

Evaluation eines Sprachsynthesystems nach dem Prinzip der Non-Uniform Unit Selection

IKP-Arbeitsbericht NF 10

Robert Hammerstingl, Stefan Breuer

rob.hammer@netcologne.de, breuer@ikp.uni-bonn.de

Abstract: Am IKP wurde basierend auf dem Bonn Open Synthesis System (BOSS) II eine automatisierte Telefonauskunft für die Firma *klickTel* entwickelt. Die Performanz dieser Anwendung zur Namenssynthese wurde evaluiert. Sowohl Funktion als auch Methode dieser Evaluation folgten den speziellen Anforderungen an diese Anwendung. Um dies zu leisten, wurde ein diagnostisches Evaluationsverfahren entwickelt und ein bestehendes globales Testverfahren wurde dem Evaluationsfokus angepasst. Zum Zeitpunkt der Evaluation wurde an einem Modul zur Signalmanipulation gearbeitet. Neben einer spektralen Anpassung der Bausteine in der näheren Umgebung der Konkatenationsstellen sollte die Dauer der genutzten Syntheseeinheiten manipuliert werden. Vor einer Implementierung dieses Moduls in die *klickTel*-Anwendung sollten mögliche Effekte auf die segmentale Verständlichkeit der Namen diagnostisch und global evaluiert werden. Zusätzlich sollte in einer globalen Evaluation untersucht werden, inwiefern die wahrgenommene Annehmlichkeit der Syntheseausgabe durch die zusätzliche Signalmanipulation beeinflusst wird. Abschließend sollte die Frage geklärt werden, ob der Trägersatz bzw. Rahmendialog, in den die Namen eingebettet werden, von der Korpusprecherin gesprochen oder synthetisiert dargeboten werden soll. Um die besondere Problematik der Evaluation eines Sprachsynthesystems nach dem Prinzip der Non-Uniform Unit Selection zu berücksichtigen, wurden die Teststimuli nicht basierend auf dem ganzen Korpus, sondern anhand ihrer Repräsentativität für die Domäne einer Anwendung ausgewählt.

1 Das Bonn Open Synthesis System (BOSS) II

Das am Institut für Kommunikationsforschung und Phonetik (IKP) entwickelte Bonn Open Synthesis System (BOSS) II bietet den Entwicklern von Sprachsynthesystemen eine open source-Softwareplattform, die von Beginn an für die Aufgaben der Non-Uniform Unit Selection entwickelt wurde [1]. Die Architektur von BOSS trennt, soweit möglich, Daten und Algorithmen. Dies ermöglicht die relativ einfache Anpassung des Systems auf verschiedene Korpora und Domänen. Die BOSS-Software ist in zwei Hauptprogramme gegliedert, den

Client und den Server. Im Client ist die Textvorverarbeitungsstufe untergebracht. Damit leistet der Client die spezifische Anpassung an die jeweilige Anwendung, sendet den Input an den Server und erhält dessen Ausgabe. Der Server erzeugt das synthetische Sprachsignal. Die zurzeit integrierten Servermodule leisten die drei Hauptaufgaben Transkription, Dauerprädiktion, Unit Selection und Synthese. Weitere Module zur Prosodieprädiktion und zur Signalmanipulation befinden sich in der Entwicklung.

