt:

`##`	für wortinitiale Labels
`$`	für wortinterne Labels
`$#`	für wortinterne, Kompositums-initiale Labels
`#`	für wortexterne Labels, z.B. Pausen, Atmen, Interpunktionszeichen

Grafik 2.1 zeigt einen kurzen Auschnitt einer solchen Segmetierungsdatei, in welchem „Nordwind und Sonne.“ gelesen wird. Der Bereich der Samples `14250` und `15318` des zugehörigen Audiofiles dieses Ausschnitts wurde bereits in Abbildung 1.1 grafisch dargestellt. Das Label `$-h` kennzeichnet dabei bei allen Plosivtypen den Zeitpunkt des Bursts. Weitere Labeleigenschaften, wie die für die Transkription von nicht verbalen und nicht vokalen Signalabschnitten, sind in dieser Arbeit von unwesentlicher Bedeutung, weswegen auf eine Erklärung dieser verzichtet wird.

```
10315 #c:
10315 ##n
11892 $'O6
14250 $t
15318 $-h
15682 $#v
17140 $"I
18012 $n
18918 $t
19388 $-h
19492 ##Q
20283 $U
20988 $n
22108 $t+
22636 ##z
23916 $'O
25336 $n
26373 $@
28465 #.
```

Abbildung 2.1: Ausschnitt des Samplenummern–und Etikettenteils der Segmentierungsdatei für read_speech/ph90/nordwind/k13nord1 [4]. Gelesen wird „Nordwind und Sonne“.

3 Detektor

Dieses Kapitel enthält einen detailierten Einblick in die Funktionsweise des Detektors und der in diesem Zusammenhang in `Matlab` implementierten Methoden.

3.1 Überblick

Wie in Kapitel 1.1 und 1.2 angeführt, ist es für kanonisch artikulierte Plosive charakteristisch, dass im für Sprache oberen Spektralbereich ($> 2000 - 3000Hz$) auf eine Phase relativ geringer Energie ein rascher Anstieg der Signalenergie (=Burst) folgt. Es ist daher naheliegend, zum Zwecke der Burstdetektion diesen Spektralbereich gesondert zu betrachten. Ausgestattet mit Vorabinformationen über den ungefähren Zeitraum eines Plosivs (gewonnen aus händischer Segmentierung oder durch vorangestellte automatisierte Signalverarbeitung) ist es damit möglich, mit relativ einfachen Werkzeugen der Signalverarbeitung Bursts zu detektieren.

In [15] wird eine mögliche Herangehensweise vorgestellt. Der Detektor arbeitet hierbei mit dem Leistungsverlauf und dessen Ableitung für den relevanten Spektralbereich. In der vorliegenden Arbeit wurde das Prinzip von [15] aufgegriffen, wobei hinsichtlich Signalaufbereitung vor dem Detektor und auch in der Funktionsweise des Detektors selbst einige Modifikationen vorgenommen wurden.

3.2 Funktionsweise

3.2.1 Signalaufbereitung

In einem ersten Schritt wird im Detektor ein Spektrogramm des zu untersuchenden Audiosegmentes erstellt. Hierfür wurde die `Matlab`-eigene Funktion `spectrogram()` herangezogen. Das Fenster der FFT wurde auf 64 Punkte bei 32 Punkten Überlappung parametrisiert.

```
1  [~,F,T,P] = spectrogram(audioData,64,32,256,Fs);
```

Listing 3.1: spectrogram

Anschließend wird ein parametrisierbarer Hochpass auf die so gewonnenen Daten angewendet, indem die Zeilen der Matrix, welche mit den Frequenzen unterhalb der Grenzfrequenz korrespondieren, ausgelassen werden.

```
1  %% ideal highpass
2
3  ind_hp = F>=f_hp;     % get index of frequencies > f_hp
4  P = P(ind_hp,:);      % write matrix with power spectral
5                        % density according to specified highpass
```

Listing 3.2: Implementierung des Hochpass

Die Leistungsdichtespektren über der Zeit werden nun aufsummiert, um zum zeitlichen Verlauf der Leistung zu gelangen. Da die Entscheidungsstufe des Detektors ihre Informationen aus dem zeitlichen Verlauf der Leistung und der Ableitung der Leistung zieht, ist das so gewonnene Signal aufgrund seiner starken Schwankungen ungeeignet (siehe Abb. 3.1).

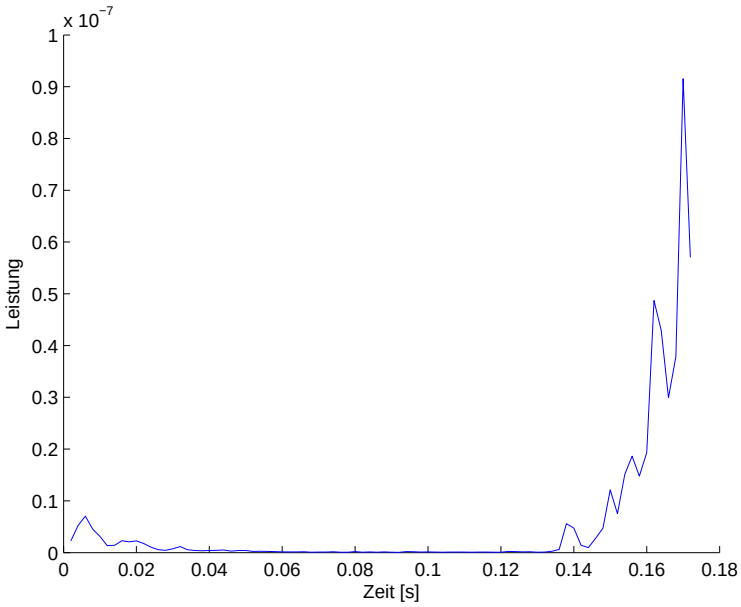


Abbildung 3.1: Exemplarischer Leistungsverlauf für /p/ mit Burst

Es wird daher in einem weiteren Schritt ein einfacher Hüllkurven-Generator, basierend auf Interpolation der lokalen Maxima des Signals, angewandt. Das Interpolationsverfahren `'pchip'` stellte sich in Versuchen als am Geeignetsten heraus, da die so gewonnene Hüllkurve für die weitere Verarbeitung die günstigsten Eigenschaften aufweist.

```
1  %% calculate upper envelope of power
2  idx = find(diff(sign(diff(power))) == -2) + 1;
3  if length(idx) <2           % if the audio sample contains no local maximum,
4      bursttime = 0;          % set bursttime 0 and abort execution, because
5      return                  % the file is effectively too short
6  end
7  power = interp1(T(idx),power(idx),T,'pchip');
```

