10.7 Akustische Muster koartikulierter Rede

Wie in den Abschnitten 7.4 und 9.4 bereits erwähnt, kommt es bei normalen Redeäußerungen zu sogenannter Koartikulation, also bzgl. der Einzellaute zu Assimilations- und Reduktionsphänomenen – selten auch zu Elaborationsphänomenen. Zur Demonstration der relativ gravierenden Koartikulationseffekte, die bei zusammenhängender Rede eigentlich immer auftreten, findet sich in Abbildung 10.17 eine Darstellung (Sonagramm, Oszillogramm und annotierte Segmentation) des Satzes „Die Rinder sind noch auf der Weide“, gesprochen von einem männlichen Sprecher mittleren Alters. Dabei ist zu beachten, dass dieser Satz sehr deutlich gesprochen wurde und als „der Hochlautung entsprechend“ wahrgenommen wird!

Wie dem Vergleich der engen phonetischen Transkription (oberste Annotationsebene) und der „Standardlautung“ (darunter) in Abbildung 10.17 zu entnehmen ist, wird die Mehrzahl der Laute durch den Einfluss der Nachbarlaute in ihrer Realisierung im kontinuierlichen Sprechbewegungsablauf verändert (vgl. auch Kap. 7.4). Neben den wechselseitigen Assimilationen finden auch einfache Reduktionen im Sinne des Nichterreichens der idealisierten Zielpositionen statt. Dieser Effekt wird mit zunehmender Sprechgeschwindigkeit immer ausgeprägter.

Im Detail erfolgt die erste Assimilation bereits im ersten Laut. Das initiale /d/ wird durch den Einfluss des nachfolgenden palatalen Vokals /i/ weiter vorne realiert als z.B. vor einem /a/ oder vor einem /u/. Das im zweiten Laut zu erwartende /i/ wird in seinem Anfangsteil durch das zentralere und dadurch weniger aufwändig zu realisierende [ɪ] ersetzt, das in seinem zeitlichen Verlauf sogar noch weiter zentralisiert wird, um die Zungenspitze in Position für das /r/ zu bringen. Der nächste eindeutig auf Koartikulation zurückzuführende Effekt ist die Elision des /d/ in <Rinder>. Das vorangehende /n/ wird homorgan gebildet; der abrupte Anstieg der Energie vom Nasal zum folgenden Vokal hin genügt der Wahrnehmung als Ersatz für einen „echten“ Burst. Da der nicht gebraucht wird, erübrigt sich auch der Verschluss des Nasenraumes durch das Velum. Bei einer Wahrnehmung im Gesamtkontext meint man hier ein /d/ zu hören, obwohl es nicht da ist; dies wird deutlich, wenn man nur den Nasal und den darauf folgenden Vokal anhört. Das der Standardlautung entsprechende /ɐ/ am Ende von <Rinder> wird durch ein [ə] ersetzt. Beide sind sich sehr ähnlich, aber das [ə] ist im Kontext eines /n/ und eines /z/ mit weniger artikulatorischem Aufwand zu erreichen. Das /t/ in <sind> wird wiederum elidiert. Der Grund ist der selbe wie bei der vorangehenden Elision in dem