1.1 Die Boss II Unit Selection und Kostenfunktionen

Als Weiterentwicklung der datenbasierten Synthese verwenden Unit-Selection-Systeme bei der Korpusynthese eine große Anzahl von Synthesebausteinen in unterschiedlichen Größen, Umgebungen und Realisierungen, um die notwendige Modifikation der Dauer und Grundfrequenz auf ein Minimum zu reduzieren. Derjenige Baustein aus einem Inventar natürlicher sprachlicher Äußerungen wird genutzt, der der prosodischen und segmentalen Zieleinheit der zu synthetisierenden Äußerung am nächsten kommt. Dabei kann bei der Non-Uniform Unit Selection innerhalb eines Systems die Größe der genutzten Einheiten von (Halb-)Phonen bis hin zu ganzen Sätzen variieren. Um die Anzahl der Konkatenationsstellen zu minimieren, wird im Korpus nach möglichst großen Einheiten gesucht, die auch der Prosodie der Zieläußerung entsprechen sollen. Erst wenn keine passenden Sätze oder Wörter gefunden werden, wird auf Bausteine unterhalb der Wortebene zurückgegriffen. Auch diese sollen wieder möglichst groß sein, um mehrere Laute im natürlichen Zusammenhang auswählen zu können. Während beispielsweise bei der Diphonsynthese nur ein Diphon aus einem Trägersatz als Synthesebaustein genutzt wird, kann bei der Korpusynthese auf jedes Wort des gesamten Trägersatzes zurückgegriffen werden, so dass das Korpus sowohl laut- als auch wortsegmentiert werden muss. Der Erfolg dieser Syntheseart hängt entscheidend von der Bestimmung der Kriterien ab, nach denen die Auswahl der Bausteininstanzen erfolgt. Die Eignung jedes im Inventar gespeicherten Bausteins in Bezug auf eine zu synthetisierende Äußerung wird für verschiedene Parameter gemessen und soll dann mit Hilfe einer Kostenfunktion global bestimmt werden. Diese ergibt sich zum einen aus den Übergangskosten, die darstellen, wie gut ein Baustein mit seinem Vorgänger bzw. Nachfolger zusammenpasst; je niedriger beispielsweise die Differenz der aufeinander folgenden Bausteine bezüglich der Energie und der Grundfrequenz ist, desto niedriger sind die Übergangskosten. Des Weiteren werden Einheitenkosten berechnet, die den Unterschied eines spezifischen Bausteins mit den an dieser Stelle benötigten Eigenschaften des zu synthetisierenden Sprachsignals darstellen. Bei BOSS II berechnet sich beispielsweise ein Teil der Einheitenkosten aus der Differenz der vorhergesagten Dauer einer Einheit und der tatsächlichen Dauer eines Kandidaten. Diejenige Bausteinssequenz wird ausgewählt, deren aufsummierte Einheiten- und Übergangskosten minimal sind. Die genaue Definition der einzelnen Kosten, wie auch ihr prozentualer Anteil an den Gesamtkosten und damit ihr Entscheidungsgewicht, muss von dem Entwickler einer bestimmten Anwendung individuell festgelegt werden.

Das primäre Syntheseelement der Einheitenauswahl von BOSS II ist das Wort. Zur nächst kleineren Einheit wird nur dann gewechselt, wenn das betreffende Wort im Korpus nicht vorhanden ist, oder die zugewiesenen Attribute eines bestimmten Wortes seine Nutzung als Element der Synthese verbieten. Die dann genutzte Einheit ist die Silbe. Findet sich auch hier kein passender Kandidat, so wird auf die kleinstmögliche Einheit, das Phon gewechselt. Für jedes Element steht nach der Vorauswahl eine größere Anzahl von Kandidaten bereit. Im zweiten Schritt der Unit Selection wird aus allen möglichen Sequenzen der unterschiedlichen

Kandidaten diejenige ausgewählt, die die gewünschte Äußerung so genau wie möglich approximiert. Die Synthesekomponente konkateniert die ausgewählten Einheiten. Zum Zeitpunkt der Evaluation wurde am synthetischen Signal keine weitere spektrale Manipulation durchgeführt.

2 Korpus und Client der *klickTel*-Anwendung

Bei der am IKP entwickelten Anwendung von BOSS II handelt es sich um eine Sprachsynthese, die im Rahmen einer automatisierten Telefonauskunft der Firma *klickTel* eingesetzt werden soll [2]. Der Anrufer erhält die von ihm gewünschte Information, nämlich Namen, Adresse und Telefonnummer einer bestimmten Person, in Form von synthetischer Sprache. Teile der Sprachausgabe dieser Anwendung, die Personen-, Orts- und Straßennamen unterliegen einer speziellen Problematik [3]. Verglichen mit einer domänenunabhängigen TTS-Anwendung kann bei dieser Anwendung ein Vorteil der inhaltsgesteuerten Sprachsynthese (CTS) genutzt werden, da die Syntax der zu synthetisierenden Sätze bekannt ist. Die Namensklassen (Vor-, Nach-, Straßen- und Ortsnamen) stehen immer an je einer bestimmten Phrasenposition, die bei der Korpuserstellung und bei der Einheitenauswahl berücksichtigt wurde.

Das für diese Anwendung erstellte Korpus enthält ca. fünf Stunden gesprochenen Textes einer Sprecherin. Die Wortgrenzen sind handsegmentiert, die Lautsegmentierung erfolgte maschinell und wurde manuell korrigiert. Der Anwendung entsprechend handelt es sich um Aussagesätze, in denen die verschiedenen Namen in bestimmten prosodischen Realisierungen zur Verfügung stehen. Nach dem Prinzip des Phrase-Slot-Fillings werden Vor-, Nach- und Straßennamen, sowie Hausnummern in den Trägersätzen mit progredienter Intonation an der Phrasengrenze, gefolgt von einer Pause realisiert. Ortsnamen werden mit finaler Intonation an der Phrasengrenze, gefolgt von einer Pause, gesprochen. Des weiteren sind Zahlen und Zahlenpaare in unterschiedlichen prosodischen Realisierungen, sowie Einheiten für die Buchstabierung enthalten. Die Anzahl aller Bausteinkandidaten des Inventars dieser Anwendung lag zum Untersuchungszeitpunkt bei 104.343 Phonen, 55.511 Silben und 34.692 Wörtern.