Listing 3.3: Hüllkurven-Generator

Da der Interpolator sowohl am Anfang wie auch am Ende des Signals aufgrund des möglichen Fehlens von notwendigen Stützstellen, das sich unter Umständen aus dem Signalverlauf ergibt, zu fehlerhafter Extrapolation neigt, wurden Hüllkurvenwerte, die vor dem ersten beziehungsweise nach dem letzten detektierten Maximum liegen, generell verworfen. In diesem Teil des Codes

befindet sich auch eine Kontrollstruktur, die von ihrer Funktionalität eher dem Detektor zuzuweisen ist. Können nicht ausreichend Stützstellen für die Generierung einer Hüllkurve gefunden werden, bricht der Algorithmus ab, da eine sinnvolle Detektion nicht möglich ist. In diesem Fall wird pauschal kein Burst detektiert[4].

```
1  % supress hull curve values before the first and after the last detected
2  % maximum to supress erroneous extrapolation
3  startidx_hullcurve = idx(1); endidx_hullcurve = idx(end);
4  power(1:startidx_hullcurve-1) = NaN; power(endidx_hullcurve+1:end) = NaN;
```

Listing 3.4: Automatisches „Abschneiden“ der Hüllkurve

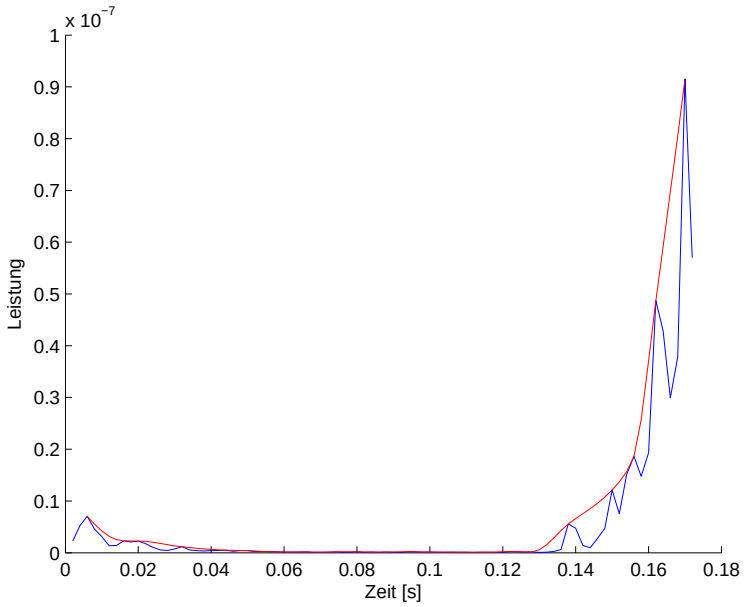


Abbildung 3.2: Hüllkurve

Es ergeben sich daher gewisse Anforderungen hinsichtlich der Länge sowie des zeitlichen Verlaufs, die das Signal zu erfüllen hat, um die einwandfreie Funktionsweise des Hüllkurven-Generators zu ermöglichen.

Bevor nun eine Entscheidung über das Vorhandensein sowie den genauen Zeitpunkt eines Bursts getroffen werden kann, wird schließlich noch die diskrete Ableitung berechnet, indem zu jedem Sample die Differenz der Leistung zu der Leistung des vorherigen Samples gebildet wird.

```
1  % for every timeslice, calculate the energy difference versus the previous
2  % timeslice (= ~discrete 1. derivative)
3  n = 2:length(power);
4  delta_power(1) = power(1);
5  delta_power(n) = power(n)-power(n-1);
```

Listing 3.5: Berechnung der Ableitung

Die so gewonnenen Signale werden schließlich an die Entscheidungsstufe übergeben.

[4] Bezüglich der Auswirkungen dieser Maßnahme auf die Gesamtergebnisse siehe Kapitel 4.3.

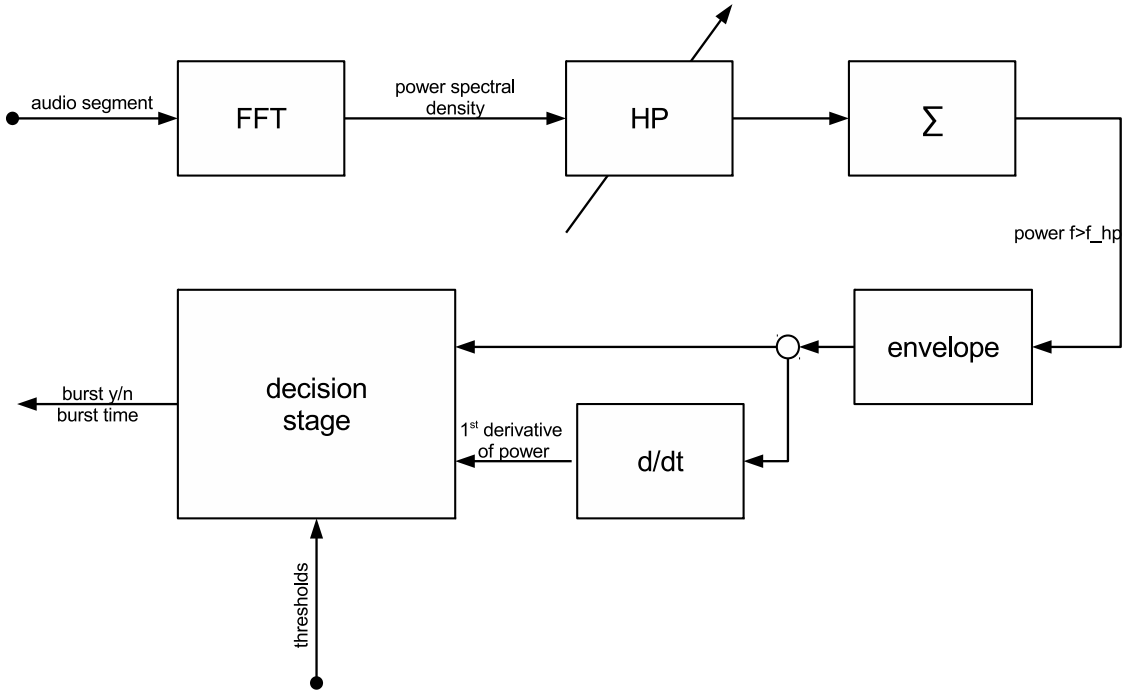


Abbildung 3.3: Blockschaltbild des Detektors: FFT... Fouriertransformation/Spektralanalyse; HP... parametrisierbarer Hochpass; $\sum$... Bildung des Leistungsverlaufs aus den Leistungsdichtespektren; d/dt... diskrete Ableitung

3.2.2 Entscheidungsstufe

Die Entscheidungsstufe basiert auf dem in [15] angeführten Prinzip, allerdings mit leichten Modifikationen. Eine wichtige Abweichung besteht darin, dass zur Verbesserung der Erkennung von fehlenden Bursts ein je nach vorliegendem Plosiv parametrisierbarer, vom Signalverlauf unabhängiger Threshold für die Ableitung der Leistung eingeführt wurde. Dieser Threshold findet in einem Vorabentscheid Anwendung. Wird der Threshold nicht überschritten, so entscheidet der Detektor, dass kein Burst vorliegt.

