GEMEINSAMKEITEN UND UNTERSCHIEDE IN DEN AUSSAGEN VON VERSUCHSPERSONEN IN EINSCHÄTZUNGSTESTS

Erwin Paulus

Institut für Nachrichtentechnik der TU Braunschweig
e.paulus@tu-bs.de

Abstract: Die Qualität von 180 synthetischen Sprachsignalen wurde von 30 Versuchspersonen auf einer sechsstufigen Skala bewertet. Je drei Signale betrafen einen von 60 Sätzen und waren von drei Varianten eines korpusbasierten Synthesesystems erzeugt worden. Die Versionen waren so beschaffen, daß deutliche Qualitätsunterschiede zu erwarten waren. Aus den Urteilen der Versuchspersonen wurden je ein MOS-Wert für die drei Versionen bestimmt. An anderer Stelle [3] sind Untersuchungen beschrieben, wie gut solche MOS-Werte aus physikalischen Merkmalen der Signale vorhergesagt werden können, und welche Möglichkeiten dadurch bestehen, aufwendige auditive Einschätzungstests durch instrumentelle Verfahren zu ersetzen. Im Gegensatz dazu werden hier die Gemeinsamkeiten und Unterschiede in den Aussagen der Versuchspersonen behandelt und die Frage betrachtet, ob es die "repräsentative Versuchsperson" gibt, aus deren Aussagen sich die kollektiven MOS-Werte vorhersagen lassen. Damit könnten sich künftige auditive Tests auf die Befragung einer einzelnen repräsentativen Versuchsperson beschränken.

1 Einleitung

Die bei der korpusbasierten Synthese erzielbare Qualität der synthetischen Sprachsignale folgt im wesentlichen aus der Vielfalt der im Korpus enthaltenen natürlichen Sprachsignale, der Qualität der diesen Signalen zugeordneten Segmentierungen und Annotierungen und aus der Strategie für die Auswahl von aufeinanderfolgenden Signalsegmenten gemäß dem jeweils zu synthetisierenden Text. Gegebenenfalls führen Änderungen des Korpus oder der Auswahlstrategie zu Änderungen der erzielbaren Sprachqualität, zu deren Beurteilung es dann geeigneter Verfahren bedarf.

Maßgebend sind auditive Tests, bei denen die Urteile möglichst vieler Versuchspersonen über möglichst viele verschiedene synthetische Signale gesammelt werden, um daraus ein Qualitätsmaß abzuleiten. Verbreitet sind sogenannte Einschätzungstests, die als Ergebnis einen MOS (*mean opinion score,* [1]) oder - wenn unter den synthetischen Signalen mehrere Systemvarianten durch Beispiele vertreten sind - MOS für mehrere Systemvarianten liefern. Wünschenswert sind instrumentelle Verfahren, bei denen das Ergebnis von auditiven Tests allein aus meßbaren Signalparametern vorhergesagt werden kann, so daß schließlich die aufwendige Befragung von Versuchspersonen entfallen kann. Allerdings sind die meisten bekannten instrumentellen Verfahren (siehe z. B. [2]) nicht dazu geeignet, die Qualität synthetischer Sprachsignale zu bewerten, da im allgemeinen die Messungen nicht direkt das zu beurteilende Signal betreffen, sondern seine Abweichungen vom zugehörigen - im Fall der Synthese in der Regel gar nicht verfügbaren - natürlichen Original.

Eine Ausnahme wurde vor einiger Zeit in [3] beschrieben. Sie betrifft ein damals in Entwicklung befindliches korpusbasiertes System, und es wurde vorgesehen, für Testzwecke (sowohl für auditive als auch für instrumentelle) nur solche Äußerungen zu synthetisieren, bei denen das zugehörige natürliche Sprachsignal im Korpus enthalten ist. Für die Synthese einer

solchen Äußerung muß dann allerdings das jeweilige natürliche Signal aus dem Korpus herausgehalten werden, so daß für jede zu Testzwecken generierte Äußerung ein anderer, jeweils um ein natürliches Signal reduzierter Korpus maßgebend ist. Auf diese Weise wird es möglich, die Abweichungen zwischen dem synthetischen und dem natürlichen Signal durch meßbare Parameter zu erfassen und aus den Meßergebnissen Schätzwerte für MOS zu gewinnen.