3 Evaluation der *klickTel*-Anwendung von BOSS II

Wie gezeigt wurde [4], kann ein Sprachsynthesesystem nur unter Berücksichtigung seines Syntheseverfahrens und seiner potenziellen Anwendung getestet werden. Sowohl die Funktion als auch die Methode der hier vorgestellten Evaluation richten sich primär nach den speziellen Anforderungen an die *klickTel*-Anwendung von BOSS II. Die Wahl von Namen als Teststimuli birgt den Vorteil, dass es unproblematisch ist, mehrsilbige Stimuli zu finden. Ob die phonemische Ausbalanciertheit, d.h. die Lauthäufigkeit in den Namen für die jeweilige Sprache repräsentativ ist, erscheint fraglich. Ebenso wenig kann man davon ausgehen, dass in einer Sprache gebräuchliche Namen deren Phonotaktik folgen, da es in Namen teilweise zu regional oder fremdsprachlich geprägten Lautfolgen kommt, die nicht der jeweiligen Hochsprache entsprechen. Diese Aspekte zeigen, dass Namen als Stimuli für segmentale Verständlichkeitsmessungen von Inhalts- und Funktionswörtern als für eine Sprache nur bedingt repräsentativ zu erachten sind. Bei einer Anwendung hingegen, die eine namensorientierte Ausgabe zum Ziel hat, erscheint die Wahl von Namen als Stimuli durchaus als sinnvoll. Das Ziel, ein solches Synthesesystem möglichst anwendungsspezifisch zu testen, spricht für die Wahl von Namen als Teststimuli. Das genutzte und hier zu berücksichtigende Syntheseprinzip, die Non-Uniform Unit Selection, unterliegt bezüglich einer Evaluation der segmentalen Verständ-

lichkeit, verglichen mit den ‚klassischen‘ datenbasierten Verfahren, einer besonderen Problematik, die im folgenden Abschnitt erläutert werden soll.

3.1 Problematik bei der Evaluation eines Non-Uniform-Unit-Selection-Systems

Für eine detaillierte Evaluation eines Sprachsynthesystems ist eine quantitative Messung der Verständlichkeit (bzw. der Verstehbarkeit) notwendig [5]. Hierbei spielt das jeweilige Syntheseprinzip eine große Rolle. Um die interne Validität zu gewährleisten, muss ein Test genügend Beispiele aller denkbaren Variationen in Bezug auf das zu messende Kriterium beinhalten. Will man die Verständlichkeit eines datenbasierten Sprachsynthesystems testen, so müssen genügend Bausteine des zugrunde liegenden Inventars betrachtet werden. Das Ziel ist zum einen, schlecht artikulierte oder fehlerhaft segmentierte Bausteine zu finden, die in der Anwendung des Synthesystems immer dann Fehler hervorrufen, wenn diese Bausteine zum Einsatz gelangen. Zum anderen soll die Verständlichkeit der Systemausgabe mit Hilfe von Hörern, die das System nicht kennen, dokumentiert werden. Ein Non-Uniform-Unit-Selection-System greift auf ein sehr großes Inventar zurück. Zudem ist die Größe der ausgewählten Einheiten variabel. Die Anzahl aller möglichen Bausteinkandidaten des Inventars dieser Anwendung liegt bei ca. 195.000 Einheiten. Will man die Verständlichkeit auch bei diesen Systemen auditiv, d.h. in Hörversuchen, ermitteln, erscheint es wegen der großen Anzahl aller möglichen Bausteinkandidaten sehr schwierig, eine repräsentative Stichprobe basierend auf dem gesamten Korpus zu finden. Unter Berücksichtigung einer begrenzten Anzahl an Testpersonen, sowie deren zeitlicher und mentaler Belastungsgrenzen, erscheint es unmöglich, alle Bausteinkandidaten eines Korpus-synthesystems perceptiv zu testen. Die vorliegende Evaluation versucht dieses Problem zu lösen, indem die Teststimuli nicht basierend auf dem ganzen Korpus einer Anwendung, sondern anhand ihrer Repräsentativität für die Domäne einer Anwendung ausgewählt werden.

Ein weiteres Problem bei der Evaluation eines Non-Uniform-Unit-Selection-Systems ist durch die unterschiedlichen Ebenen der Einheitenauswahl begründet. Je nachdem, ob ein Wort komplett durch das Synthesekorpus abgedeckt wird, oder ob es durch Elemente tieferer Selektionsebenen synthetisiert wird, schwankt die Synthesequalität erheblich [6]. Diesem Zusammenhang wird in der folgenden Evaluation Rechnung getragen, indem die Verständlichkeit in Bezug auf die Anzahl der Konkatenationen betrachtet wird.