Die Notwendigkeit für einen weiteren Vorabentscheid resultiert aus der Funktionsweise des Hüllkurven-Generators. Sollte diese Stufe nicht in der Lage sein, mehr als ein lokales Maximum zu detektieren, so ergeben sich nicht ausreichend Stützstellen für eine Hüllkurve, der Algorithmus wird somit abgebrochen und es wird ein pauschaler Negativentscheid gefällt.

An dieser Stelle kommen die plosivabhängig gesetzten Parameter zur Bestimmung der Thresholds der Leistung und deren Ableitung zum Einsatz. Es wird durch die händische Segmentierung vorausgesetzt, dass das Maximum der Leistung im betrachteten Zeitfenster mit dem Burst koinzidiert. Der Detektor bestimmt daher die absoluten Maxima der Leistung und deren Ableitung im betrachteten Zeitfenster sowie den Zeitpunkt, bei dem das Maximum der Leistung auftritt.

In Abhängigkeit von den detetektierten Maxima werden nun die Thresholds berechnet, indem die Extremwerte mit den zugehörigen Parametern multipliziert werden. In einer `for`-Schleife passiert letzten Endes der finale Entscheid, ob ein Burst vorliegt. Die Schleife beginnt am Zeitpunkt des Leistungsmaximums und arbeitet sich in Richtung Anfang des Audiosignals vor. Sobald die

Thresholds für Leistung als auch deren Ableitung unterschritten werden, bricht die Schleife ab und detektiert den Burst. Werden die Thresholds nicht unterschritten, so wird auch kein Burst detektiert.

```
1  %start at maximum power and proceed left
2  for n = abs_max:-1:1
3      if power(n) < thresh_power && abs(delta_power(n)) < thresh_delta
4          burstsample = n;
5          break
6      end
7  end
```

Listing 3.6: Schleife des Detektors

Abschließend wird das Sample, bei dem der Burst detektiert wurde, mit einem variablen Offset versehen. Dies geschieht aus zweierlei Gründen. Einerseits wird durch die Verwendung der Hüllkurve der Zeitpunkt des Energieanstiegs nach vorne verschoben. Andererseits entsteht zwischen Ja/Nein-Entscheidungspräzision und der Detektion des Bursts an der gewünschten Stelle in der `for`-Schleife bei dem vorliegenden Verfahren ein Optimierungsproblem. Beides lässt sich jedoch teilweise umgehen, indem man den Burst generell zu früh detektiert und in weiterer Folge einen, vom vorliegenden Plosiv abhängigen, Offset addiert.

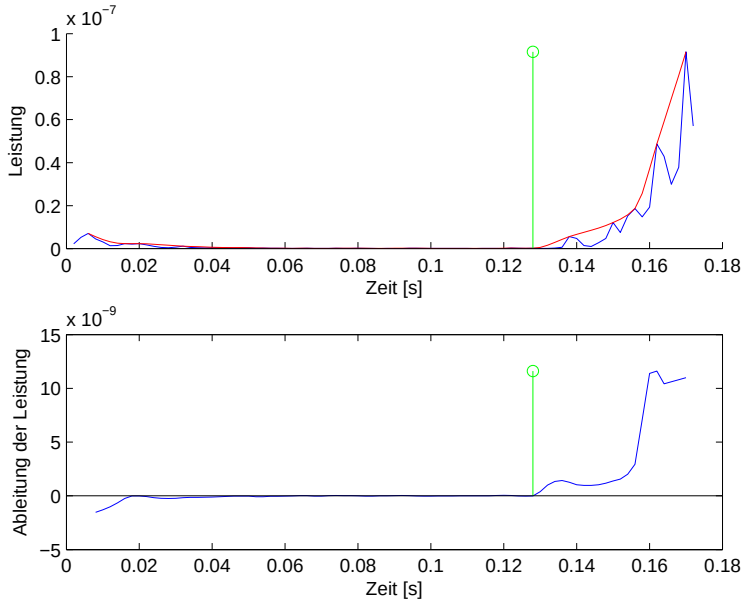


Abbildung 3.4: Signalverläufe mit detektiertem Burst (Offset wurde noch nicht addiert).

3.3 Parametrisierung

Der Detektor verfügt über eine Anzahl von Parametern, die plosivabhängig gesetzt werden können, um auf das Ergebnis Einfluss zu nehmen. Dadurch wird es ermöglicht, die vorliegende Implementierung auch auf andere Datensätze zu trainieren.

Die Parametrisierung erfolgt händisch, im File `detektor.m` befindet sich für diesen Zweck eine Zelle speziell für die Deklaration. Die Einträge für die einzelnen Plosive sind in eine `switch`-Kontrollstruktur eingebettet, der Fall dass kein Plosiv deklariert wurde, wird über einen Default-Eintrag abgefangen. Die Default-Werte befinden sich unter dem `switch`-Konstrukt und können

ebenso einfach manipuliert werden. Die Übergabe der Plosivart an den Detektor erfolgt als ASCII-codierter String (am Beispiel von `/t/` erhält man 116).

```
case 116 % 't'

    % threshold = max * multiplier
    power_multiplier = 0.1;
    delta_multiplier = 0.03;

    % absolute delta threshold under which no burst will be detected
    delta_threshold = 10^-7;

    % detector offset
    detect_offst = 115;
```

Listing 3.7: Parametrisierung für /t/

In dem Code Snip 3.7 sind die Parameter ersichtlich:

- **Multiplikatoren:** Die Multiplikatoren legen fest, welchen Bruchteil vom detektierten Maximalwert des Signals der Threshold ausmachen soll. Sie bestimmen somit die Empfindlichkeit, wobei kleinere Werte eine geringere Empfindlichkeit bedeuten.
- **Delta Threshold:** Der Delta Threshold bestimmt, welche Steigung der Leistungsverlauf mindestens erreichen muss, damit überhaupt ein Burst detektiert wird. Dieser Parameter ist absolut, das heißt er wird nicht vom Signalverlauf beeinflusst. Wie die Multiplikatoren kann dieser Parameter für jeden Plosiv separat gesetzt werden.
- **Detektor Offset:** Offset, der zur detektierten Burstzeit addiert wird. Dies dient dazu, um dem eher auf Ja/Nein-Entscheid parametrisierten Detektor eine zuverlässigere Detektion des Burstzeitpunktes zu ermöglichen. Dieser Parameter steht in Wechselwirkung mit den Multiplikatoren und sollte nach deren Änderung separat optimiert werden.