Sowohl auditive Einschätzungstests als auch etwaige instrumentelle Verfahren eignen sich vor allem für die Beurteilung der Auswirkungen von Änderungen an einem in Entwicklung befindlichen System. Für den Vergleich von fertigen Systemen, die womöglich auf vollkommen verschiedenen Ansätzen beruhen, z. B. eine korpusbasierte Synthese gegen eine lautweise regelbasierte Synthese, sind allenfalls auditive Tests geeignet, und zwar auch nur dann, wenn alle zu vergleichenden Systeme in einem gemeinsamen Test durch Beispiele vertreten sind. Die Ergebnisse (MOS), die aus getrennt für einzelne Systeme durchführten Tests stammen, sind für Vergleichszwecke kaum geeignet.

Im Zuge der weiteren Überprüfung des in [3] beschriebenen instrumentellen Verfahrens wurden noch mehrfach umfangreiche auditive Einschätzungstests durchgeführt, deren Ergebnisse dazu dienen sollten, das instrumentelle Verfahren zu „eichen", d. h. die Werte bestimmter Parameter festzulegen (zu „erlernen"), so daß MOS möglichst gut vorhergesagt werden können. Das dabei verwendete korpusbasierte Sprachsynthesesystem ist z. B. in [4] beschrieben und wurde vom Institut für Phonetik und Kommunikationsforschung (IPK) der Universität Bonn zur Verfügung gestellt. Um von vorneherein die Eignung des instrumentellen Verfahrens für vergleichende Untersuchungen verschiedener Systemvarianten und –versionen sicherzustellen, wurden zwei Varianten in je drei Versionen des verwendeten Systems erzeugt. Die zwei Varianten unterscheiden sich in der parametrischen Signaldarstellung der Sprachsignale im Korpus. Die drei Versionen jeder Variante ergaben sich durch einfache Eingriffe in die Auswahlkomponente (siehe unten). Beide Varianten mit ihren je drei Versionen wurden gemeinsam in die auditiven Tests einbezogen. Insgesamt waren also in einem Test sechs verschiedene MOS zu bestimmen.

Für die „beste Version" (b) einer Variante wurden keine Eingriffe vorgenommen, für die „mittlere Version" (m) wurde für jedes Segment des synthetischen Sprachsignals nicht das beste von der Auswahlkomponente angebotene Segment, sondern das zweitbeste verwendet. Auf ähnliche Weise wurde eine „schlechteste Version" (w) durch Verwendung des drittbesten Segments erzeugt. Damit war die Erwartung verbunden, daß im auditiven Test für b der beste und für w der schlechteste MOS resultiert.

Im folgenden werden die Ergebnisse eines Tests analysiert, bei dem beide Varianten gemeinsam einbezogen waren. In den Abschnitten 2 und 3 werden vorerst nur Testergebnisse für eine der beiden Systemvarianten berücksichtigt. Das sind die Urteile von 30 Versuchspersonen über drei synthetisierte Versionen von 60 verschiedenen Sätzen, bzw. 180 Urteile je Versuchsperson und insgesamt 5400 Urteile. In Abschnitt 2 werden einige Einzelheiten zur Durchführung des Tests mitgeteilt. Anschließend wird in Abschnitt 3 untersucht, wieweit sich die Versuchspersonen in ihren Urteilen zu den einzelnen synthetischen Sätzen einig waren. Danach wird in Abschnitt 4 der Frage nachgegangen, ob sich eine „repräsentative Versuchsperson" finden läßt, deren individuelle Urteile schon allein dazu ausreichen, brauchbare Schätzwerte für die von einem Kollektiv von Versuchspersonen gelieferten MOS zu gewinnen. Dabei werden dann zu Prüfzwecken auch die Testergebnisse für die andere Systemvariante - das sind wiederum 5400 Urteile - mitverwendet. Im Erfolgsfall könnte sich ein auditiver Einschätzungstest auf die Befragung einer repräsentativen Versuchsperson beschränken, unabhängig davon, aus welchen Gründen (z. B. zur „Eichung" eines instrumentellen Verfahrens) er notwendig wird. Abschnitt 5 bringt schließlich einen kritischen Ausblick.

