Zwei hybride Unit Selection-Strategien im Vergleich

Martin Barbisch, Bettina Säuberlich, Antje Schweitzer
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland
Email: {martin.barbisch,bettina.saeuberlich,antje.schweitzer}@ims.uni-stuttgart.de

Einleitung

Dieser Beitrag beschreibt zwei Varianten einer Unit Selection-Sprachsynthese, die im Rahmen von SmartWeb implementiert und eingesetzt wurden. SmartWeb ist ein vom BMBF gefördertes Forschungsprojekt, bei dem Anfragen an das *Semantic Web* intuitiv, d.h. durch natürliche Sprache, ggf. kombiniert mit Gestik, gestellt werden können. Die Antwort erfolgt ebenfalls natürlichsprachlich, und wird akustisch durch die hier vorgestellte Synthese präsentiert.

Im Laufe des Projekts wurde alternativ zu einem früher implementierten Auswahlverfahren [2] ein neues Verfahren entwickelt, so dass nun bei gleicher Datenbank, gleicher Textvorverarbeitung und gleicher symbolischer Synthese zwei unterschiedliche Auswahlverfahren zur Verfügung stehen. Beide Varianten liefern subjektiv sehr gute Qualität bei sehr guter Verständlichkeit. Die beiden Varianten wurden mit Hilfe eines Perzeptionsexperiments miteinander verglichen.

Der Beitrag ist wie folgt gegliedert. Zunächst werden die beiden Auswahlverfahren genauer beschrieben. Anschließend werden Ergebnisse einer vergleichenden Evaluierung der beiden Ansätze präsentiert, gefolgt von einer kurzen Zusammenfassung.

Die Auswahlverfahren im Vergleich

Die beiden hier vorgestellten Auswahlverfahren sind hybride Verfahren, die Aspekte zweier existierender Unit Selection-Ansätze verbinden, nämlich des sog. *Phonological Structure Matching* (PSM, [3]) und des *Automatic Clustering* (AC, [1]). Im Folgenden werden zunächst PSM und AC kurz erläutert, bevor die beiden hybriden Auswahlverfahren genauer beschrieben werden.

Beim PSM wird für jede zu synthetisierende Äußerung top-down von der Phrasenebene über die Wort- und die Silbenebene bis zur Segmentebene in der Datenbank nach passenden Kandidaten gesucht. Für alle Einheiten, für die keine Kandidaten gefunden wurden, wird auf der nächstniedrigeren Ebene weiter gesucht. PSM garantiert so die Auswahl zusammenhängender längerer Einheiten aus der Datenbank, sofern diese vorhanden sind.

Dagegen wird beim AC immer auf der Segmentebene nach Kandidaten gesucht; längere zusammenhängende Stücke aus der Datenbank werden nur indirekt dadurch begünstigt, dass sie keine Verkettungskosten verursachen. Um die Kandidatenmengen einzugrenzen, werden beim AC offline alle Segmente in Cluster zusammengefasst, und zwar in einer automatischen Prozedur, die für jedes Phonem einen Entscheidungsbaum erstellt, dessen Blätter die Cluster repräsentieren, und an dessen Knoten Eigenschaften des linguistisch-phonologischen Kontexts abgefragt werden. Dies geschieht derart, dass die akustische Ähnlichkeit der Realisierungen innerhalb eines Cluster maximiert wird. Es werden im Prinzip diejenigen linguistisch-phonologischen Kontexteigenschaften ermittelt, die den größten Einfluss auf die akustische Realisierung haben. Die Cluster können später durch Pruning verkleinert werden, indem Segmente, die akustisch relativ zum Mittelpunkt des Clusters aus dem Rahmen fallen, ausgeschlossen werden. Online wird dann bei der Suche nach geeigneten Kandidaten für ein bestimmtes Segment einer zu synthetisierenden Äußerung zunächst das Cluster bestimmt, das dem gewünschten linguistisch-phonologischen Kontext entspricht. Die Realisierungen innerhalb dieses Clusters werden als Kandidaten übernommen.