Um die Parametrisierung bzw. das Debuggen des Detektors zu erleichtern, besteht die Möglichkeit, einen Plot der Signalverläufe mit der erstellten Hüllkurve und dem detektierten Burstzeitpunkt (siehe Abb. 3.4) erstellen zu lassen. Dazu muss an oberster Stelle im Code des Detektors die Flag `do_plot` gesetzt werden (Listing 3.8.

```
%% set plot option for debugging/parametrizing

% plot option if do_plot = 1, debug plot will be provided
do_plot = 0;
```

Listing 3.8: plot-Flag

3.3.1 Vorgehensweise bei der Parameteroptimierung

Ausgehend vom Material des KielCorpus wurde ein exemplarisches /t/ am Wortanfang mit klar ausgeprägtem Burst während der Entwicklung des Detektors zu Testzwecken herangezogen. Unter Zuhilfenahme des Debug-Plots des Detektors wurden die Multiplikatoren so gewählt, dass eine Detektion des Bursts genau am Übergang zwischen Verschluss und Burst zustandekommt. Die so gefundenen Werte bilden den Ausgangspunkt für die weitere Parametrisierung.

Anschließend wurde, um die Funktion des Detektors unter weniger eng abgesteckten Bedingungen zu überprüfen, ein Versuchlauf mit dem gesamten „Nordwind“-Corpus vorgenommen. Die Ergebnisse im Bereich der True Positives waren bereits durchaus gut, allerdings war der Detektor zu empfindlich und lieferte unbefriedigend viele False Positives.

In folgenden Versuchen stellte sich heraus, dass eine Anpassung der Multiplier zu keiner namhaften Verbesserung führen kann, da jegliche Änderung hauptsächlich einen Tradeoff von True

Positives zu True Negatives bewirkt. Durch manuelle Auswertung von Teilen des Corpus wurde schließlich augenscheinlich, dass eine merkliche Korrelation zwischen Vorhandensein eines Bursts sowie der Magnitude der Ableitung der Leistung besteht, besonders, wenn einzelne Typen von Plosiven (/t/s, /p/s, etc.) spezialisiert und nicht allgemein betrachtet werden. Somit war es logisch und zielführend, einen Threshold für das Minimum der Leistungs-Ableitung einzuführen (Delta Threshold).

Dieser Delta Threshold wurde auf der Basis von Versuchen mit dem gesamten „Nordwind"-Corpus gewählt und so lange angehoben, bis sich eine maximale Verbesserung der True Negatives ohne eintretende Verschlechterung der True Positives einstellte. Um die Performance weiter zu verbessern, wurden anschließend Subsets für alle Plosivtypen einzeln untersucht und die Parameter für jeden Plosivtyp separat fein abgestimmt.

Zuletzt wurde noch der Offset Parameter eingeführt, dessen Notwendigkeit sich aus der Anforderung an den Detektor ableitet, den Burstzeitpunkt nicht genau am Beginn des Bursts zu setzen. Aufgrund der Verwendung einer Hüllkurve liegt der detektierte Zeitpunkt von vornherein zu früh, da die Hüllkurve in den meisten Fällen im Vergleich zum eigentlichen Signal früher anfängt, zu steigen. Die Funktionsweise des Algorithmus in Hinblick auf die Präzision des Ja/Nein-Entscheids wird aber maßgeblich durch die Multiplikatoren beeinflusst. Eine Optimierung der Burstzeitdetektion unabhängig von der Präzision beim Ja/Nein Entscheid ist daher über die Multiplikatoren allein unmöglich. Die notwendige Konsequenz war daher die Einführung eines Offsets.

Wie bei den vorhergehenden Parametern wurde auch hier ein Ausgangspunkt auf Basis der Ergebnisse auf dem gesamten „Nordwind"-Corpus erstellt, indem der Offset händisch auf das Minimum der Abweichung zwischen detektiertem und in der Transkription manuell gesetztem Burstzeitpunkt trainiert wurde. Dieses Verfahren wurde wiederum auf den Subsets für alle Plosivtypen einzeln wiederholt, um das Ergebnis weiter zu verbessern.

Die letzen Endes gewählten Parameter werden in Tabelle 3.1 aufgelistet:

	MP_{power}	MP_{delta}	TH_{delta}	$Offset$
p	0.002	0.0001	0	160
t	0.1	0.03	10^{-7}	115
k	0.1	0.03	10^{-7}	115
b	0.1	0.03	$3 * 10^{-8}$	117
d	0.1	0.8	10^{-8}	110
g	0.1	0.03	0	115

Tabelle 3.1: Auflistung der Parameterwerte

4 Ergebnisse

4.1 Datenset

Als Ausgangsmaterial für die Evaluierung der Performance des Algorithmus wurde ein Subset des Kiel Corpus herangezogen. Es handelt sich dabei um gelesene Sprache, und zwar um den Text „Nordwind und Sonne“(111 Wörter). Dabei wurde dieser Text von insgesamt 16 verschiedenen Sprechern — von neun Männern und sieben Frauen aus verschiedenen Altersgruppen (jedoch alles Erwachsene), vorgetragen.

Versuche wurden sowohl auf dem gesamten Nordwind Set als auch auf Subsets dessen durchgeführt, um die Performance für einzelne Plosive, bei bestimmter Stellung des Plosivs im Wort und für stimmlos bzw. stimmbehaftet artikulierte Plosive offenzulegen.

Die Sets sind im Speziellen[5]:

- **Gesamt:** Dieses Set enthält sämtliche relevanten Plosive aller Sprecher im „Nordwind“-Ordner des KielCorpus.
- **Initial:** Sämtliche Plosive des Gesamtsets, die sich im Anfang des Wortes in dem sie vorkommen befinden.
- **Medial:** Sämtliche Plosive des Gesamtsets, die sich in der Wortmitte befinden.
- **Final:** Sämtliche Plosive des Gesamtsets, die sich am Wortende befinden.

4.2 Auswertungsparameter

4.2.1 Trefferquote

Zur Beurteilung der Trefferquote werden drei Parameter herangezogen, wobei die manuelle Burstdetektion immer als die korrekte angenommen wurde:

[5] Die größe dieser Subsets, d.h. wie oft die jeweiligen Plosive an der jeweiligen Stelle vorkommen ist den Tabellen 4.1 und 4.2 in der eingeklammerten Zahl zu entnehmen.

- $\mathbf{P}_1$ bezeichnet die Wahrscheinlichkeit, dass der Detektor richtig entschieden hat unter der Bedingung, dass ein Burst vorhanden war.
- $\mathbf{P}_2$ bezeichnet die Wahrscheinlichkeit, dass der Detektor richtig entschieden hat unter der Bedingung, dass kein Burst vorhanden war.
- $\mathbf{P}$ bezeichnet die Wahrscheinlichkeit, dass der Detektor richtig entschieden hat ob ein Burst vorhanden war oder nicht

4.2.2 Temporale Genauigkeit

Zur Beurteilung des temporalen Fehlers $\Delta\mathbf{b}$ bei der Burstdetektion wird der Mittelwert der Beträge der jeweiligen Differenz zwischen detektiertem Burstzeitpunk und dem in der Datenbank transkribiertem Zeitpunkt in Samples (Samplefrequenz 16 kHz) gebildet, wenn beide vorhanden. Gleichung 4.1 soll die Berechnung veranschaulichen, wobei n die Anzahl der richtig erkannten Bursts darstellt.

$$\Delta\mathbf{b} = \frac{1}{n}\sum_{i=1}^{n} |t_{i,detektiert} - t_{i,Datenbank}| \tag{4.1}$$

4.3 Auswertung

Die Folgenden Tabellen listen die Gesamtergebnisse sowie die Ergebnisse für einzelne Teile des Sets auf. Die abgebildeten Werte geben die in Kapitel 4.2 angeführten Quotienten an, die Werte in eckigen Klammern beziffern die Häufigkeit des Auftretens im Korpus laut manueller Transkription (**nicht** die Häufigkeit der richtigen Erkennung).

