OPTIMIERUNG EINER TRAININGSBASIERTEN PROSODIEGENERIERUNG FÜR SPRACHSYNTHESE

Oliver Jokisch, Michael Hofmann

Institut für Akustik und Sprachkommunikation, TU Dresden
oliver.jokisch@ias.et.tu-dresden.de

Abstract: Ausgehend vom trainingsbasierten, silbenorientierten *Integrated Model of German Prosody* (IGM, Mixdorff und Jokisch, 2003) diskutiert der Beitrag Ansätze und Ergebnisse der weiteren Optimierung. IGM schätzt in einem Schritt je Silbe 8 Modellparameter zur Intonations-, Dauer- sowie Intensitätssteuerung und nutzt dazu ein multi-layer feed-forward neural network (MFN) mit einem Eingangsvektor aus 24 linguistischen und phonetischen Merkmalen. Der Beitrag untersucht Ansätze zur Erweiterung der Trainingsdaten, zur evolutionären Strukturoptimierung des MFN sowie zur Optimierung einzelner, zu schätzender Modellausgabeparameter.

1 Einleitung

Prosodiegenerierung (Akzentuierung, Intonation, zeitliche Strukturierung) hat einen großen Einfluss auf Verständlichkeit sowie wahrgenommene Natürlichkeit von Sprache und stellt eine Schwachstelle in Text-to-Speech-Systemen dar. Trainingsbasierte Prosodiemodelle, z. B. auf Basis neuronaler Netze, sind ein flexibles Mittel, um verschiedene Sprachen, Sprecher bzw. Sprechstile abzubilden und werden seit etwa einer Dekade näher untersucht. Nutzer kritisieren nach wie vor z. B. Monotonieeffekte, Betonungsfehler oder Signalartefakte. Außerdem steigen Rechenkomplexität und Speicheranforderungen im Vergleich zur konventionellen, regelbasierten Modellierung.

Ausgehend vom trainingsbasierten, silbenorientierten *Integrated Model of German Prosody* (IGM, Mixdorff und Jokisch, 2003 [3]) diskutiert der Beitrag Ansätze und Ergebnisse der weiteren Optimierung. IGM schätzt in einem Schritt je Silbe 8 Modellparameter zur Intonations-, Dauer- sowie Intensitätssteuerung und nutzt dazu ein multi-layer feed-forward neural network (MFN) mit einem Eingangsvektor aus 24 linguistischen und phonetischen Merkmalen. Folgende Aspekte einer Optimierung der Prosodiemodellierung werden dabei berücksichtigt:

- Weitere Qualifizierung der Eingangsmerkmale, z. B. durch semantische Information.
- Evolutionäre Strukturoptimierung des neuronalen Netzes.
- Genauere Analyse der zu schätzenden Modellausgabeparameter, z. B. die Untersuchung typischer Intensitätsverläufe für das Deutsche.

Diese Maßnahmen führen u. a. zu einer Aufwandsreduktion der Netzstruktur bei gleich bleibender Vorhersage-Performanz bezüglich der Modellausgabeparameter.

2 Prosodiemodellierung mit dem IGM-Modell

2.1 Trainingsdaten

Die analysierten Daten sind Teil eines deutschen Nachrichtenkorpus, welches am IMS der Universität Stuttgart generiert wurde (48 Minuten, männlicher Sprechers, Deutschlandfunk, zwei Aufnahmetage). Die Datenbank enthält 356 Sätze einschließlich 5.726 Wörtern, 13.151 Silben bzw. 29.362 Phonemen. Die Nachrichten wurden teils im Abstand von 30 min. wie-

derholt: Abgesehen von Wortersetzungen gibt es identische Phrasen in der Datenbasis, welche ebenfalls untersucht wurden. Der Sprechstil ist konsistent. Es sind keine spontan-sprachlichen Äußerungen enthalten, so dass dieses Korpus für das Zielmodell geeignet ist.