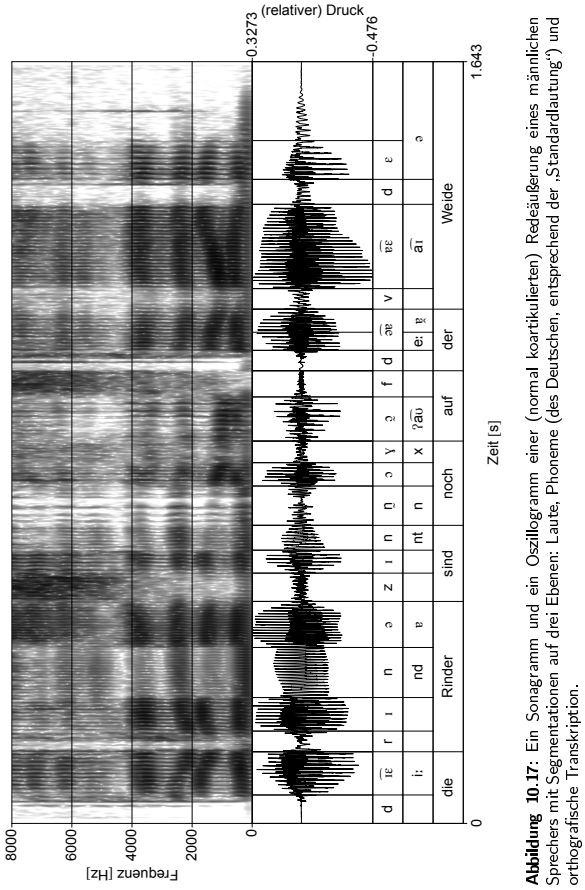

Abbildung 10.17: Ein Sonagramm und ein Oszillogramm einer (normal koartikulierten) Redeäußerung eines männlichen Sprechers mit Segmentationen auf drei Ebenen: Laute, Phoneme (des Deutschen, entsprechend der „Standardlautung") und orthografische Transkription.

Wort <Rinder>. Stattdessen erhält die zweite Hälfte des (nun geminatischen) /n/ eine Elaboration in Form einer Laryngalisierung bzw. Glottalisierung [n̰]. Solche unregelmäßigen und langsameren Glottisschläge finden sich regelmäßig als Ersatz eines Glottal Stops, der wiederum der Standardlautung entsprechend am Anfang von Wörtern (als „harter Stimmeinsatz") gesprochen wird, die mit einem Vokal beginnen. Der eigentlich stimmlose Frikativ /x/ wird aufgrund seiner vokalischen und damit beiderseitig stimmhaften Einbettung durch sein stimmhaftes Pendant [ɣ] ersetzt – obwohl dieser Laut kein Phonem des Deutschen ist. Der Diphthong /a͡ʊ̯/ und der eigentlich zu erwartende Glottal Stop davor werden zu einem homogenen [ɔ̯]-Laut assimiliert: Das [ɔ] befindet sich im Vokalraum mittig zwischen [a] und [ʊ], den beiden Endpunkten des zu realisierenden Diphthongs. Die Laryngalisierung des [ɔ] ersetzt den Glottal Stop. Die restlichen Reduktionen in diesem Satz sind als Zentrierungen (innerhalb des Vokalraumes) der eigentlichen Phoneme interpretierbar. Solche Zentrierungen vermindern den artikulatorischen Aufwand und können ohne Aussprachedeutlichkeitsverlust vorgenommen werden, sofern die phonologische Unterscheidbarkeit aufgrund des Kontextes erhalten wird.

10.8 Akustische Ausprägungen von Stimmqualität

Neben diversen Klassifizierungsversuchen verschiedener Stimmgebungsarten (vgl. z.B. die Phonationsarten in Abb. 6.12 und ihre Erhebung über EGG-Messungen) kann Stimmqualität auch auf metrischen Skalen gemessen werden. Allerdings besteht keine „eins-zu-eins"-Beziehung solcher Maße und perzeptiven Qualitäten, wie beispielsweise zwischen der F_0 und der Stimmtonlage. Zu solchen Messungen eignet sich auch das akustische Sprachsignal, da in ihm das Quellsignal enthalten ist, das nach herrschender Meinung ganz oder wenigstens größtenteils „die Stimme" ausmacht. Und nicht zuletzt nehmen auch Menschen zunächst das an den Mundlippen abgestrahlte Sprachsignal wahr, anhand dessen sie dann Einschätzungen der Stimmqualität vornehmen. Technisch gesehen wäre es optimal, das Quellsignal allein, separiert von der Filterfunktion des Ansatzrohres, zu betrachten. Eine solche akustische Quelle-Filter-Trennung ist machbar – erzeugt aber ihrerseits zumeist Fehler ähnlicher Art und Größenordnung wie die zu messenden stimmlichen Unterschiede. Die meisten stimm-qualitativen Maße unterlaufen gewissermaßen dieses Problem durch ihre Schlichtheit – wobei dafür in Kauf zu nehmen ist, dass in diesen Maßen eben Quell- und Filtereigen-