Tabelle 4.1 zeigt die zusammengefassten Ergebnisse von $\mathbf{P}$, $\mathbf{P}_1$, $\mathbf{P}_2$ und $\Delta\mathbf{b}$ (siehe Kapitel 4.2) für den gesamten „Nordwind“ Teil des Kiel Korpus. Die Werte für die drei verschiedenen Trefferquoten sind mit wenigen Ausnahmen immer um die 90 % und höher. Für alle im Nordwind-Teil vorkommenden Plosive beträgt die Wahrscheinlichkeit, dass der Detektor richtig entschieden hat 93 %. Lediglich die Wahrscheinlichkeit $\mathbf{P}_1$, zum einen bei Plosiven am Wortende (82 %) und zum anderen bei stimmhaften, mittig im Wort gelegenen Plosiven (67 %), ist deutlicher niedriger als der Durchschnitt. Die Werte für $\Delta\mathbf{b}$ schwanken im Großen und Ganzen nicht sehr stark, lediglich eine Verschlechterung, d.h. Anstieg der mittleren Abweichung des manuellen zum automatischen detektieren Bursts bei am Wortende gelegenen Fällen lässt sich erkennen.

Die Tabelle 4.2 weist die selbe Struktur, wie Tabelle 4.1 auf, jedoch werden die Werte der Parameter für alle einzelnen Plosive getrennt betrachtet. Deutlich schlechter als in anderen Fällen ist die Trefferquote des Burst-Detektors nur für $\mathbf{P}_1$, sowohl bei wortmedialen `/d/`s (55 %) als auch wortfinalen `/k/`s (69 %). Der Wert für $\Delta\mathbf{b}$ reißt besonders bei wortfinalen `/p/`s nach oben aus (135), jedoch basiert diese Berechnung nur auf einer Stückzahl von 5. Außerdem sind die schlechten Werte für $\mathbf{P}_2$ bei initialen `/g/`s und für $\mathbf{P}_1$ bei medialen `/p/`s aus eben denselben Grund wenig aussagekräftig.

Für alle Plosive auf allen Positionen wurde 161 mal kein Burst detektiert, weil nicht ausreichend Stützstellen für die Berechnung einer Hüllkurve generiert werden konnten (siehe Punkt 3.2.2). Mit der Entscheidung des Detektors, dass hierbei kein Burst vorhanden ist, lag dieser zu 98 % richtig. Zum Δb für ebenfalls alle Plosive auf allen Positionen sei bemerkt, dass die Abweichung in 68 % der Fälle kleiner gleich war als der Durschnitt (43).

Gesamt	**total**	**stimmlos**	**stimmhaft**
P	**0.93** [1579]	0.95 [842]	0.89 [737]
$\mathbf{P}_1$	**0.91** [822]	0.92 [368]	0.89 [454]
$\mathbf{P}_2$	**0.95** [757]	0.98 [474]	0.89 [283]
$\Delta\mathbf{b}$	**43** [744]	49 [339]	38 [405]
Initial			
P	0.92 [528]	0.98 [64]	0.92 [464]
$\mathbf{P}_1$	0.93 [407]	1.00 [16]	0.93 [391]
$\mathbf{P}_2$	0.90 [121]	0.98 [48]	0.85 [73]
$\Delta\mathbf{b}$	39 [379]	25 [16]	39 [363]
Medial			
P	0.92 [801]	0.96 [536]	0.85 [265]
$\mathbf{P}_1$	0.89 [338]	0.95 [275]	**0.67** [63]
$\mathbf{P}_2$	0.94 [463]	0.97 [261]	0.91 [202]
$\Delta\mathbf{b}$	46 [302]	48 [260]	29 [42]
Final			
P	0.94 [250]	0.93 [242]	1.00 [8]
$\mathbf{P}_1$	0.82 [77]	**0.82** [77]	-
$\mathbf{P}_2$	0.99 [173]	0.99 [165]	1.00 [8]
$\Delta\mathbf{b}$	58 [63]	58 [63]	-

Tabelle 4.1: Zusammengefasste Auswertung für den gesamten Nordwind-Teil des Kiel-Corpus [4]. Für die Definition von $\boldsymbol{P}$, $\boldsymbol{P}_1$,$\boldsymbol{P}_2$ und $\Delta\boldsymbol{b}$ siehe Punkt 4.2 .

Gesamt	/**p**/	/**t**/	/**k**/	/**b**/	/**d**/	/**g**/
P	0.96 [50]	0.96 [673]	0.92 [119]	0.92 [112]	0.89 [494]	0.87 [131]
$\mathbf{P}_1$	0.92 [12]	0.92 [263]	0.91 [93]	0.97 [35]	0.89 [351]	0.88 [68]
$\mathbf{P}_2$	0.97 [38]	0.98 [410]	0.92 [26]	0.90 [77]	0.91 [143]	0.86 [63]
$\Delta\mathbf{b}$	76 [11]	50 [243]	44 [85]	50 [34]	37 [311]	35 [60]
Initial						
P	-	0.98 [48]	1.00 [16]	0.94 [64]	0.92 [352]	0.88 [48]
$\mathbf{P}_1$	-	-	1.00 [16]	1.00 [33]	0.93 [313]	0.89 [45]
$\mathbf{P}_2$	-	0.98 [48]	-	0.87 [31]	0.85 [39]	0.67 [3]
$\Delta\mathbf{b}$	-	-	25 [16]	51 [33]	38 [290]	37 [40]
Medial						
P	0.97 [32]	0.96 [424]	0.95 [80]	0.90 [48]	0.82 [134]	0.87 [83]
$\mathbf{P}_1$	1.00 [6]	0.94 [208]	0.95 [61]	0.50 [2]	**0.55** [38]	0.87 [23]
$\mathbf{P}_2$	0.96 [26]	0.98 [216]	0.95 [19]	0.91 [46]	0.93 [96]	0.87 [60]
$\Delta\mathbf{b}$	26 [6]	48 [196]	50 [58]	38 [1]	27 [21]	32 [20]
Final						
P	0.94 [18]	0.96 [201]	0.74 [23]	-	1.00 [8]	-
$\mathbf{P}_1$	0.83 [6]	0.85 [55]	**0.69** [16]	-	-	-
$\mathbf{P}_2$	1.00 [12]	0.99 [146]	0.86 [7]	-	1.00 [8]	-
$\Delta\mathbf{b}$	**135** [5]	54 [47]	39 [11]	-	-	-

Tabelle 4.2: Gesonderte Auswertung für den gesamten Nordwind-Teil des Kiel-Corpus [4]. Für die Definition von $\boldsymbol{P}$, $\boldsymbol{P}_1$,$\boldsymbol{P}_2$ und $\Delta\boldsymbol{b}$ siehe Punkt 4.2 .