2.2 Algorithmus

Mixdorff und Jokisch entwickelten ein integriertes Modell der deutschen Prosodie (IGM), welches die Wechselwirkungen zwischen melodischen und rhythmischen Eigenschaften der Sprache berücksichtigt. Die prosodischen Parameter Silbendauer, Grundfrequenz f0 (bzw. die entsprechenden Fujisaki-Steuerparameter), Pausendauer sowie Intensität der Silbe werden in einem Schritt trainiert, wobei die Silbe als rhythmische Basiseinheit betrachtet wird.

Das IGM schätzt für akzentuierte Silben die Fujisaki-Modellparameter: *accent amplitude Aa, onset time T1* und *offset time T2* bzw. deren relative Werte. Für die erste Silbe einer prosodischen Phrase werden zusätzlich *magnitude Ap* und *onset time T0* berechnet. Die sprecherabhängige *base frequency Fb* und die Zeitkonstanten *alpha* sowie *beta* werden als konstant betrachtet. Die Lautdauern werden entsprechend den Lauteigenschaften aus den übergeordneten Silbendauern berechnet (vgl. auch Elastizitätsmodell von Campbell [1]).

Um potentielle Wechselwirkungen zwischen Intonation und Rhythmus abzubilden, werden die genannten prosodischen Parameter mit einem einzigen multi-layer feed-forward neural network (MFN, vgl. Abb. 1) geschätzt, wobei 24 linguistisch-phonetische Eingangsmerkmale genutzt werden. In früheren Studien wurde nachgewiesen, das MFN prinzipiell in der Lage sind, sowohl prosodische Verläufe direkt [2] als auch Steuerparameter für ein Modell [4] (z. B. Fujisaki-Parameter) zu schätzen.

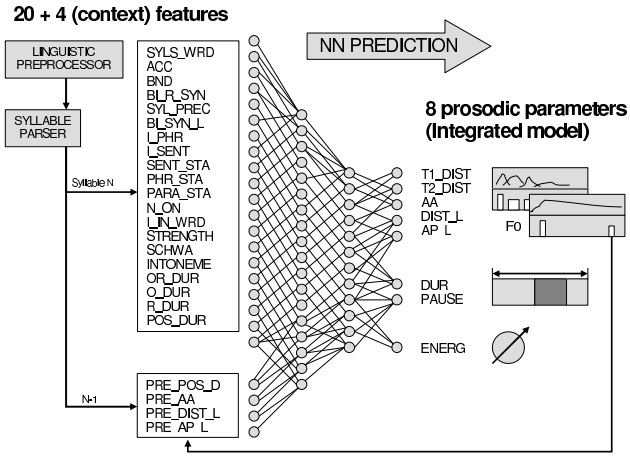

Abbildung 1 – MFN-Struktur (neuronaler Netzwerkkern des IGM).

3 Optimierung des IGM-Modells

Das Modell IGM nutzt einen ausgewählten Satz an linguistischen sowie phonetischen Eingangsmerkmalen (Syntax, Phrasierung, Akzentuierung, Lautklasse, etc.) und wurde mit Resynthese- bzw. Synthese-Stimuli getestet [3]. Die Performanz ist für bestimmte Syntheseapplikationen hinreichend. Die Autoren identifizierten folgende Ansatzpunkte für eine Optimierung des IGM:

- Berücksichtigung zusätzlicher Eingangsinformation: Das Modell berücksichtigt bisher z. B. keine Elemente von Bedeutung, *semantic focus* oder abgeleitete Merkmale.
- Strukturoptimierung des Neuronalen Netzwerkes: Die neuronale Struktur des Netzwerkkerns des IGM wurde entsprechend den Erfahrungen mit ähnlichen Mustererkennungs- bzw. Prediktionsaufgaben implementiert. Es existiert kein geschlossenes Re-

gelsystem, um die Anzahl von verdeckten Schichten bzw. Neuronen und andere Topologieparameter zu konfigurieren. Ein evolutionärer Algorithmus soll Redundanzen in der Topologie reduzieren und ggf. die Signifikanz verbleibender Neuronen, Eingangsmerkmale, etc. in weiteren Trainingszyklen erhöhen.

- Verringerung der Rechenkomplexität bzw. des Speicherverbrauchs: IGM kann derzeit nicht in einem Embedded Text-to-Speech (TTS-)System implementiert werden.