3.2 Funktion der Evaluation

Da das Ziel der klickTel-Anwendung die synthetisierte Ausgabe von Namen ist, ist die segmentale Verständlichkeit der Sprachsynthese von primärer Bedeutung. Die zu beurteilenden Vor-, Nach-, Orts- und Straßennamen werden in einen für diese Anwendung typischen Trägersatz gebettet. Die Verständlichkeitsfehler werden im Verhältnis zur Anzahl der Konkatenationen und zur Fehlerklasse (Einfügung, Auslassung oder Substitution) analysiert. Am IKP wurde zum Zeitpunkt der Evaluation ein Modul zur Signalmanipulation entwickelt. Neben einer spektralen Anpassung der Bausteine in der näheren Umgebung der Konkatenationsstellen sollte die Dauer der genutzten Syntheseeinheiten manipuliert werden. Um die verschiedenen Effekte dieser Dauermanipulation vorab zu testen, wurden die Stimuli auf folgende Weise manipuliert: Auf Basis der von der Dauerprädiktion auf Phonebene geschätzten Dauerwerte für die Zieläußerung wurden die zur Synthese genutzten Einheiten angegli-

chen. Vor einer Implementierung dieses Moduls in die klickTel-Anwendung sollten mögliche Effekte auf die segmentale Verständlichkeit von Namen diagnostisch evaluiert werden. Zusätzlich sollte in einer globalen Evaluation untersucht werden, inwiefern die wahrgenommene Annehmlichkeit der Syntheseausgabe durch die zusätzliche Signalmanipulation beeinflusst wird. Abschließend sollte die Frage geklärt werden, ob der Trägersatz, in den die Namen eingebettet werden, von der Korpus Sprecherin gesprochen oder synthetisiert dargeboten werden soll. Möglicherweise wird der abrupte Wechsel von einer natürlichen zu einer synthetischen Stimme vom Hörer negativ beurteilt. Andererseits wirkt sich der im Verbund eingesprochene Trägersatz möglicherweise positiv auf die empfundene Natürlichkeit der gesamten Äußerung aus. Diesbezügliche Effekte auf die wahrgenommene Annehmlichkeit der Stimme sollten global evaluiert werden.

3.3 Methode der Evaluation

Für die Auswahl der Stimuli wurden die Namen im Korpus der klickTel-Anwendung untersucht. Diese Namen wurden in Orts-, Straßen- und Personennamensklassen zusammengefasst. Zur Untersuchung dieser Namen wurde ein Programm erstellt, das Triphon-Klassen innerhalb des Korpus bildet. Als Triphon-Klasse wurde ein Lautcluster von drei phoxsy-Einheiten¹ [7], [8] definiert. Da bei der phoxsy-Einheitendefinition zeitlich schwer segmentierbare Lautkombinationen zu einem Multiphon-Segment zusammengefasst werden, können die hier definierten Triphon-Klassen im phoxsy-Format aus mehr als drei Phonen bestehen. Dann wurden die häufigsten Triphon-Klassen innerhalb der Namen des zugrunde liegenden Korpus maschinell gesucht und nach Häufigkeit sortiert. Somit konnten für die Personen-, Straßen- und Ortsnamensklassen die häufigsten Triphon-Klassen ermittelt werden. Es wird davon ausgegangen, dass diese Triphon-Klassen, die am häufigsten in den Namen des Korpus vorkommen, auch relativ häufig in der Anwendung zum Einsatz gelangen. Dann wurden die häufigsten Triphon-Klassen der jeweiligen Namensklasse in möglichst selten vorkommenden Namen des zugrunde liegenden Telefonbuchs (klickTel 2000) gesucht. Als Stimuli wurde für jede Triphon-Klasse ein möglichst seltener, möglichst zweisilbiger, im deutschen Sprachraum gebräuchlicher Name gewählt. Auf diese Weise wurden für die Personennamensklasse 102 Stimuli (51 Vor- und 51 Nachnamen) und für die Straßen- und Ortsnamensklasse jeweils 51 Stimuli ausgewählt.

3.3.1 Funktionaler Verständlichkeitstest:

Der erste Teil der Evaluation bestand aus einem funktionalen Verständlichkeitstest. Die Versuchspersonen (VPen) hörten die verschiedenen Namen in zwei verschiedenen Synthesevarianten. Die segmentale und globale Verständlichkeit wurde mit und ohne zusätzliche Dauermanipulation derselben Stimuli gemessen. Somit wurden zwei verschiedene Stimulisets getestet:

- Set 1) das Stimuliset nicht manipuliert (original)
- Set 2) das Stimuliset dauermanipuliert

¹ Bei phoxsy, den phone extensions for synthesis, handelt es sich um Phon- und Multiphon-segmente, die als Basiseinheiten der Konkatenation in BOSS dienen.