Der Vorteil einer Kombination der beiden Algorithmen liegt darin, dass PSM die Auswahl längerer Einheiten begünstigt und so die Anzahl der Verkettungsstellen reduziert, während AC geeignet ist, die z. T. sehr großen Kandidatenmengen einzuschränken. Große Kandidatenmengen ergeben sich erfahrungsgemäß hauptsächlich auf der Segmentebene. Auf Silben-, Wort- und Phrasenebene werden die Kandidatenmengen seltener so groß, da deutlich mehr unterschiedliche Einheiten des betreffenden Einheitentyps in der Datenbank vorhanden sind[1]. Daher wird bei der ursprünglich implementierten Variante (im Folgenden PSM/AC genannt) PSM auf Silben-, Wort- und Phrasenebene eingesetzt, weil auf diesen höheren Ebenen seltener große Kandidatenmengen zu erwarten sind, die die Effizienz des Systems beeinträchtigen könnten. Auf Segmentebene jedoch wird AC verwendet.

Beim alternativ entwickelten Verfahren erfolgt die Einheitensuche wie bei PSM top-down, und es werden ebenfalls Entscheidungsbäume eingesetzt, um die Einheiten in der Datenbank zu Clustern zusammenzufassen. Allerdings wird die Clustertechnik auf allen Ebenen, nicht nur auf der Segmentebene, verwendet, und die Struktur der Entscheidungsbäume wird für jeden Einheitentyp von Hand vorgegeben. Dieser Ansatz wird im Folgenden Manual Clustering (MC) genannt.

Der Vorteil des MC gegenüber dem AC liegt darin, dass beim AC in manchen Fällen die Auswahl von

[1] Im vorliegenden Korpus gibt es z. B. bei ca. 107 000 Segmenten nur 84 Phoneme (Unterscheidung glottalisierter und nichtglottalisierter Vokalvarianten sowie Integration von fremdsprachlichen Phonemen). Dies entspricht einer durchschnittlichen Anzahl von 1 274 Tokens pro Type, während auf Silbenebene auf ca. 41 000 Tokens 3 350 Types kommen, d. h. ca 12,2 Tokens pro Type.

in der Datenbank aufeinander folgenden (und deshalb störungsfrei zu verkettenden) Segmenten schon deshalb ausgeschlossen ist, weil sie entweder einem Cluster zugeordnet wurden, das im vorliegenden Kontext nicht berücksichtigt wird, oder weil sie beim Prunen entfernt wurden. Zudem fließt bei der Erstellung der Entscheidungsbäume kein phonetisches Wissen ein. Es ist z. B. im Gegensatz zum MC nicht möglich, bestimmten linguistisch-phonologischen Eigenschaften, deren Relevanz für die Verkettung bekannt ist (Koartikulationseinflüsse bestimmter angrenzender Phoneme etc.) höhere Priorität zuzuweisen.

Außerdem macht beim MC der generelle Einsatz von Entscheidungsbäumen auf allen Einheitenebenen eine einheitliche Verwaltung der Einheiten in einer Baumstruktur und den effizienten Zugriff auf die Einheiten durch Indexierung über die einzelnen Baumebenen möglich. Die Struktur der Bäume ist insofern vorgegeben, als auf jeder Ebene des Baums ein Merkmal des linguistisch-phonologischen Kontexts abgefragt wird (z. B. der Artikulationsort des nachfolgenden Phonems oder die Silbenbetonung); die Reihenfolge dieser Merkmale wird für jeden Einheitentyp manuell vorgegeben.

Das MC Verfahren ist sehr flexibel. Zum einen ist es sehr leicht möglich, den Suchbaum umzustrukturieren, falls sich eine Merkmalsreihenfolge als nicht optimal herausstellt. Außerdem kann zwischen Segmenten, Diphonen und Demiphonen als Basiseinheit gewählt werden. Es kann zur Laufzeit ein Schwellwert für die Mindestanzahl von Kandidaten pro Cluster vorgegeben werden, da die Suche im Baum an der Stelle abgebrochen wird, an der die Gesamtzahl der in den darunter liegenden Clustern vorhandenen Kandidaten das vorgegebene Minimum unterschreiten würde. In diesem Fall werden alle Kandidaten der betreffenden Cluster berücksichtigt.

Die Vorteile des MC sind seine hohe Flexibilität und die einheitliche Behandlung aller Einheitstypen. Im Gegensatz zum AC werden außerdem durch das Clustern nicht von vornherein bestimmte Kandidaten ausgeschlossen. Zudem fließt bei der manuellen Erstellung der Entscheidungsbäume phonetisches Wissen mit ein.