4 Zusätzliche Qualifizierung der IGM-Eingangsmerkmale

Die semantische Erweiterung der Eingangsmerkmale des IGM ist in der Praxis kompliziert, da die entsprechenden Informationen nicht nur in der Trainingsdatenbasis zuverlässig markiert werden, sondern auch während der Kannphase im TTS-System vorliegen müssen. Die zusätzliche Qualifizierung der Eingangsmerkmale entspricht deshalb einem iterativen Prozess des systematischen Testens.

4.1 Markierung von extremen Werten

Es wird getestet, welchen Einfluss die automatische Markierung von extremen Werten eines bestimmten Ausgabeparameters für das Netztraining hat. Dazu werden unterschiedliche Durchläufe durchgeführt, in welchen die Menge der zu markierenden Datensätze variiert wird. Für jeden Ausgabeparameter wird ein separates Netz trainiert. Dabei werden je Parameter vier Konfigurationen untersucht:

- Verwendung des ursprünglichen MFN ohne Änderungen (MFN),
- Eliminierung aller Datensätze aus Trainings- und Testmenge, bei denen der Parameter außerhalb eines Vielfachen n der Standardabweichung S liegt (Weglassen > |nS|),
- Ein zusätzlicher Eingang, ob der Parameter außerhalb eines bestimmten Vielfachen n der Standardabweichung liegt (ein Eingang > |nS|),
- Zwei zusätzliche Eingänge, ob der Parameter positiv oder negativ außerhalb eines bestimmten Vielfachen der Standardabweichung liegt (zwei Eingänge, > +nS / < -nS).

	Anzahl		
Parameter	Tag 2	Input > |3S|	
AA	2600	29	1.1%
T1_DI	2569	14	1.3%
T2_DI	2587	20	0.8%
AP_L	891	5	0.6%
DIST_	891	5	0.6%
PAUSE	493	8	1.6%
DUR	11368	102	0.9%
ENERG	11368	101	0.9%

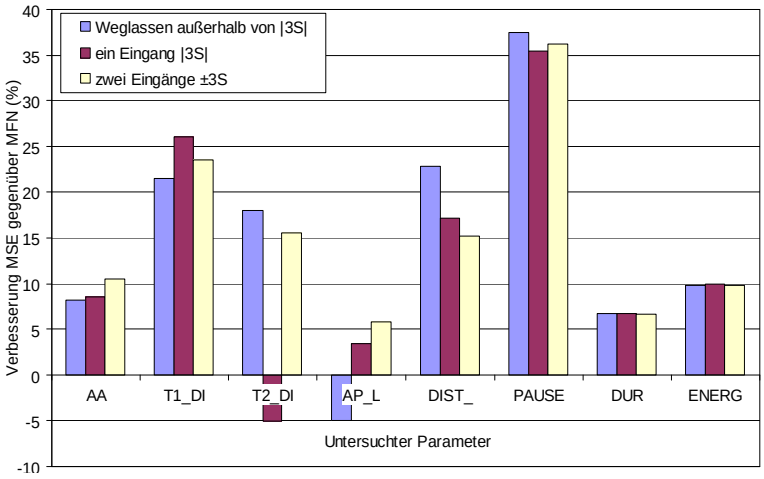

Abbildung 2 - Automatische Markierung von Extremwert-Daten

Das Training wird mit Markierungen für die ein- bis dreifache Standardabweichung durchgeführt (nur Trainingsdaten, 2. Aufnahmetag). Abbildung 2 zeigt die Ergebnisse für die dreifache Streuung. Bis auf einige *Ausreißer* beim Training ist gut zu erkennen, dass alle drei Konfigurationen ähnliche Verbesserungen der Prediktion nach sich ziehen. Besonders markant ist diese Beobachtung bei den zeitlichen Parametern *T1, DIST (T0)* bzw. *PAUSE*, wo sich der Fehler (*MSE*) um ca. 20% bis 30% reduziert. In Zusammenhang mit der geringen Anzahl an

Datensätzen, welche dabei markiert werden, ist zu vermuten, dass diese Extremwerte teilweise auf Label-Ungenauigkeiten in der Datenbasis zurückzuführen sind.