2 Besonderheiten des Einschätzungstest

Jede hälfte des Versuchsmaterial bestand aus 180 synthetischen Sprachsignalen, i. e. die drei Versionen b, m, w von 60 verschiedenen Sätzen aus Verbmobil-Szenarien [5][6] mit jeweils zwischen vier und vierzehn Silben. In allen 180 Fällen wurde ein und dieselbe weibliche Stimme eingesetzt. Die akustische Darbietung erfolgte in ruhiger räumlicher Umgebung über Kopfhörer. Die Reihenfolge, in der die Signale zur Beurteilung angeboten wurden, war zufällig; und niemals folgten synthetische Versionen desselben Satzes unmittelbar aufeinander. Die einzelne Darbietung umfaßte zweimal hintereinander das natürliche Sprachsignal gefolgt vom synthetischen Sprachsignal. Die Versuchsperson konnte beliebig oft eine Wiederholung der Darbietung veranlassen, bevor sie ihr Urteil abgab.

Zu beurteilen war die allgemeine Qualität des synthetischen Signals im Vergleich zum natürlichen Signal auf einer sechsstufigen Skala zwischen 1 für „exzellent" und 6 für „schlecht". Die gerade Stufenzahl wurde vorgesehen, um einer etwaigen Vorliebe für ein neutrales mittleres Urteil entgegenzuwirken. (3 auf einer fünfstufigen, 4 auf einer siebenstufigen Skala.) Jede von 30 Versuchsperson durfte sich die Abgabe ihrer Urteile auf mehrere Sitzungen aufteilen, die nicht länger als 20 Minuten dauern sollten, um die Gefahr von störenden Ermüdungserscheinungen gering zu halten. Alle Versuchspersonen waren Studentinnen bzw. Studenten an der TU Braunschweig im Alter von unter 30 Jahren mit Deutsch als Muttersprache.

3 Geschlossenheit der Versuchspersonen

Um einen ersten Eindruck davon zu erhalten, wieweit sich die Versuchspersonen in ihren Urteilen einig sind, wurde untersucht, in wie vielen der 180 Fälle sich eine Mehrheit für 1, 2,...bzw. 6 entscheidet und welchen minimalen und maximalen Umfang diese Mehrheit haben kann. Das Untersuchungsergebnis ist in der Tabelle 1 zusammengefaßt.

Urteil		1	2	3	4	5	6
Fallzahl		9	24	36	39	54	18
einig	min	9	9	9	9	10	9
	max	23	20	19	23	19	30

Tabelle 1: Anzahl der Fälle, in denen das häufigste Urteil 1,2,...bzw. 6 ist, sowie minimale und maximale Häufigkeit dieses Urteils in den betreffenden Fällen.

Die Tabelle zeigt für jedes der sechs möglichen Urteile in der zweiten Zeile die Anzahl der Fälle, und in der dritten und vierten Zeile die Grenzen zwischen denen der Umfang der betreffenden Mehrheit variiert. Zum Beispiel haben sich in 36 der 180 Fälle mindestens 9 und höchstens 19 der 30 Versuchspersonen für das Urteil 3 entschieden. Was das Urteil 6 angeht, sind sich in einem von 9 Fällen, in denen dieses Urteil mehrheitlich gefällt wurde, sogar alle 30 Versuchspersonen einig. Diese umfassende Einigkeit ist allerdings die Ausnahme. In der Regel liegt der maximale Umfang der Mehrheit zwischen 19 und 23. Der minimale Umfang ist nahezu immer 9, in einem Fall mit 10 geringfügig größer. Grob gesprochen wird das mehrheitliche Urteil von mindesten einem Drittel und höchstens von zwei Dritteln der Versuchspersonen getragen.