schaften gleichzeitig und untrennbar verwoben abgebildet sind. Zum Zweck der Vergleichbarkeit erhobener Daten sollte also die Filterfunktion des Ansatzrohres möglichst konstant gehalten werden. Ein in Tonhöhe und Lautstärke für mehrere Sekunden möglichst konstant gehaltenes /a/ ist deshalb die Standardäußerung zur medizinisch-akustischen Stimmqualitätsuntersuchung. Die wichtigsten Signalparameter, die stimmqualitative Unterschiede abbilden – Perturbationsmaße, Maße der spektralen (Geräusch-) Energieverteilung und Modulationsmaße –, werden nachfolgend kurz erläutert. Sie korrelieren mit Empfindungsgrößen wie Knarrigkeit, Rauigkeit, Heiserkeit, Behauchtheit und Zittrigkeit und im weiteren Sinne mit Krankheit und Alter (vgl. Kap. 16).

10.8.1 Perturbationsmaße: Jitter und Shimmer

Perturbationsmaße korrelieren mit der Empfindung einer Stimme als mehr oder weniger „rau“ oder auch „heiser“. Amplitudenperturbationen (Shimmer und Derivate davon) sind perzeptiv von Grundfrequenzperturbationen (Jitter und dessen Derivaten) schwer zu unterscheiden (vgl. z.B. Kreiman & Gerratt, 2003). Dies betrifft noch vermehrt ihre begriffliche Denotation.

Perturbationsmaße erfassen die mittleren absoluten Abweichungen der Periodendauern bzw. der Amplituden von einer Grundfrequenzperiode zur nächsten (vgl. Abb. 10.18). Damit sind sie einfachen statistischen Streuungsmaßen (also der Standardabweichung und mehr noch der AD-Streuung, also der „mittleren absoluten Abweichung“, vgl. z.B. Bortz & Schuster, 2010, S. 31) sehr ähnlich. Der entscheidende Unterschied besteht in der Einbeziehung der geordneten seriellen Abfolge der Perioden bei den Perturbationsmaßen:

$$absoluter\ Jitter\ [\mu s] = \frac{\sum_{i=1}^{n-1} |T_i - T_{i+1}|}{n-1} \tag{10.4}$$

Da die Abweichungen von einer Periode zur nächsten natürlich größer sind, wenn die Periodendauern selbst größer sind, empfiehlt sich die Relativierung der mittleren Abweichung an der mittleren Periodendauer $\overline{T}$:

$$relativer\ Jitter\ [\%] = \frac{absoluter\ Jitter}{\overline{T}} \tag{10.5}$$

Der erste als Stimmqualitätsmaß veröffentlichte Perturbationsparameter war die „Relative Average Perturbation“ (RAP, Koike, 1973). Im Unterschied zum relativen Jitter wird hier jeweils die Dauer von drei aufeinander

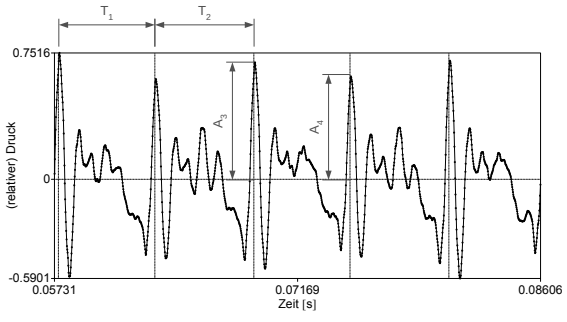

Abbildung 10.18: Berechnung von Stimmperturbationsmaßen: 5 (Quasi-) Perioden eines von einer Frau gehaltenen [a]s mit einer mittleren Periodendauer ($\overline{T}$) von ca. 5,75 ms und (akustisch berechneten) Glottisverschlusszeitpunkten (die mit senkrechten gestrichelten Linien markiert sind). Relativer Jitter = 2,653%; relativer Shimmer = 6,637%.

folgenden Perioden gemittelt, bevor davon die absolute Differenz zur mittleren der drei bestimmt wird. Durch eine solche Mittlung der Dauer von drei (oder mehr) Perioden erreicht man, dass starke kurzfristige Änderungen der Dauer (z.B. von einer Periode zur nächsten) nicht so ins Gewicht fallen wie längerfristige Änderungen. Es erfolgt eine „Glättung der Perturbationskontur“.