5 Evolutionäre Strukturoptimierung des Neuronalen Netzwerks

Evolutionäre Algorithmen (EA) sind Methoden zur stochastischen Optimierung, welchen den natürlichen Prozess von Evolution, Selektion und Variation simulieren. Selektion basiert auf dem „Überleben des Tüchtigsten“. Einzelne Lösungen kämpfen um Ressourcen und Vermehrung. Die Rekombination und Mutation von Genomen nennt man Variation. EA werden in vielen Bereichen angewandt, eingeschlossen Sprach- und Sprechererkennung.

In [6] gibt Takagi einen breiten Überblick über die Nutzung von Interactive Evolutionary Computation (IEC) für die Optimierung von Systemen, wobei als Grundlage die subjektive menschliche Bewertung dient. Außer IEC gibt es nur wenige Fälle, in welchen EA auf prosodische Problemstellungen angewandt werden. Kürzlich untersuchte Kruschke die evolutionäre Optimierung der automatischen Extraktion von Fujisaki-Parametern aus einer Sprachdatenbank [7]. In [5] untersuchten die Autoren die evolutionäre Strukturoptimierung für das neuronale Netzwerk des IGM-Modells bezüglich Prediktionsqualität und Aufwandsminimierung.

5.1 Multikriterielle Optimierungsprobleme (MOP)

Viele real existierende Probleme besitzen mehr als einen Zielparameter. Die Optimierung mehrerer (kontinuierlicher) Parameter verursacht eine unendliche Anzahl von optimalen Problemlösungen, die so genannten Pareto-optimalen Lösungen. Diese Menge an Lösungen wird auch Pareto-optimale Front genannt. Das Konzept der Pareto-Dominanz beschreibt die Beziehung zwischen einzelnen Lösungen. Eine Lösung dominiert eine andere Lösung, falls sie diese in mindestens einem Parameter übertrifft, während kein anderer Parameter schlechter ist. In Abbildung 3 werden alle Lösungen im dunkelgrauen Rechteck (unten links) von *B* dominiert. *B* selbst wird aber von jeder Lösung im Rechteck rechts oben dominiert. Die Lösungen *F* und *C* sind weder von *B* dominiert noch dominieren sie *B*, obwohl sie genauso wenig optimal sind. Lösungen wie *A*, die auf der Pareto-optimalen Front liegen, sind optimal.

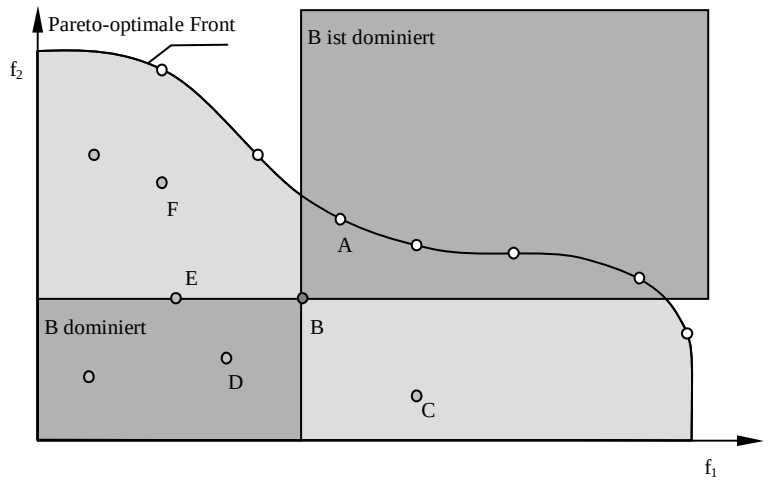

Abbildung 3: Pareto-optimale Front [9].