Es bleibt zu fragen, wie die doch erhebliche Uneinigkeit zwischen den Versuchspersonen zustande kommt. Zerfällt das Kollektiv der Versuchspersonen in mehrere Gruppen derart, daß die Übereinstimmung innerhalb einer Gruppe deutlich höher ist als zwischen Versuchspersonen aus verschiedenen Gruppen? Um diese Frage zu klären, wurden die 180 Urteile jeder Versuchsperson als 180-dimensionale Vektoren bzw. als Punkte in einem 180-di-

mensionalen Raum aufgefaßt und die derart aus 30 Punkten bestehende „Punktwolke" einer einfachen Clusteranalyse unterzogen.

Im allgemeinen dient die Clusteranalyse dazu, die Anzahl der Gipfel (Moden) einer mehrdimensionalen Wahrscheinlichkeitsdichtefunktion zu bestimmen und die gegebenen Punkte den einzelnen Moden zuzuordnen. Im vorliegenden Fall wurde dazu ein graphentheoretisch motiviertes Verfahren [7],[8] herangezogen, das damit startet, jeden Punkt mit seinem nächsten Nachbarn zu verbinden (*single linkage*). Dadurch entstehen Teilgraphen des Minimalen Spannenden Baumes (*minimum spanning tree*) und jeder Teilgraph verbindet eine Untermenge der Punkte, die gleichsam einen „Splitter" aus der Punktwolke darstellt. Die Anzahl der Splitter ist meist deutlich höher als die etwaige Anzahl der Moden, und daher müssen sie meist erst schrittweise so lange zu immer größeren Untermengen zusammengefaßt werden, bis ein geeignetes Abbruchkriterium erfüllt ist. Im vorliegenden Fall hat sich die Zusammenfassung als trivial erwiesen, da sich schon zu Beginn ein einziger überraschend großer Splitter ergab, neben dem nur noch zwei weitere sehr kleine Splitter zu finden sind. Einer davon besteht aus nur zwei Punkten, die unschwer als Ausreißer zu erkennen sind, der andere aus drei Punkten, die allesamt so geringe Entfernung zu Punkten des großen Splitters haben, daß sie ohne weiteres diesem zugeschlagen werden können. Damit ist die Uneinigkeit der Versuchspersonen nicht auf verschieden Moden der Dichtefunktion zurückzuführen; vielmehr ist von einer einzigen Mode auszugehen, und die Uneinigkeit ist durch die „Breite" des zugehörigen Gipfels der Dichtefunktion zu erklären.

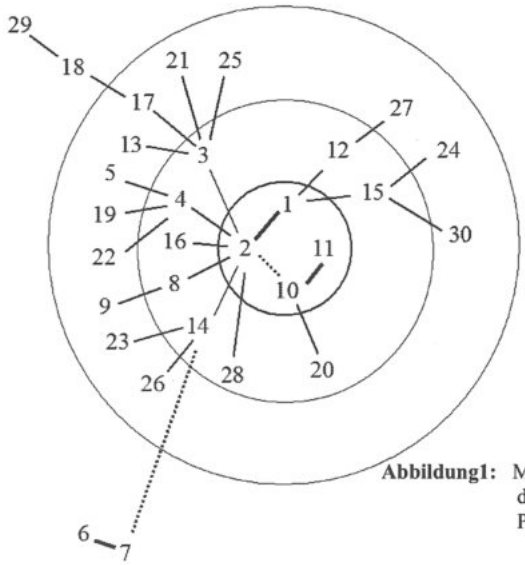

Abbildung1: Minimaler Spannender Baum zwischen den 30 Versuchspersonen (aufgefaßt als Punkte im 180-dimensionalen Raum)