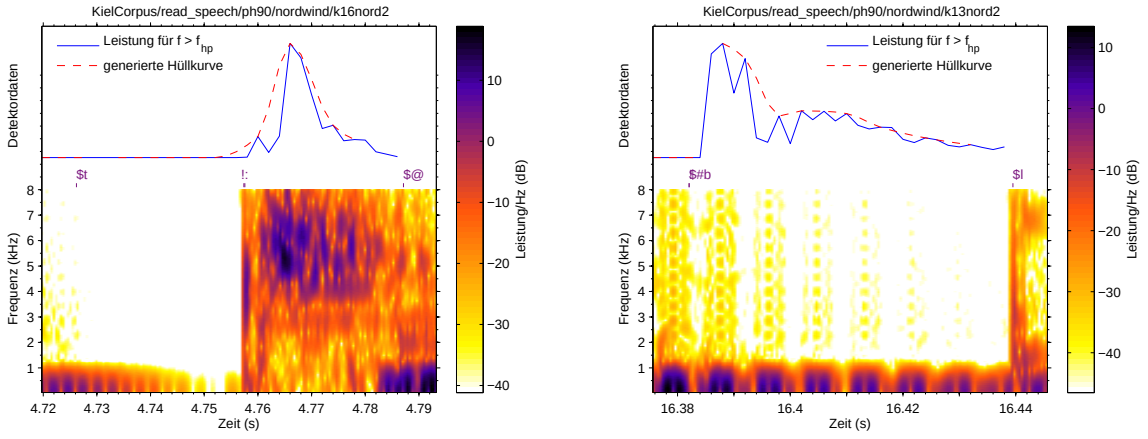
4.4 Analyse anhand repräsentativer Beispiele

Dieses Kapitel untersucht anhand verschiedener repräsentativer Beispiele die Funktionweise des Detektors. Dabei wird aufgezeigt wieso gewisse Signalverarbeitungsstufen im Detektor überhaupt notwendig sind — auch werden Schwachstellen des Detektors angesprochen. Außerdem wird erklärt wieso die unter Kapitiel 4.3 aufgeführten Ergebnisse teilweise nicht mehr verbesserungsfähig sind.

4.4.1 Korrekte Detektion

Abbildung 4.1(a) zeigt das Spektogramm und die Detektordaten eines korrekt manuell und automatisch erkannten, vorhandenen Bursts. Die beiden Labels für die Burstzeitpunkte sind dabei nahezu auf der selben Position. Obowhl der gefensterte Leistungsverlauf $P(f > f_{hp})$ beim Burst nicht monoton, sondern ruckartig ansteigt, wird durch die Hüllkurve gewährleistet, dass der richtige Burstzeitpunkt erkannt wird.

Abbildung 4.1(b) zeigt einen Fall, bei welchen sowohl manuell als auch automatisch, kein Burst gelabelt wurde. Man erkennt bei dem auf den Plosiv `$#b` folgenden Label `$l` zwar einen für einen Burst charakteristischen Spektogrammverlauf, jedoch zählt dieses Ereignis zeitlich bereits nicht mehr zum vorausgehenden Plosiv, weswegen auch kein Label für den Burst gesetzt wird.



(a) (b)

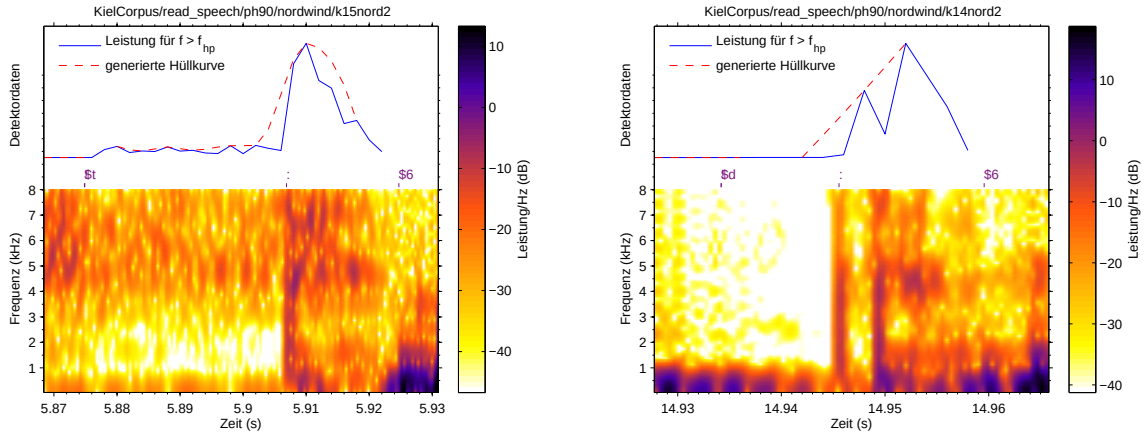
*Abbildung 4.1: Spektogramme und Dekektordetails korrekt detektierter Beispiele (aus [4]). (a) Korrekt manuell und automatisch erkannter, vorhandener Burst. Das Label **!** steht für den automatisch und das Label **:** für den manuell gelabelten Burstzeitpunkt. (b) Plosiv für welchen kein Burst automatisch detektiert und auch keiner manuell gelabelt wurde.*

4.4.2 Fehlerhafte Detektion

Abbildung 4.2(a) zeigt ein Beispiel, wo ein vorhandener Burst vom Detektor nicht erkannt wurde. Aufgrund vorausgehender Friktion wurde in der Entscheidungsstufe des Detektors der Threshold für die Leistung nicht unterschritten und somit kein Burst detektiert (siehe Kapitel 3.2.2).

Die Grafik 4.2(b) zeigt das Spektogramm und die Detektordetails eines wortmedialen `/d/`s. Hierbei wird der Burst vom Detektor nicht erkannt, weil bereits die Generation der Hüllkurve nicht sinngemäß funktioniert, da die Verschlusszeit (i.a.W. die Zeitspanne vom Plosivlabel bis zum Burstlabel) sehr kurz ist. Bei nahe zu allen Fällen in denen ein Burst bei einem medialen

/d/ fälschlicherweise nicht detektiert wurde, passierte dies aufgrund der selbigen Problematik [6].