Evaluierung

Die beiden Auswahlverfahren wurden in einem Perzeptionsexperiment miteinander verglichen[2]. Dabei wurden 26 Probanden (17 m/9 w) jeweils insgesamt 30 synthetisierte Stimulus-Paare vorgespielt, die bewertet werden sollten. Die Bewertungen „A klingt besser als B“, „B klingt besser als A“ und „A klingt gleich gut wie B“ waren möglich. Mehrmaliges Anhören war erlaubt, und es gab kein Zeitlimit für die Beurteilung. Das Experiment wurde unbeaufsichtigt durchgeführt. Um eine Einschätzung der Zuverlässigkeit der Beurteilungen zu erhalten, waren die Stimuli in 8 Stimuluspaaren identisch, ansonsten waren für jedes Paar die Stimuli durch die beiden unterschiedlichen Verfahren generiert worden. Die Abfolge innerhalb eines Paares war zufällig (AB oder BA). Einige Paare wurden jedem Probanden in beiden Abfolgen präsentiert. Den Teilnehmern war bekannt, dass die Stimuli in manchen Fällen identisch waren.

Die Ergebnisse zeigen einen leichten Vorteil für die PSM/AC-Stimme (49,8% PSM/AC vs. 40,7% MC vs. 9,4% unentschieden, prozentuale Angaben bezogen auf die 572 Bewertungen 22 unterschiedlicher Stimuluspaare). Ein χ^2-Test bestätigt die Signifikanz ($\chi^2(1,N=572)=154.0315$, $p \ll 0.05$). Eine genauere Inspektion zeigt, dass der Effekt weniger auf persönlichen Vorlieben der Probanden beruht: ein signifikanter Unterschied in der Bewertung zeigt sich nur bei 3 Probanden, die die PSM/AC-Stimme besser bewertet haben, bei Anpassung des Signifikanzniveaus auf $p = 0.05/26 \simeq 0.002$ auf Grund der 26 unabhängigen χ^2-Tests. Hingegen ist die Bewertung eher abhängig von den jeweiligen Stimuluspaaren; hier finden sich signifikante Unterschiede in der Bewertung bei 16 von 22 Stimuluspaaren, bei einem Signifikanzniveau von $p = 0.05/22 \simeq 0.002$. Dabei wurde 6 mal die MC-Stimme und 10 mal die PSM/AC-Stimme besser bewertet. Dies bedeutet, dass sich bei der Mehrheit der Stimuluspaare die Probanden in ihren Bewertungen relativ einig waren - es wurde meist dieselbe Stimme bevorzugt. Dies scheint daran zu liegen, dass in einzelnen Fällen suboptimale Einheiten in der Datenbank kleine Diskontinuitäten bei der Verkettung verursachen. Die betreffenden Einheiten werden aber nicht immer von beiden Verfahren ausgewählt, so dass sie zufällig in der einen oder in der anderen Variante auftreten.

Zusammenfassung und Ausblick

Es wurden zwei hybride Auswahlverfahren für Unit Selection-Synthese vorgestellt. Eine perzeptive Evaluierung zeigte, dass der Unterschied zwischen beiden Varianten in der momentanen Implementierung nicht sehr deutlich ist, wenn auch die PSM/AC-Stimme etwas besser bewertet wird. Als nächster Schritt erscheint eine diagnostischere Evaluierung sinnvoll, so dass die speziellen Stärken und Schwächen der beiden Varianten genauer ermittelt werden können.

Literatur

[1] Black, A. W. und Taylor, P.: Automatically clustering similar units for unit selection in speech synthesis. Eurospeech Proceedings (1997), 601–604.

[2] Schweitzer, A., Braunschweiler, N., Dogil, G., Klankert, T., Möbius, B., Möhler, G., Morais, E., Säuberlich, B., Thomae, M.: Multimodal speech synthesis. In W. Wahlster: SmartKom: Foundations of Multimodal Dialogue Systems, 411–435, Springer, 2004.

[3] Taylor, P. und Black, A. W.: Speech synthesis by phonological structure matching. Eurospeech Proceedings (1999), 623–626.

[2] Herzlichen Dank an Andreas Madsack, Sylvia Riester und Christian Scheible für die Durchführung des Experiments.