Abbildung 1 zeigt grob schematisch den gesamten Minimalen Spannenden Baum als planaren Graphen. Die „Kerne" der drei oben erwähnten Teilgraphen (Splitter) werden durch diejenigen Kanten des Graphen gebildet, die zwei Punkte, die zueinander nächste Nachbarn sind, verbinden. Im Bild sind das die Kanten zwischen den Punkten 1 und 2, 10 und 11 bzw. 6 und 7. (Die Punkte 6 und 7 sind die beiden schon erwähnten Ausreißer.) Die beiden Kanten zwischen den Punkten 2 und 10 bzw. 7 und 14 vervollständigen die drei schon genannten Teilgraphen zum Minimalen Spannenden Baum. Die im Bild eingetragenen konzentrischen Kreise grenzen Punktmengen voneinander ab, die eine Entfernung von 1, 2 bzw. 3 und größer zum „Kernbereich" haben. Der Kernbereich wird durch den kleinsten, d. h. den innersten Kreis begrenzt. Zum Beispiel hat der Punkt 3 die Entfernung 1 zum Kernbereich,

der Punkt 17 die Entfernung 2, der Punkt 18 die Entfernung 3 und der Punkt 29 die Entfernung 4. Jeder Kante außerhalb des Kernbereichs ist dabei die Länge 1 zugeordnet. Für die Bestimmung des Minimalen Spannenden Baumes war allerdings die Euklidische Distanz zwischen den Punkten maßgebend. Jeder Kante könnte demgemäß auch die Euklidische Distanz zischen den beiden durch sie verbundenen Punkten zugeordnet werden. In der Regel ist das die Distanz eines Punktes zu seinem nächsten Nachbarn. Ausnahmen bilden nur die den Kanten zwischen 2 und 1 und zwischen 7 und 14 zuordenbaren Distanzen. Abbildung 2 zeigt für alle 30 Punkte (Versuchspersonen) die Quadrate der Euklidischen Distanzen zu ihrem jeweiligen nächsten Nachbarn in aufsteigender Reihenfolge. (Diese Reihenfolge ist auch für die Nummern der Punkte bzw. Versuchspersonen in allen Tabellen und Abbildungen maßgebend.) Abbildung 2 zeigt für jeden Punkt (Versuchsperson) auch das Quadrat der Distanz zum jeweils am weitesten entfernten Punkt; dessen Nummer ist in Tabelle 2 zusätzlich zur Nummer des nächsten Nachbarn angegeben.

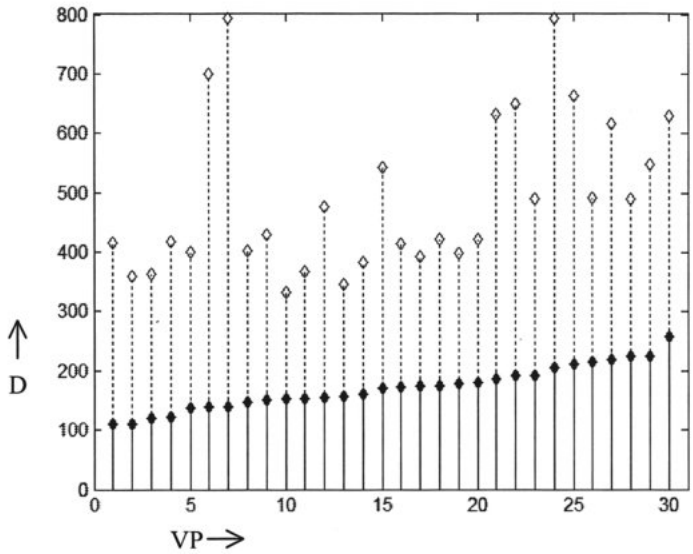

Abbildung 2: Quadratische Entfernungen D zum nächsten Nachbarn und zum weitesten Nachbarn der 30 Versuchspersonen (VP)