Beide Stimulisets wurden in einem applikationstypischen Trägersatz dargeboten, bei dem nur die zu beurteilenden Namen ausgetauscht wurden. Es wurde die offene Antwortform gewählt, die Versuchspersonen sprachen die synthetisierten Namen nach und schrieben sie auf. Die Transkription der Stimuli wurde mit den mündlichen VP-Antworten verglichen. Wie eine Antwort von der Transkription des Stimulus ab, wurde der Verständlichkeitsfehler auf dem Auswertungsblatt transkribiert. Zur späteren Analyse enthielt das Auswertungsblatt, neben der orthographischen und der transkribierten Repräsentation der Stimuli, die Angabe über die Orte der Konkatenationen der synthetisierten Namen. Zur Kontrolle der Fehlertranskription wurde die orthographische Antwort der VP herangezogen. Bei Einfügungen und Auslassungen wurden orthographische und mündliche Antwort mit dem gleichen Gewicht berücksichtigt. Bei Substitutionen entschied die mündliche Antwort der VP.

3.3.2 Präferenztest

Im zweiten Teil der Evaluation wurde ein Präferenztest [9],[10] durchgeführt. Den Versuchspersonen wurden Auszüge der beiden beschriebenen Stimulisets, zuzüglich einer weiteren Bedingung, dargeboten. Zum Vergleich mit der Darbietung in dem synthetisiert gesprochenen Trägersatz sollten die Stimuli in einen natürlichen Trägersatz gebettet beurteilt werden. Es ergaben sich vier Stimulikonditionen:

- Kondition 1: Set 1) in natürlichem Trägersatz (nat_orig)
- Kondition 2: Set 1) in synthetisiertem Trägersatz (synth_orig)
- Kondition 3: Set 2) in natürlichem Trägersatz (nat_manip)
- Kondition 4: Set 2) in synthetisiertem Trägersatz (synth_manip)

Den Versuchspersonen wurde ein Satz in zwei unterschiedlichen Konditionen präsentiert und es sollte die bevorzugte Variante benannt werden. Jede Version eines Satzes wurde mit allen anderen Versionen desselben Satzes verglichen. Alle möglichen Kombinationen, einschließlich der umgekehrten Reihenfolge (AB-BA) der Satzpaare, wurden getestet.

4 Ergebnisse

4.1 Ergebnisse und Bewertung des funktionalen Verständlichkeitstests

Da die Synthese von Zahlen einen wichtigen Teil der klickTel-Anwendung darstellt, war auch sie ein Ziel der vorliegenden Evaluation. Die Auswertung der Ergebnisse des Verständlichkeitstests zeigte, dass bei keiner Versuchsperson ein Verständlichkeitsfehler bezüglich der Hausnummern in den Stimulisätzen auftrat. Dieses Ergebnis wird damit begründet, dass innerhalb der Hausnummern keine Konkatenationsstellen auftraten, also auf Wortebene synthetisiert wurde. Bei der weiteren Auswertung der Ergebnisse des Verständlichkeitstests wurden die Hausnummern nicht berücksichtigt. Für die Bewertung der segmentalen Verständlichkeit der Namen wird das Maß Phonemfehlerrate (Phf) genutzt [11]. Dieses ist wie folgt definiert:

$$\text{Ph}_f = \frac{\text{Einfügungen} + \text{Auslassungen} + \text{Substitutionen}}{\text{Anzahl der gesamten Phoneme} \cdot \text{Anzahl der Versuchspersonen}} \cdot 100\%$$

Für die Bewertung der globalen Verständlichkeit der Namen wird das Maß Namensfehler-rate (Naf) definiert. Ein Name gilt als global falsch verstanden, sobald sich die transkribierte VP-Antwort eines Namens um mindestens ein Zeichen von der kanonischen Transkription unterscheidet:

$$Naf = \frac{\text{Namen mit Verständlichkeitsfehler} \geq 1}{\text{Anzahl aller Namen} (\cdot \text{Anzahl der Versuchspersonen})} \cdot 100 \%$$

Für die beiden Stimulisets ergaben sich folgende Fehlerraten:

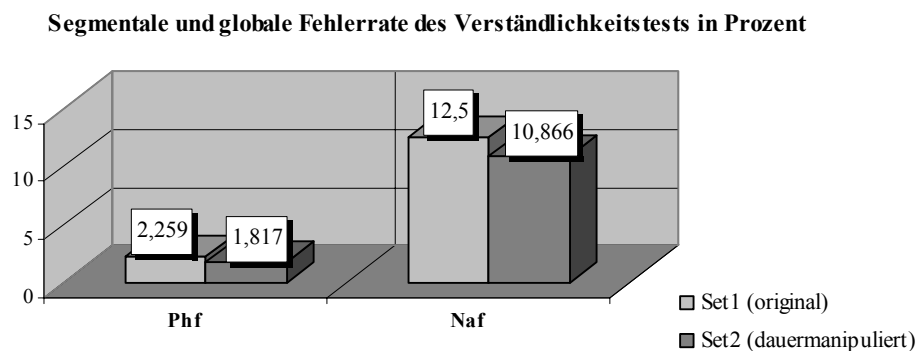


Abbildung 4.1: Segmentale und globale Fehlerrate des Verständlichkeitstests in Prozent.