5.2 Strength Pareto Evolutionary Algorithm (SPEA)

Um die Pareto-optimale Front zu finden, existieren verschiedene Algorithmen. Unter Berücksichtigung der zahlreichen Eingangsgrößen und der Topologie des IGM fiel die Wahl auf den Strength Pareto Evolutionary Algorithm (SPEA[8]). Dabei werden alle bisher gefundenen

nichtdominierten Lösungen extern als Elite gespeichert. Die Fitness einer neuen Lösung wird dann nur unter Beachtung der Beziehungen zu Mitgliedern der Elite betrachtet. Die Anzahl der Lösungen in der Elite wird durch Verfahren wie Clustering klein gehalten. Dominierte Lösungen werden entfernt.

Die Struktur eines neuronalen Netzes ist durch zwei Extreme begrenzt: Auf der einen Seite kann das Netz so klein sein, dass es unfähig ist, alle Trainingsmuster und die darin liegenden Zusammenhänge vollständig zu lernen, auf der anderen Seite kann es zu groß sein, um die angebotenen Daten zu generalisieren und wird jeden Datensatz einzeln lernen. Es gibt drei mögliche Ansätze für die Optimierung des beschriebenen Netzwerkkerns (MFN) im IGM:

1. die Topologie des Netzwerks,
2. die Verringerung der Anzahl der Eingänge,
3. sowie die Vergrößerung der Anzahl der Eingänge.

Die dritte Methode würde einen Neuentwurf des gesamten Modells erfordern und wird daher hier nicht weiter verfolgt. Durch die Nutzung von EA sollen folgenden Optimierungsziele erreicht werden:

- Minimierung des quadratischen Fehlers (RMSE),
- Minimierung der Anzahl der Verbindungen,
- Minimierung der Anzahl der Eingänge,
- Minimierung der Anzahl der verdeckten Knoten.

Um das beschriebene multikriterielle Optimierungsproblem zu lösen, wird das Konzept der Pareto-Dominanz genutzt.

5.3 Experimente und Ergebnisse der evolutionären Optimierung

Das ursprüngliche MFN besteht aus 24 Eingängen, 744 Verbindungen, 30 verdeckten Knoten und 8 Ausgängen. Der minimal beobachtete Overall-RMSE nach dem Training beträgt 0.139.

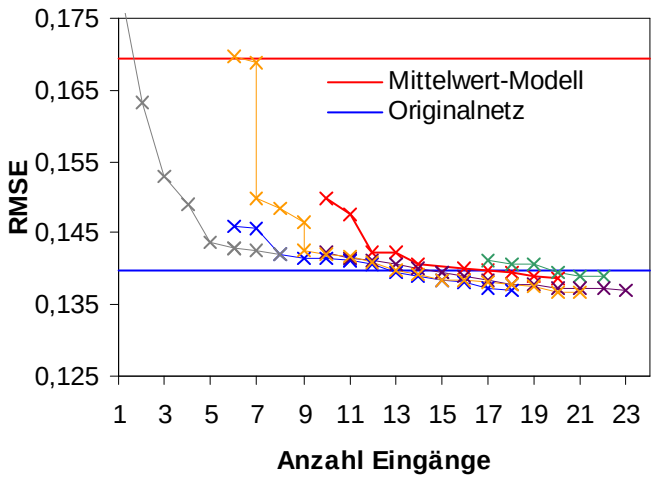

Abbildung 4: Einige Optimierungsdurchläufe im Vergleich zur ursprünglichen Netzleistung [9].

Die erste Methode, eine Vergrößerung der MFN-Struktur, zeigt nur wenig signifikante Ergebnisse. Zwei größere Konfigurationen (mit 40 und 50 verdeckten Knoten) verhalten sich ähnlich wie das ursprüngliche Netz. Es scheint, dass die bisherige Netzstruktur der Aufgabenstellung angemessen ist. Die Verwendung von SPEA zur Reduktion der generellen Netzwerk-

Topologie durch Löschen von Knoten und Verbindungen (wobei alle Durchläufe von einem neuen Training mittels Backpropagation gefolgt sind) zeigt keine signifikante Änderung des resultierenden RMSE. Unter Beachtung der Zeit pro Evolutionsdurchlauf von einigen Stunden wurde das Experiment nach einigen hundert Iterationen beendet.