Entscheidend für eine verlässliche Perturbationsmessung ist eine durchgängig richtige Grundfrequenzbestimmung, die ihrerseits verwendet wird, um die einzelnen Perioden voneinander zu trennen, also die Glottisverschlusszeitpunkte (vgl. Abb. 10.18) zu berechnen. Gerade bei stark perturbierten Äußerungen kommt es deshalb zu dem misslichen Umstand, dass die Grundfrequenz nur unzureichend oder gar nicht bestimmt werden kann und deshalb die Perturbationsmaße größere Fehler aufweisen oder sogar gar nicht extrahiert werden können.

Die Amplituden-Perturbationen (Shimmer-Maße) sind ganz entsprechend definiert, unter der Maßgabe, dass nicht die Periodendauern T_i die Grundelemente der Berechnung sind, sondern die Amplituden A_i (vgl.

Abb. 10.18). Hierbei ist allerdings zu beachten, dass verschiedene Implementationen nicht nur Unterschiede in der Bestimmung einzelner Perioden machen, sondern dass auch „Amplitude“ verschieden interpretiert wird. Die gängigsten Varianten sind: (1) die maximale Elongation der Zeitfunktion innerhalb einer Periode (wie in Abb. 10.18), oder (2) die minimale oder auch (3) das Integral über eine Periode.

10.8.2 Harmonics-to-Noise Ratio

Die vorherrschende Wirkung von Geräuschanteilen – insbesondere von höherfrequenten Geräuschanteilen – im Stimmklang ist „Heiserkeit“ oder auch „Behauchung“. Die allen Implementierungen gemeinsame Grundidee für ein entsprechendes akustisches Maß, den „Harmonics-to-Noise Ratio“ (HNR), ist wiederum relativ einfach:

$$HNR = \frac{harmonische\ Energie}{turbulente\ Energie} = \frac{1}{NHR} \tag{10.6}$$

Schwieriger zu beantworten ist die Frage, wie die harmonischen Energieanteile von den turbulenten zu unterscheiden sind. In Praat (vgl. Boersma, 1993) beispielsweise ist auch die Implementierung möglichst einfach gehalten. Dort wird der HNR abgeleitet vom Ausmaß derselben Autokorrelation r_{auto} des (Sprach-) Signals, die auch zur Abschätzung der Grundfrequenz berechnet wird:

$$HNR_{Praat} = \frac{r_{auto}}{1 - r_{auto}} \tag{10.7}$$

Je ähnlicher sich also die Signalfrequenzperioden im analysierten Ausschnitt sind, umso höher fällt der HNR aus. Harmonische und turbulente Anteile aller Frequenzen fließen gleichermaßen ein. Jedoch operiert die Autokorrelation ausschließlich im Zeitbereich. Der Autokorrelationskoeffizient fällt regelmäßig höher aus, wenn im Zeitsignal (pro Periode) weniger Elongations-Richtungs-Änderungen (vereinfacht Nulldurchgänge) stattfinden. Ein solches Auf und Ab der Intensität innerhalb einer Grundfrequenzperiode hat aber zunächst nichts mit dem Anteil turbulenter Energie zu tun, sondern hängt von der Filterfunktion des Ansatzrohres ab und der Phasenlage der einzelnen Teilschwingungen – wobei sich letztere gewöhnlich gänzlich der Wahrnehmung entzieht. Eine leicht perturbierte Äußerung, die alleine aufgrund der Phasenlage ihrer harmonischen Frequenzkomponenten mehr Nulldurchgänge aufweist (auditiv also unverändert ist), korreliert schlechter

mit sich selbst und wird deshalb von Praat mit einem geringeren HNR bewertet. Konkret führt das dazu, dass HNR-Werte in Praat von [a]-Lauten, die generell mehr Nulldurchgänge aufweisen, mitunter 90-mal(!) kleiner[4] sind als die von [u]-Lauten desselben gesunden Sprechers (vgl. Boersma & Weenink, 2016, Manualeintrag „Harmonicity“). Zudem steigt der Autokorrelationskoeffizient regelmäßig an den Lautübergängen aufgrund der Änderung der Lautqualität an. Solche Unterschiede in einem „Geräuschmaß“, die nicht auf Unterschiede im Geräuschanteil zurückzuführen sind, sind schwer vermittelbar und führen dazu, dass Praats HNR nur bei sehr kontrollierten Äußerungen und für sehr kurze Zeitintervalle (innerhalb von Segmentgrenzen) als Stimmwirkungsmaß einsetzbar ist.