Abbildung 4.1 zeigt, dass die Phonemfehlerrate der klickTel-Anwendung von BOSS II ohne zusätzliche Dauermanipulation bei 2,259% lag. Dieser Wert zeigte sich deutlich niedriger als erwartet und wird hier als – für synthetische Sprache – durchaus positiv bewertet. Die segmentale Performanz der Anwendung konnte durch die zusätzliche Dauermanipulation noch geringfügig auf Phf = 1,817% verbessert werden. Dieser Unterschied erwies sich als nicht signifikant. Die globale Namensfehlerrate der synthetisierten Namen ohne zusätzliche Dauermanipulation lag bei 12,5%. Die segmentale Verbesserung durch die zusätzliche Dauermanipulation zeigte sich in der globalen Namensfehlerrate des Systems etwas deutlicher mit der Verringerung von Naf um 1,634 Prozentpunkte auf 10,866%. Auch dieser Unterschied erwies sich als nicht signifikant. Da die Verständlichkeit von Namen als, im Vergleich mit der Verständlichkeit von Wörtern, problematischer anzusehen ist [3], wird auch dieses Ergebnis hier positiv bewertet.

Fehlerklassen des segmentalen Verständlichkeitstests in Prozent

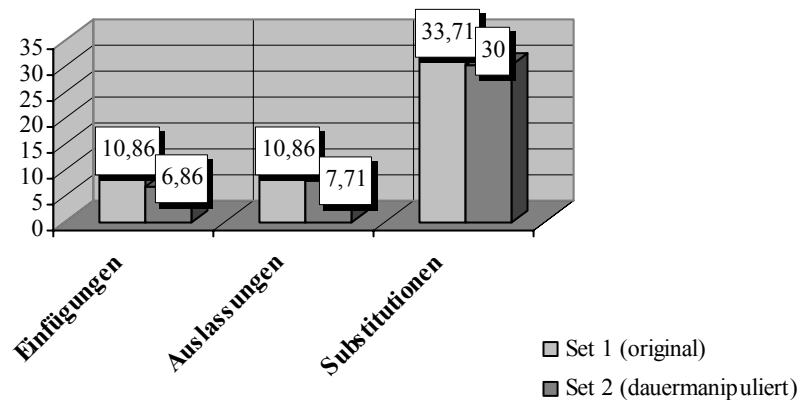


Abbildung 4.2: Fehlerklassen des segmentalen Verständlichkeitstests in Prozent.

Wie Abbildung 4.2 zeigt, wurden die meisten segmentalen Verständlichkeitsfehler in Form von Substitutionen gemacht, das heißt ein Phon wurde mit einem anderen Phon verwechselt. Die Analyse zeigte, dass die Unterschiede zwischen den Substitutionen und den anderen beiden Fehlerklassen hochsignifikant waren. Sowohl Abbildung 4.2, als auch die Untersuchung der segmentalen Verständlichkeitsfehler bzgl. Fehlerorte (Vor-, Nach-, Orts- und Straßennamen) zeigte in der Tendenz über alle Orte und Klassen die bessere segmentale Verständlichkeit der dauermanipulierten Stimuli gegenüber den originalen Namen. Die auditive und spektrale Analyse der Substitutionsfehler ergab, dass bei den Substitutionen die häufigste Ursache der segmentalen Verständlichkeitsfehler eine falsche Kennzeichnung der ausgewählten Bausteine war, gefolgt von der Verursachung durch eine falsche Segmentierung.

Zur weiteren Analyse der segmentalen Verständlichkeitsfehler wurden in den Transkriptionen der Stimuli auch die Konkatenationsstellen angegeben. Es wurde das Verhältnis zwischen der Anzahl der Konkatenationen und den Orten der Fehler betrachtet. Die Analyse der segmentalen Verständlichkeitsfehler in Abhängigkeit zur Konkatenationsstellendichte erfolgte auf Satzebene und die Anzahl der Verständlichkeitsfehler beider VPen-Gruppen wurde je Satz addiert. Da die 51 Stimulisätze unterschiedlich viele Phone enthielten, wurde eine Normierung eingeführt, wozu die Anzahl aller Verkettungsstellen eines Satzes durch die Anzahl aller Phone dieses Satzes dividiert wurde. Dabei wurde nicht zwischen Verkettungsstellen von Wörtern und Lauten unterschieden. Die Abbildung 4.3 zeigt die Anzahl der Fehler in Abhängigkeit zur Konkatenationsstellendichte:

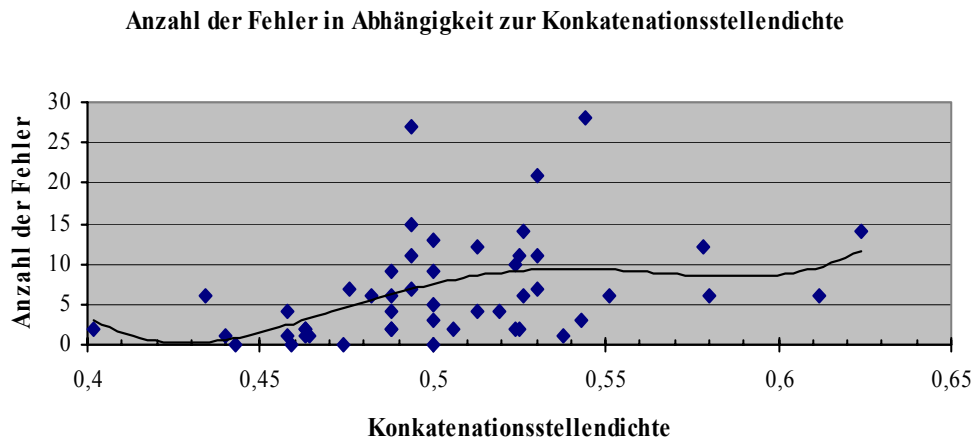


Abbildung 4.3: Anzahl der segmentalen Verständlichkeitsfehler in Abhängigkeit zur Konkatinationsstellendichte. Die durchgezogene Linie zeigt die polynomische Trendlinie (Ordnung 4) zur grafischen Darstellung des Trends in den Daten. Ein Datenpunkt in der Abbildung steht für einen Stimulussatz.

Die Anzahl der Konkatinationsstellen lag zwischen 33 und 53 je Satz. Die Anzahl der Phone lag zwischen 76 und 91 je Satz. In der Abbildung 4.3 zeigt sich, dass die Anzahl der segmentalen Verständlichkeitsfehler mit zunehmender Konkatinationsstellendichte wuchs. Insbesondere alle Sätze mit einer KSD unter 0,49 wiesen eine Fehlerzahl kleiner 10 auf. Die gehäufte Anzahl der Fehler zwischen $KSD > 0,49$ und $KSD < 0,55$ erklärt sich dadurch, dass die meisten Testsätze innerhalb dieser Werte lagen. Auch innerhalb dieser KSD-Werte zeigt die Tendenz die Steigerung der segmentalen Verständlichkeitsfehler mit wachsender Konkatinationsstellendichte. Dieses Ergebnis weist auf eine systembedingte Problematik der Sprachsynthese nach dem Prinzip der Non-Uniform Unit Selection hin. Da die Größe der Bausteine uneinheitlich ist, ist auch die segmentale Verständlichkeit der resultierenden Sprachausgabe uneinheitlich; sie sinkt mit wachsender Anzahl der Konkatinationsstellendichte.

4.2 Ergebnisse und Bewertung des Präferenztests

Die Präferenzen der VPen bezüglich der vier verschiedenen Stimulikonditionen wurden gemessen, indem gezählt wurde, wie oft eine VP eine Kondition bevorzugte. Hochsignifikante Unterschiede wurden für drei der vier Testkonditionen von allen VPen wahrgenommen. Die Unterschiede zwischen Kondition 1 (nat_orig) und Kondition 3 (nat_manip) waren nicht signifikant. Alle VPen bevorzugten die beiden Konditionen 1 und 3 mit den natürlich gesprochenen Trägersätzen im Vergleich zu den Konditionen 2 und 4 mit den synthetisierten Trägersätzen deutlich. Die Präferenzen für Kondition 1 (nat_orig) und Kondition 3 (nat_manip) variierten bei den VPen; offensichtlich wurden diese beiden Stimulikonditionen als nicht unterschiedlich genug wahrgenommen, um zu einer einheitlichen Präferenz zu gelangen.