Die zweite Methode, eine Reduzierung der Eingangsanzahl, zeigt signifikante Ergebnisse (Abbildung 4). Ein Netzwerk mit nur sechs Eingängen erreicht einen RMSE von ca. 0.145 – lediglich 4.4% schlechter als das Originalnetz. Die verbleibenden sechs Eingänge sind: *BI_SYN_R* (Break-Index rechts), *SYL_PREC* (Silben-Anzahl in der vorhergehenden Phrase), *BI_SYN_L* (Break-Index links), *SCHWA* (Schwa-Laut), *INTONEME* (Intonem-Typ) und *O_DUR* (Summe der durchschnittlichen Lautdauern im Silben-Onset). Die Eingänge des MFN werden in [9] beschrieben. Korrelationsanalysen in [3] bestätigen, dass ca. 80 bis 95% der Vorhersageleistung von nur fünf bis acht Eingabeparametern abhängen. Subjektive Hörtests zeigen, dass keine signifikanten, hörbaren Unterschiede zwischen den optimierten bzw. den ursprünglich geschätzten Grundfrequenz- und Dauer-Verläufen existieren.

5.4 Inkonsistenz der Eingangsdaten

Ein weiteres evolutionäres Experiment beschäftigte sich mit der korrekten Auswahl von Trainings- und Testmustern. Während des ursprünglichen Trainings und der Adaption des IGM wurden die Bedingungen, unter denen der benutzte Teil des Stuttgarter Nachrichtenkorpus aufgenommen wurde, als konstant angenommen. Da der EA jedoch eine spezifische Vorliebe für bestimmte Trainings- und Testmengen-Kombinationen zeigt, wurden Inkonsistenzen in den Eingabedaten entdeckt. Die Datenbank enthält Sprachsignale von zwei verschiedenen Tagen. Die Signale des ersten Tages sind auf die Spitzenamplitude normiert, was zu einem durchschnittlichen RMS-Wert von ca. -17 dB führt. Die Signale des zweiten Tages wurden nicht verändert und erreichen nur einen durchschnittlichen RMS-Wert von -21 dB – eine Abnahme um 4 dB. Die beobachtete Inkonsistenz kann z. B. auch durch unterschiedliche Sprechstile an beiden Tagen erklärt werden.

6 Weitere Analyse der zu schätzenden IGM-Modellparameter

Die Basiseinheit bei der Modifikation der prosodischen Parameter mittels IGM ist die Sprechsilbe. Dieser Ansatz entspricht der humanen Spracherzeugung und wird auch in vergleichbaren Prosodiemodellen verfolgt. Nichtsdestotrotz muss das Modell auch die natürliche Parametrisierung im Subsilben-Bereich (primär Lautebene) korrekt widerspiegeln. Bei der Steuerung der Grundfrequenz f0 wird eine qualifizierte Parametrisierung durch die nach geschaltete Fujisaki-Formel und bei der silbenbasierten Dauermodellierung durch die nachfolgende Lautdauerverteilung gemäß Elastizitätsmodell von Campbell [1] erreicht.

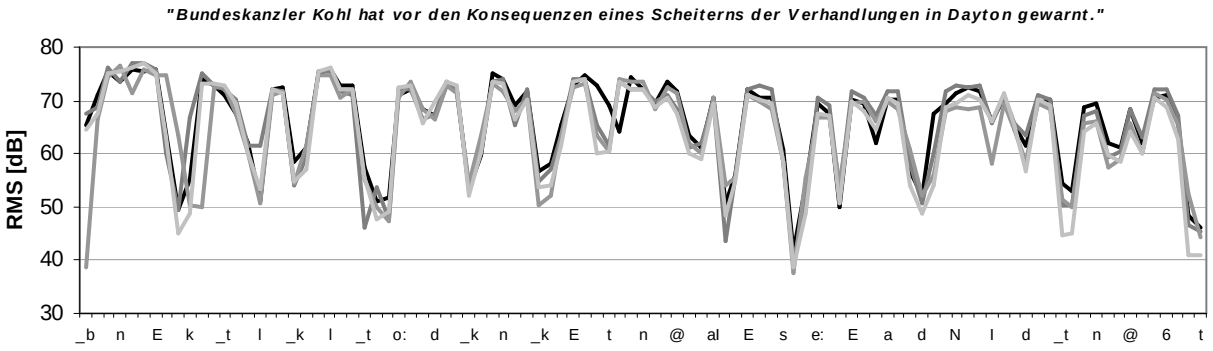

Abbildung 5: Intra-individuelle Streuungen des Intensitätsverlaufs für einen typischen Satz.