4 Suche nach einzelnen repräsentativen Versuchspersonen

Indem sich das Kollektiv der Versuchspersonen gleichsam als Stichprobe (hier vom Umfang 30) für eine Zufallsgröße mit breiter eingipfeliger Verteilungsdichtefunktion herausgestellt hat, bestehen gute Aussichten dafür, in dieser Stichprobe auch „repräsentative" Versuchspersonen zu finden, deren „individuelle MOS" mit den entsprechenden „kollektiven MOS" gut übereinstimmen. (Ein individueller MOS berücksichtigt nur die Urteile einer einzelnen Versuchsperson, ein kollektiver die Urteile aller Versuchspersonen.) Repräsentative Versuchspersonen stammen gleichsam aus dem Gipfelbereich der Dichtefunktion, und da dort die Dichte naturgemäß hoch ist, darf erwartet werden, daß sogar mehrere davon gefunden werden können. Es ist zu vermuten, daß die gesuchten Versuchspersonen am ehesten in dem in Abbildung 1 gezeigten Kernbereich zu finden sind. Als Kriterium für die Suche wurde aber nicht diese besondere Lage verwendet, sondern die mittlere absolute Abweichung e der 180 individuellen Urteile jeder Versuchsperson von den jeweils zugehörigen 180 mittleren (kollektiven) Urteilen aller 30 Versuchspersonen. Im allgemeinen werden MOS auch dann

schon als brauchbar angesehen, wenn sie sich auf die Urteile von weniger als 30 Versuchspersonen stützen. Die aus den vorliegenden Urteilen bestimmbaren MOS sollten daher brauchbar sein und können als repräsentativ für alle MOS gelten, die bei vergleichbaren Einschätzungstest mit anderen Kollektiven von Versuchspersonen (kleiner oder größer als 30) zu gewinnen sind.

VP	NN	WN
1	2	29
2	1	7
3	2	7
4	2	7
5	4	7
6	7	24
7	6	24
8	2	7
9	8	7
10	11	7
11	10	7
12	1	7
13	3	30
14	2	24
15	1	7
16	2	30
17	3	7
18	17	30
19	4	7
20	10	7
21	3	7
22	4	7
23	14	24
24	15	7
25	3	7
26	14	24
27	12	7
28	2	7
29	18	24
30	15	7

Tabelle 2: Nächster Nachbar (NN) und weitester Nachbar (WN) der 30 Versuchspersonen (VP)

VP	e	f
1	0.56	0.63
2	0.49	0.48
3	0.47	0.49
4	0.54	0.67
5	0.57	0.54
6	0.93	0.82
7	1.06	1.02
8	0.57	0.56
9	0.60	0.61
10	0.55	0.57
11	0.53	0.50
12	0.59	0.55
13	0.55	0.57
14	0.58	0.58
15	0.67	0.58
16	0.67	0.60
17	0.59	0.63
18	0.65	0.62
19	0.62	0.59
20	0.66	0.61
21	0.72	0.76
22	0.73	0.67
23	0.77	0.75
24	0.88	1.01
25	0.76	0.78
26	0.79	0.74
27	0.76	0.69
28	0.77	0.87
29	0.89	0.88
30	0.86	0.72

Tabelle 3: Individuelle mittlere absolute Abweichung e bzw. f vom kollektiven Mittelwert für jeweils 180 Sätze

Je kleiner die mittlere absolute Abweichung e ist, desto eher kann eine Versuchsperson als repräsentativ gelten. Der Wert 0 für diese Abweichung darf aber nicht als Kriterium verwendet werden, da die mittleren Urteile aller Versuchspersonen im allgemeinen keine ganzen Zahlen sind. Im Sinne der Rundung scheint die Forderung nach einer Abweichung unter 0,5 sinnvoll zu sein. Tabelle 3 zeigt für jede Versuchspersonen den zugehörigen Wert von e in der zweiten Spalte. Nur für die Versuchspersonen 2 und 3 liegt dieser Wert mit 0.49 bzw. 0,47 knapp unterhalb von 0,5. Damit sind zumindest zwei möglicherweise repräsentative Versuchspersonen gefunden, von denen nur eine (Versuchsperson 2) aus dem Kernbereich ist. Eine scharfe Grenze zwischen repräsentativ und nichtrepräsentativ ist aber nicht zu erkennen und darf – angesichts des „breiten" Gipfels der oben erwähnten Dichtefunktion - gar nicht erwartet werden. Die untersuchte mittlere absolute Abweichung ist innerhalb des Kollektivs der Versuchspersonen quasi kontinuierlich veränderlich. (Von den schon früher erwähnten Ausreißern 6 und 7 abgesehen.) So liegen die Werte für die drei übrigen Versuchspersonen

aus dem Kernbereich (1, 10, und 11) mit 0.56, 0,55 bzw. 0,53 nur knapp über 0,5 und auch bei der Versuchsperson 13 wird 0,5 mit 0,55 nur knapp überschritten. Wenn die Grenze mit 0,55 angenommen wird, dann sind die Versuchspersonen 2, 3, 10, 11 und 13 möglicherweise repräsentativ.