(a) (b)

*Abbildung 4.2: Spektogramme und Dekektordetails falsch detektierter Beispiele (aus [4]). (a) Plosiv mit vorhanden Burst, jedoch ohne Verschluss, sondern mit Friktion. Automatisch wurde kein Burst erkannt, manuell jedoch schon. Das Label **:** steht für den manuell eingetragenen Burstzeitpunkt. (b) Nicht automatisch detektierter Burst, obwohl einer vorhanden und auch manuell gelabelt ist. Das Label **:** steht für den manuell eingetragenen Burstzeitpunkt.*

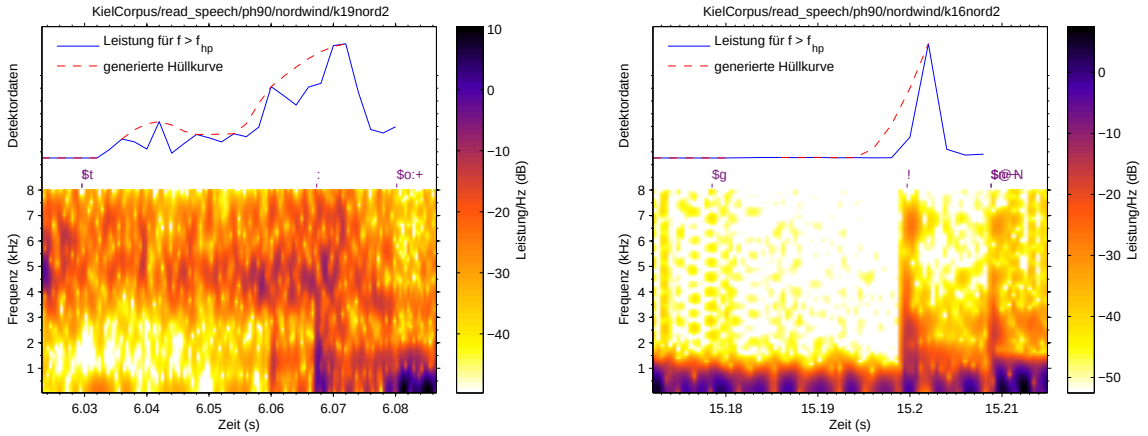
4.4.3 Grenzfälle und Fehler in der Datenbank

Abbildung 4.3(a) zeigt ein Beispiel, wo manuell ein Burst gelabelt wurde, aber es jedoch fraglich ist, ob man in diesem Fall (noch) von einen Burst sprechen kann. Der Detektor hat keinen Burst detektiert — hätte er einen detektiert, wäre der Zeitpunkt aufgrund der Hüllkurve jedoch früher (bei circa 6.056 s) festgelegt worden. Der Threshold für die Leistung wird an dieser Stelle aber nicht unterschritten, da der Verschluss nicht stark genug war, und somit kein Burst festgestellt.

Abbildung 4.3(b) zeit einen Plosiv, bei dem der Burst in der Datenbank nicht gekennzeichnet wurde, obowhl dieser eindeutig vorhanden ist. Der Detektor entscheidet aber richtig und labelt einen Burst.

Da es relativ viele dieser Grenzfälle, sowie einige eindeutige Fehler im Kiel-Corpus gibt, ist es folglich nicht möglich, dass die automatische Detektion in 100 Prozent der Fälle mit der manuellen übereinstimmt. Werte über 90 % der unter Punkt 4.2.1 erklärten Auswertungsparameter sind somit bereits als sehr gut einzustufen.

[6] Mit einer Vergrößerung des Audiosegments für den Detektor, speziell bei mittleren /d/s (nach vorne um 6 ms und nach hinten um 3 ms), konnte hier eine erhebliche Verbesserung der Trefferquote $\mathbf{P}_1$ von 55 auf 85 % erzielt werden (siehe Kapitel 4.3).



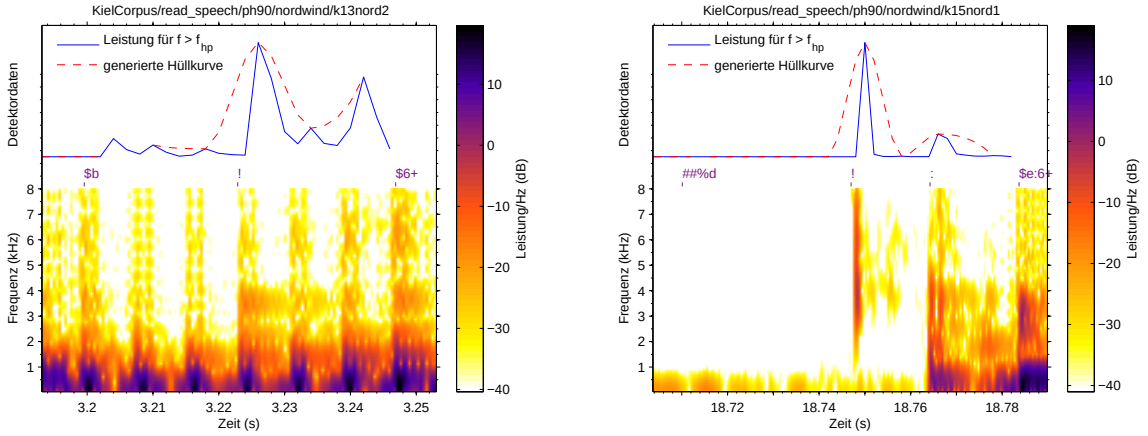
(a) (b)

Abbildung 4.3: Spektogramme und Dekektordetails (aus [4]) für (a) einen Grenzfall, wo es nicht eindeutig ist, ob man von einem Burst sprechen kann oder nicht (das Label **:** *steht für den manuellen Burstzeitpunkt) und für (b) einen Plosiv, bei welchen im Kiel-Corpus der Burst nicht gelabelt wurde, obwohl eindeutig vorhanden (das Label* **!** *steht für den detektieren Burstzeitpunkt).*

4.4.4 Multiple Bursts

Abbildung 4.4(a) zeigt einen Plosiv mit Multiple Bursts, d.h. es finden mehrere kurze, aufeinanderfolgende Bursts in einem Plosiv statt. Hier wurde deswegen manuell auf ein Burstlabel verzichtet, da in der Segmentierungsdatei des Kiel-Corpus maximal ein Burstlabel pro Plosiv zulässig ist. Der Detektor jedoch hat ein Burstlabel beim energetisch stärksten Burst gesetzt.

Abbildung 4.4(a) zeigt einen Plosiv mit einen „doppelten“ Burst. Der Transkribient setzte das Burstlabel auf den letzeren Burst, der Detektor jedoch auf den ersteren — wiederum dem energetisch stärksten. Dieses Beispiel zeigt, wieso es wegen großer Variabilität der Transkribienten bei Multiple Bursts nur bedingt möglich ist die Werte unter Punkt 4.3 für die mittlere Abweichung des Burstzeitpunktes $\mathbf{\Delta b}$ weiter zu verbessern.