Ein ambitionierterer Ansatz, die hochfrequente turbulente Energie in einer Äußerung zur niederfrequenten harmonischen Energie ins Verhältnis zu setzen, wurde von Klasmeyer (1999) vorgestellt. Sie adressiert die Probleme der Autokorrelations-Methode von Boersma: (1) die Filterung des „die Stimme“ ausmachenden Primärschalls durch das Ansatzrohr und (2) die Problematik unterschiedlicher Phasenlagen der harmonischen Teilkomponenten. Zur Lösung des ersteren wird zunächst eine inverse Filterung am Sprachsignal berechnet, um das Quellsignal nachzubilden (vgl. oben). Dazu bedarf es allerdings einer Modellierung (der Eigenresonanzen) des Ansatzrohres. Und wie bei jeder Modellierung entspricht das nicht exakt der Realität, weshalb erste Fehler entstehen. Das Problem der unterschiedlichen Phasenlagen wird durch die Verwendung des komplexwertigen Betrags der Amplituden der Teilschwingungen im Rahmen einer Hilbert-Transformation gelöst. Die daraus akkumulierten, zeitlich sehr präzisen Intensitätsverläufe (Hilberteinhüllenden) können für beliebige Frequenzbänder berechnet werden. Eine Korrelation der Hilberteinhüllenden eines tieffrequenten Bandes, von dem anzunehmen ist, dass nahezu all seine Energie harmonisch ist, mit einem höherfrequenten Band erlaubt die Schätzung des HNR. „In der Praxis stellte sich jedoch heraus, dass die Korrelationen der einzelnen Bänder auch bei rein harmonischen Signalen nie hoch wurde“ – was größtenteils an den Artefakten der inversen Filterung liegen dürfte (vgl. Klasmeyer, 1999, S. 110).

Auch im Multi-Dimensional Voice Program der Firma Kay (MDVP, KAY, 1993) wird diese Hilbert-Transformation wohl bei der Berechnung des NHR angewandt. Da es sich um proprietäre Software handelt, ist jedoch unklar, wie die Trennung von periodischen und Geräusch-Anteilen genau vorgenommen wird. An belastbarer Information bleibt, dass

[4] 20dB geringer

der NHR unter MDVP als „inharmonische Energie im Frequenzbereich 1500-4500 Hz zu harmonischer im Bereich 70-4500 Hz“ definiert ist. Davon abgeleitete weitere spektrale Energieverhältnisse aus dieser Software sind der „Voice Turbulence Index“ (VTI) und der „Soft Phonation Index“ (SPI), ein Verhältnismaß zweier harmonischer Energiebänder, das zum Ziel hat, die Härte der Phonation zu denotieren. Kritik an diesem System formuliert Paul Boersma (2008): "The MDVP algorithm is so imprecise that it never yields harmonics-to-noise ratios that are better than 10 dB.“ Eine quell-offene Implementierung der Grundidee hinter diesen drei MDVP-Energieverteilungsmaßen als Praat-Skript findet sich bei Brückl (2011).

10.8.3 Modulationen – Stimm-Tremor

Die vorherrschende Wirkung von Modulationen der Periodendauer oder der Amplitude ist „Zittrigkeit“.

Generell bezeichnet man periodische Schwankungen der Grundfrequenz oder der Amplitude, die unterhalb der Grundfrequenz liegen, als Stimm-Modulationen. Im Gegensatz zu den oben erläuterten Perturbationen handelt es sich also ausschließlich um regelmäßig wiederkehrende Abweichungen der Periodendauer oder der Amplitude. Als Stimm-Tremor wird nur eine Untergruppe der Modulationen bezeichnet, nämlich die unwillkürlichen Modulationen mit einer Frequenz von ca. 1,5 bis 15 Hz.

(Stimm-) Tremor kann als Folge einer vermehrten Verzögerung in neuronalen Regelungsprozessen (vgl. Wiener, 1958, in Kap. 2) zur Kontrolle der Phonation auftreten. Zu solchen Verzögerungen kommt es u.a. aufgrund eines alters- (vgl. Kap. 16) oder auch krankheitsbedingten Mangels an Neurotransmittern.