Alle bisherigen Überlegungen stützen sich nur auf eine Hälfte des gesamten verfügbaren Versuchsmaterials. Um die oben getroffene Auswahl von möglicherweise repräsentativen Versuchspersonen zu festigen, kann nun die andere Hälfte herangezogen werden. Die dritte Spalte der Tabelle 3 entspricht genau der zweiten, außer daß für die mittlere absolute Abweichung f nun die für die andere Systemvariante abgegebenen Urteile der Versuchspersonen maßgebend sind. Bei den Versuchspersonen 2 und 3 bleibt f ebenso wie e knapp unter 0,5. Bei Versuchsperson 11 wird in beiden Fällen die oben angenommene obere Grenze 0,55 unterschritten. Die Versuchspersonen 2, 3 und 11 sind damit als repräsentativ gefestigt. Bei strenger Anwendung der oberen Grenze können die Versuchspersonen 10 und 13 nicht als repräsentativ angenommen werden, da in beiden Fällen f mit 0,57 diese Grenze überschreitet. Sie werden aber trotzdem auch für die nächste Prüfung zugelassen.

Die von repräsentativen Versuchspersonen stammenden individuellen Urteile müssen jeweils für sich allein genommen schon dafür ausreichen, gute Schätzwerte für MOS zu gewinnen, die aus den Urteilen aller Versuchspersonen abgeleitet worden sind. Das kann nun an den jeweils für die drei Versionen b, m, w der 60 betrachteten Sätze überprüft werden. Tabelle 4

	Variante 1				Variante 2			
VP	b	m	w	E	b	m	w	F
2	2.68	4.30	4.42	0.09	2.73	4.33	4.78	0.12
3	2.75	4.23	4.43	0.11	2.60	4.50	5.15	0.19
11	2.58	4.28	4.87	0.13	2.42	4.45	5.13	0.16
10	2.87	4.30	4.75	0.10	2.37	3.97	4.55	0.27
13	2.85	4.38	4.77	0.11	2.28	4.28	4.88	0.08
MOS	2.71	4.32	4.64		2.51	4.30	4.88	

Tabelle 4: Mittelwerte der Urteile getrennt nach Versuchspersonen (VP) und Versionen b, m, w, sowie die insgesamte Abweichung E bzw. F vom jeweiligen MOS für zwei Systemvarianten

zeigt in der letzten Zeile die (kollektiven) MOS für die drei Versionen b, m, w für die beiden betrachteten Varianten des Synthesesystems. Jeder dieser MOS ist ein Mittelwert über 1800 Urteile (von 30 Versuchspersonen über jeweils 60 Sätze) . In den darüberliegenden Zeilen stehen die individuellen MOS der als repräsentativ angenommenen Versuchspersonen 2, 3 und 11 und zusätzlich auch die der Versuchspersonen 10 und 13. Jeder individuelle MOS ist ein Mittelwert über 60 Urteile einer einzelnen Versuchsperson. Neben den individuellen MOS ist auch die mittlere absolute Abweichung E bzw. F der drei individuellen MOS von den zugehörigen kollektiven MOS angegeben. Die Einträge für F können wieder zum Zweck der Überfürüfung verwendet werden. Bei Anwendung eines strengen Maßstabs können 2, 11 und 13 dabei als bestätigt gelten, während 3 und vor allem 10 weniger gut abschneiden. Zusammen-fassend läßt sich sagen, daß zumindest die Versuchspersonen 2, 11 und 13 eine Vorhersage kollektiver MOS mit einer mittleren Genauigkeit von ± 0,1 erlauben.