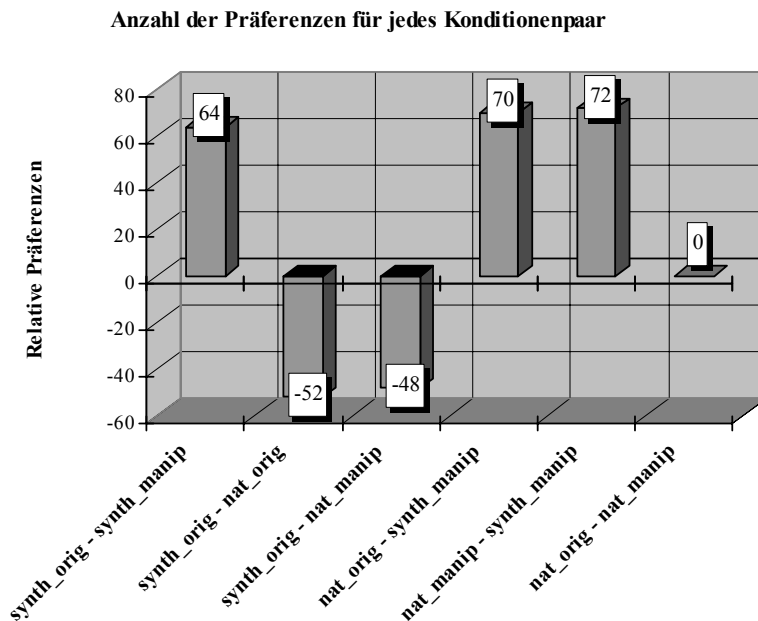


Abbildung 4.4: Anzahl der Präferenzen für jedes Konditionenpaar des Präferenztests. Die Nulllinie zeigt gleiche Wahrnehmung; positive und negative Werte zeigen die Präferenz, bzw. Ablehnungen für bzw. gegen die erste Kondition des jeweiligen Paares.

Die Abbildung 4.4 zeigt deutlich die Bevorzugung der Stimuli in natürlich gesprochenen Trägersätzen der Kondition 1 (nat_orig) und der Kondition 3 (nat_manip). Bei den synthetisierten Trägersätzen wurden die Stimuli der Kondition 4 (synth_manip) mit den dauermanipulierten Trägersätzen deutlich weniger präferiert als die originalen Stimuli der Kondition 2 (synth_orig). Die Ergebnisse des globalen Präferenztests zeigen, dass die zusätzliche Dauer-manipulation der synthetisierten Namen in natürlichen Trägersätzen nicht zu einer negativen Bewertung bzgl. der Annehmlichkeit der synthetischen Stimme führte. Die Ergebnisse zeigen deutlich, dass der im Verbund eingesprochene Trägersatz sich positiv auf die empfundene Annehmlichkeit der gesamten Äußerung auswirkt.

Literatur

- [1] STÖBER, K. (2003) *Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese*, Lang: Frankfurt/M., Sprache, Sprechen und Computer (6).
- [2] BREUER, S.; ABRESCH, J. (2003) „Unit Selection Speech Synthesis for a directory enquiry service“, in: *Proc. ICPHS*, Barcelona.
- [3] SPIEGEL, M.; MACCI, M.; GOLLHARDT, K. (1993) „Synthesis of names by a demisyllable based speech synthesizer.“, in: *Proc. Eurospeech, Berlin*, pp: 279-282.
- [4] SONNTAG, G. (1999). *Evaluation von Prosodie*, Shaker Verlag: Aachen.
- [5] HESS, W.; KRAFT, V.; PORTELE, T. (1994). „Zum Problem der Evaluierung von Sprachsynthesensystemen – dargestellt am Beispiel der Synthesekomponenten in VERBMOBIL“, in: *Fortschritte der Akustik, DAGA 94* (DEGA, Oldenburg), S.: 103-116.
- [6] STEFFENS, S.; STÖBER, K.; HESS, W.; PAULUS, E. (2000) „Anwendungsbezogene Einschätzung und Verbesserung der Qualität synthetischer Sprache“, *Tagungsband DAGA'2000*, Oldenburg, S.:206-207.
- [7] BREUER, S.; ABRESCH, J.; WAGNER, P.; STÖBER, K. (2001) „BLF – Ein Labelformat für die maschinelle Sprachsynthese mit Boss II“, in: Hess, W.; Stöber, K.; (Hrsg.). *Elektronische Sprachsignalverarbeitung*, 12. Konferenz, Studentexte zur Sprachkommunikation 22, Bonn, S.: 100-106.
- [8] BREUER, S.; ABRESCH, J. (2004) „Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis“, in *Proc. ICSLP 2004*, Jeju Island, Korea.
- [9] BENOÎT, C.; EMERARD, F. ; SCHANBEL, B. ; TSEVA, A. (1991). „Quality comparisons of prosodic and of acoustic components of various synthesizers“, in: *Proc. Eurospeech*, Genova, Italy, vol.2, pp: 875-878.
- [10] KRAFT, V.; PORTELE, T. (1995) „Quality Evaluation of Five German Speech Synthesis Systems“, in: *acta acustica* Vol. 3, S.:351-365.
- [11] BELHOULA, A. (1996) *Ein regelbasiertes Verfahren zur maschinellen Graphem-nach-Phonem-Umsetzung von Eigennamen in der Sprachsynthese*, Fortschr.-Ber. Reihe 10, Nr. 432, VDI Verlag: Düsseldorf.