(a) (b)

Abbildung 4.4: Spektogramme und Dekektordetails von Plosiven mit Multiple Bursts (aus [4]). (a) Plosiv mit multiple Bursts — manuell wurde hier kein Burstlabel gesetzt. Das Label **!** *steht für den automatisch gelabelten Burstzeitpunkt. (b) Plosiv mit einem „doppelten“ Burst. Das Label* **!** *steht für den automatisch und das Label* **:** *für den manuell gelabelten Burstzeitpunkt.*

5 Conclusio

5.1 Diskussion der Ergebnisse

Die Motivation für die vorliegende Arbeit lag darin, einen Teil eines Tools zur automatischen Generierung einer Transkription eines Sprachcorpus zu implementieren, um damit die zukünftige Erstellung von Sprachcorpora zu erleichtern. Generell ist der Detektor in seiner vorliegenden Form bereits verwendbar unter dem Aspekt zwischen Vorhandensein bzw. Fehlen eines Bursts zu entscheiden. Die erreichten Trefferwahrscheinlichkeiten liegen in weiten Bereichen um oder sogar über 90%. Die Trefferquoten in den den Bereichen mit weit über 90% haben aus der Sicht der Autoren kaum noch Verbesserungspotential. Hier kommt oftmals die Fehlerquote der manuellen Transkription zu tragen, vor allem wenn der Plosiv unklar artikuliert wurde, in anderen Worten wenn ein Grenzfall vorliegt ob nun tatsächlich ein Burst vorhanden ist oder nicht.

Für einige Kombinationen aus Plosivtyp und Stellungen im Wort ergeben sich weniger zufriedenstellende Ergebnisse. Unseren Untersuchungen zufolge liegen die schlechten Ergebnisse hauptsächlich dann vor, wenn die Hüllkurvengeneration augrund des Signalverlaufs ungenau oder fehlerhaft wird. Dies liegt Beispielsweise vor, wenn vor dem Burst lange kein Maximum im gefensterten Leistungsverlauf detektiert wird. Die resultierende Hüllkurve fällt damit viel zu flach vom Burst in Richtung Verschluss ab, wodurch die Qualität der Detektion massiv beeinträchtigt wird. Dieser Umstand wirkt sich auch oft nachteilig auf die zeitliche Detektionsgenauigkeit aus. Ferner basiert die Berechnung des Burstzeitpunktes auf einer relativ groben Fensterung (4 ms lange Fenster mit je 2 ms Überlappung), weshalb eine allzugenaue Detektion des Zeitpunktes a priori nicht möglich ist.

Im Vergleich mit ähnlichen Projekten wie [15] und [7] schneidet der hier vorgestellte Detektor durchaus zufriedenstellend ab. Gegenüber [15] konnte vor allem bei der Präzision der Entscheidung, dass kein Burst vorliegt eine signifikante Verbesserung um 30 bis 45 % erreicht werden, wobei die Übereinstimmung, dass ein Burst vorliegt etwa gleich hoch ist. Beim Detector in [7] muss jedoch berücksichtigt werden, dass spontanes Sprachmaterial herangezogen wurde, was die Detektion erschwert und folgich die Trefferquoten verschlechtert.

5.2 Ausblick

Für die Anwendung des hier entwickelten Detektors sollten in weiterer Folge aussagekräftige Versuche mit spontaner Sprache unternommen werden. Da der Detektor hauptsächlich für die Erstellung des Graz Corpus of Austrian German herangezogen werden soll, ist eine ausgiebige Parametrisierung auf Basis eines Testsets mit entsprechender Sprache unumgänglich.

Optimierungsmöglichkeiten, sowohl im Bereich der Trefferquote sowie der zeitlichen Genauigkeit, existieren aus unserer Sicht aufgrund der mit dem Nordwind-Korpus gewonnenen Erkenntnisse im Bereich der Hüllkurvengenerierung des gefilterten und gefensterten Leistungsverlaufs. Beispielsweise wäre es denkbar nicht nur alle lokalen Maxima im gefensterten Leistungsverlauf, sondern alle Punkte für die die zweite Ableitungs kleiner null oder kleiner als ein zu wählender Threshold ist zu interpolieren. Auch spezialisiertere Algorithmen könnten hier im Vergleich zu dem sehr einfachen zur Anwendung kommenden Verfahren durchaus bessere Ergebnisse erzielen.

Eine signifikant bessere zeitliche Genauigkeit des Burstzeitpunktes — abgesehen von Multiple Bursts (Kapitel 4.4.4) — könnte folgendermaßen erreicht werden: Anstelle des im Kaptitel 3.2.2 beschriebenen Offsets könnte man basierend auf den vorläufigen Wert des Burstzeitpunktes (ohne Offset), der in der Regel zu früh ist, den schlussendlichen Burstzeitpunkt dadurch gewinnen, indem man das ursprüngliche Sprachsignal in einer zweiten Bearbeitungsstufe mit einen feiner gefensterten Leistungsverlauf (d.h. kürzere Fenster bis zu 0.5 ms) analysiert.

Außerdem wäre es empfehlenswert eine gesonderte Behandlung für multiple Bursts zu entwickeln, bzw. diese überhaupt als solche zu detektieren.

Literaturverzeichnis

[1] [Online]. Available: http://de.wikipedia.org/wiki/Plosiv

[2] K. Johnson, *Acoustic and Auditory Phonetics*. Wiley Blackwell, 2012.

[3] [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[4] [Online]. Available: http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html

[5] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1999.

[6] H. Mitterer and M. Ernestus, "Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch," *Journal of Phonetics*, no. 34, pp. 73–103, 2006.

[7] B. Schuppler, "Automatic analysis of acoustic reduction in spontaneous speech," Ph.D. dissertation, Radboud Universiteit Nijmegen, 2011.

[8] A. Juneja and C. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 2, no. 123, pp. 1154–1168, 2007.

[9] A. Kapoor and J. B. Allen, "Peceptual effects of plosive feature modification," *J. Acoust. Soc. Am.*, vol. 1, no. 131, pp. 478–491, 2012.

[10] O. E. Scharenborg, "Narrowing the gap between automatic and human word recognition," Ph.D. dissertation, Radboud Universiteit Nijmegen, 2005.

[11] L. Chin-Hui, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, J. Biing-Hwang, and L. R. Rabiner, "An Overview on Automatic Speech Attribute Transcription (ASAT)," in *INTERSPEECH 2007*, pp. 1825–1828.

[12] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.*, vol. 5, no. 100, pp. 3417–3430, 1996.

[13] L. Jung-Won, C. Jeung-Yoon, and K. Hong-Goo, "Classification of stop place in consonant-vowel contexts using feature extrapolation of acoustic-phonetic features in telephone speech," *J. Acoust. Soc. Am.*, vol. 2, no. 131, pp. 1536–1546, 2012.

[14] C. Cucchiarini and H. Strik, "Automatic Phonetic Transcription: An overview," in *15th ICPhS Barcelona*, pp. 347–350.

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