5 Ausblick

Für 30 zufällig für Einschatzungstests verfügbare Personen wurden erhebliche Unterschiede in den abgegebenen Urteilen festgestellt. Das Kollektiv dieser Versuchspersonen erwies sich aber als unstrukturiert; d. h. es konnten keine Gruppen gefunden werden, deren Urteile jeweils systematisch anders ausfallen als die Urteile der anderen Gruppen. Das wurde zum Anlaß genommen, nach einzelnen repräsentativen Versuchspersonen zu suchen, deren individuelle Urteile weitgehend mit den über das Kollektiv gemittelten Urteilen übereinstimmen. Tatsächlich wurden mehrere Versuchpersonen gefunden, die als repräsentativ geeignet erscheinen; umsomehr als ihre Eignung an Versuchsergebnissen überprüft wurde, die noch nicht für die Suche verwendet worden waren. Die zur Überprüfung verwendete Daten haben allerdings den Schönheitsfehler, daß sie von genau denselben Versuchspersonen stammen, von denen auch die für die Suche verwendeten Daten geliefert wurden, und daß auch die den synthetischen Signalen zugrundeliegenden Sätze (nicht die synthetischen Signale selbst!) sowohl bei der Suche als auch bei der Überprüfung verwendet wurden. Außerdem ist nur eine einzige Stimme für alle Signale eingesetzt worden.

Bis zu einem gewissen Grad ließen sich die Schönheitsfehler auch ohne neuerliche Einschätzungstests verringern. Um eine nach „Lern-„ und „Testdaten" getrennte Auswertung der vorhandenen Daten zu ermöglichen, könnte das Kollektiv der Versuchspersonen in zwei Gruppen gespalten werden, wobei nur eine Person, nämlich diejenige, die als möglicherweise repräsentativ angesehen wird, beiden Gruppen angehören muß. Ebenso ließe sich die Menge der Sätze in zwei disjunkte Teilmengen aufspalten, und auch die Struktur der Menge der zugehörigen synthetischen Signale ließe sich für Lern- und Testzwecke unterschiedlich gestalten, z. B. indem einerseits die b-Version aus Variante 2 mit den m- und w-Versionen der Variante 1 und andererseits die b-Version aus Variante 1 mit den m- und w-Versionen der Variante 2 kombiniert werden.

Mehrere Stimmen könnten dagegen nur durch neuerliche Einschätzungstests berücksichtigt werden, und nur unter der Bedingung, daß das Synthesesystem mehrere Stimmen bereitstellt. Offen bleibt auch die Frage, ob eine auf dem hier beschriebenen Weg gefundene repräsentative Versuchsperson auch für andersartige Einschätzungstests – z. B. mit 5 oder 7 Stufen oder ohne Vergleichsmöglichkeit mit dem natürlichen Sprachsignal - als repräsentativ gelten kann.

Literatur

[1] ITU_T Recommendation P.800, ITU, Genf 1996

[2] Möller, S.: Assessment and Predediction of Speech Quality in Telecommunications. Kluwer Academic Publishers 2000.

[3] Paulus, E., und J. Steffens: Instrumentelle und auditive Qualitätsbeurteilung bei der Korpusbasierten Sprachsynthese. In: Kristian Kroschel (Hrsg.), Elektronische Sprach-signalverarbeitung . Tagungsband der ESSV '03, Karlsruhe, Sept. 2003, 211 – 218

[4] Stöber, K.: Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese. Sprache, Sprechen und Computer, Bd. 6, Peter Verlag 2003

[5] Wahlster, W. (Ed): Verbmobil - Foundations of Speech-to-Speech-Translation. Springer Verlag 2000

[6] Steffens, J., and E. Paulus: Speech Synthesis Quality Assessment . In: [5], 592 – 598

[7] Zahn,C: Graph-Theoretical Methods forDetecting and Describing Gestalt Clusters. IEEE Trans. Comp. 20/1, Jan. 1971, 68-86

[8] Paulus, E: Nonsupervised Pattern Classification Based on the Nearest Neighbour Rule In: Proc. 1st Int. Conf. Pattern Recognition, Washington, D. C. , 1973, 546 - 549