Hingegen wurden der Einfluss einer Intensitätssteuerung in der Sprachsynthese, insbesondere für das Deutsche, und damit auch die Wechselwirkungen zwischen globalen und lokalen Einflussfaktoren bisher kaum untersucht. Jokisch und Kühne [10] untersuchten typische Intensitätsmuster im Stuttgarter Nachrichtenkorpus und dabei vor allem folgende Aspekte:

- Intra-individuelle Streuung bei der Wiederholung von Phrasen,
- Phonemposition in der Silbe,
- Vokalkontext-Einfluss,
- Interaktion des Intensitätsverlaufs mit anderen prosodischen Parametern.

Abbildung 5 zeigt die erstaunlich geringen **intra-individuellen** Streuungen im Intensitätsverlauf eines männlichen Sprechers aus der erwähnten Datenbasis, wenn Nachrichtenäußerungen im Abstand von 30 min. identisch wiederholt werden.

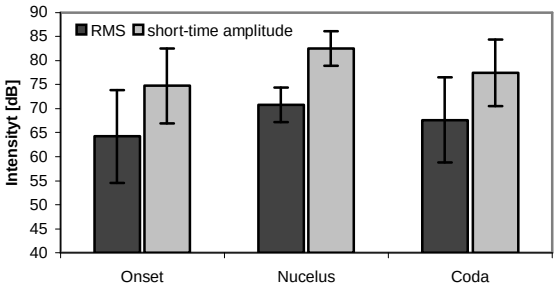

Abbildung 6: Lautintensität an verschiedenen Silbenpositionen.

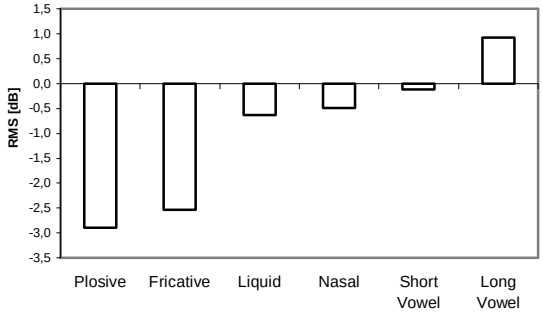

Abbildung 7: Kontexteinfluss auf die Intensität von Langvokalen, abhängig vom Folgelaut.

Position sowie **Kontext** (z. B. innerhalb der Silbe) haben ebenfalls großen Einfluss auf die Intensitätsverteilung in der Lautebene, wie Abbildungen 6 und 7 beispielhaft verdeutlichen.

Den stärksten Einfluss auf die Intensitätsverteilung hat gemäß [10] die **Interaktion mit dem Parameter Grundfrequenz f0**. Der Korrelationskoeffizient nach Neyman-Pearson zwischen f0 und Kurzzeitamplitude bzw. RMS erreicht für den untersuchten Nachrichtenkorpus Werte von 0.69 bzw. 0.70. Abbildung 8 visualisiert die Korrelation zwischen f0-Verlauf und den Kurzzeitamplituden- bzw. RMS-Verläufen für den Ausschnitt aus einer typischen Äußerung.