Brückl (2012) stellt einen als Praat-Skript implementierten Algorithmus vor, mit dem Stimm-Tremor-Maße aus akustischen Aufzeichnungen gehaltener Vokal-Äußerungen extrahiert werden können. Die Prinzipien/Besonderheiten dieser Methode sind die Begradigung natürlicher Deklinationen und die anschließende Zyklizitätsanalyse[5] von Amplituden- und Grundfrequenz-*Konturen*.

In Abbildung 10.19 sind diese Berechnungsschritte (für Amplitudentremor) dargestellt: Aufbauend auf einer Grundfrequenzanalyse und der daraus ableitbaren Trennung einzelner Glottisperioden wird pro Periode eine Amplitude bestimmt (Abb. 10.19, oben). Da die Perioden unterschiedliche

[5] „Zyklizität“ beschreibt bei Konturen das Pendant zur „Periodizität“ bei Oszillogrammen.

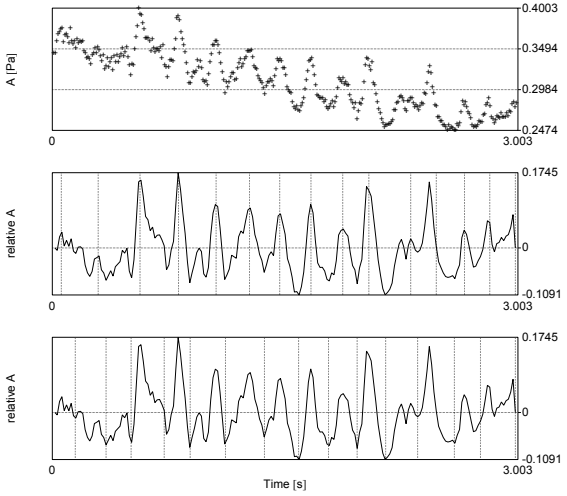

Abbildung 10.19: Schritte einer Amplitudentremoranalyse eines 3 Sekunden dauernden quasi-stationären Teils eines gehaltenen /a/s eines 74-jährigen männlichen Sprechers, der an Parkinson erkrankt ist: Der Graph oben zeigt die Amplituden(-pro-Periode)-Kontur, der mittlere die im festen Zeitschritt abgetastete, normalisierte und de-deklinierte Kontur inklusive der Zeitpunkte gefundener Maxima. Das untere Bild zeigt dieselbe Kontur mit den Minima. ATrF=4,64Hz, ATrI=6,462%.

Dauern haben können, muss die Amplitudenkontur mit einer festen Frequenz erneut abgetastet werden. Die (lineare) Deklination der Kontur, die gewöhnlich bei natürlichen Äußerungen vorzufinden ist, wird entfernt. Nun erfolgt die Bestimmung der Zyklizität der Kontur, die nötig ist, um die (Extrema der) Konturperioden zu finden (Abb. 10.19, Mitte und unten). Der Amplitudentremor-Intensitätsindex (ATrI) ist ein Mittelwert der Be-

träge der Ordinaten dieser Kontur zu den (mithilfe der Tremorfrequenz) geschätzten Zeitpunkten von Kontur-Extrema.

Die daraus abzuleitenden wichtigsten Maße zur Beschreibung von Stimm-Tremor sind die (1) Tremor-Frequenzen (FTrF und ATrF), also die tiefen Frequenzen mit denen sich bestimmte Konturabschnitte wiederholen und (2) die Tremor-Intensitäten (FtrI und ATrI), welche das Ausmaß der Elongationen der Konturen erfassen. Die Tremorfrequenzen sind relativ unkorreliert mit der Stimmwirkung „Zittrigkeit"; ebenso mit dem Sprecheralter oder einer neuronalen Krankheit. In den Intensitätsindizes hingegen wird viel vom Konstrukt „Zittrigkeit" erfasst, womit sich diese Maße als Indikator für das Alter des Sprechers eignen oder beispielsweise auch zur (Früh-) Diagnostik von Parkinson. Aber auch das Ausmaß der Periodizität der Konturen (die Tremor-Zyklizitätsindizes) ist für die Wahrnehmung von „Zittrigkeit" relevant.