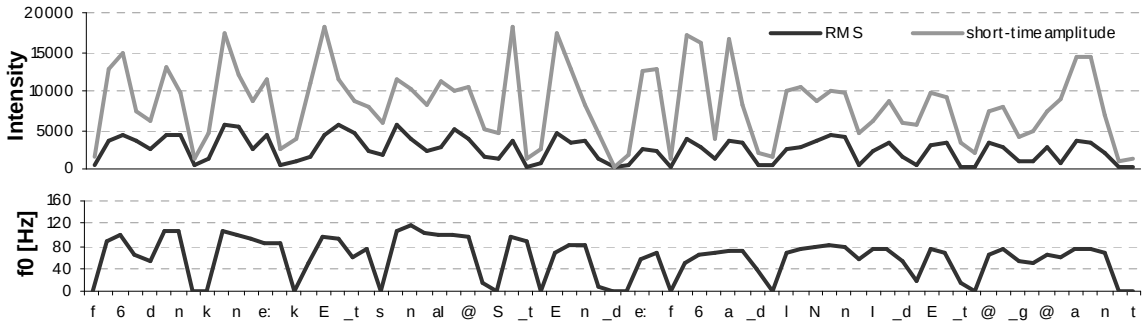

Abbildung 8: Interaktion zwischen Grundfrequenz f0 und Sprechintensität (Satzausschnitt).

7 Zusammenfassung

Die Analyse der Optimierungsergebnisse zeigt den großen Einfluss der Trainingsdaten. Weitere Untersuchungen sind notwendig, um mittels zusätzlicher Eingangsinformation bzw. korrigierter Datensätze die Performanz des IGM zu steigern. Für diese Untersuchung ist vor allem die Analyse von Extremwerten in den Daten, einschließlich ihrer Ursachen interessant.

Die vorgeschlagene evolutionäre Optimierung mit Hilfe von SPEA ist geeignet, Netzwerktopologien zu überprüfen und zu optimieren, z. B. für die Verwendung in Embedded TTS-Systemen. Allerdings ist es nicht möglich, die Vorhersagequalität des IGM damit zu erhöhen. Datenbasierte Modelle wie das IGM benötigen zusätzliches Wissen zur Verbesserung des Ergebnisses, z. B. semantische Informationen.

Die Untersuchung zur Intensitätsparametrisierung zeigt, dass eine Detaillierung der silbenorientierten Intensitätssteuerung des IGM auf der Lautebene sinnvoll ist. Aufgrund der Korrelation zwischen Grundfrequenz- und Intensitätsverlauf wird eine Kopplung zwischen diesen beiden Parametern angestrebt. Grundfrequenz- und Dauermodellierung auf Sub-Silbenebene sind durch die Integration etablierter quantitativer Modelle hinreichend qualifiziert.

Literatur

[1] W.N. Campbell, “Syllable-based segmental duration.” *In: Bailly, G. and Benoît, C. (ed.), Talking Machines: Theories, Models, and Designs*, 211-224, Elsevier Science, 1992.

[2] C. Traber, “F0 generation with a database of natural f0 patterns and with a neural network.” *In: Bailly, G. and Benoît, C. (ed.), Talking Machines: Theories, Models, and Designs*, 287-304, Elsevier Science, 1992.

[3] H. Mixdorff and O. Jokisch, “Evaluating the quality of an integrated model of German prosody”, *International Journal of Speech Technology (IJST) vol. 6 (issue 1)*, 45-55, Kluwer Academic Publishers, 2003.

[4] O. Jokisch, H. Mixdorff, H. Kruschke, U. Kordon, “Learning the parameters of quantitative prosody models”, *Proc. ICSLP 2000*, Beijing, 645-648, 2000.

[5] O. Jokisch and M. Hofmann, „Evolutionary Optimization of an Adaptive Prosody Model“, *Proc. ICSLP*, Jeju, Korea, 2004 (in Druck).

[6] H. Takagi, “Interactive evolutionary computation: Fusion of the Capabilities of EC optimization and human evaluation”, *IEEE Proc., vol. 89, no. 9*, 1275-1296, 2001.

[7] H. Kruschke and A. Koch, “Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization”, *Proc. ICASSP*, vol. 1, 524-527, Hong Kong, 2003.

[8] E. Zitzler, “Evolutionary algorithms for multiobjective optimization: methods and applications”, *PhD thesis, ETH Zurich*, 1999.

[9] F. Kossebau, “Evolutionary optimization of a trainable prosody computation, *Diploma thesis, TU Dresden*, 2003 (in German).

[10] O. Jokisch and M. Kühne, “An investigation of intensity patterns for German”, *Proc. EUROSPEECH*, 165-168, Geneva, 2003.