RELIABILITÄTSMASSE FÜR DIE AUTOMATISCHE TRANSKRIPTION

REINHOLD GREISBACH

Institut für Phonetik
Universität zu Köln
Greinstr. 2, D-5000 Köln 41

ZUSAMMENFASSUNG

Die Beurteilung der Leistungsfähigkeit von automatischen Transkriptionsverfahren verlangt nach Methoden, die die Validitätsproblematik bei phonetischen Transkriptionen berücksichtigen können. Reliabilitätsmaße scheinen diese Voraussetzung zu erfüllen. Bereits im Hinblick auf zukünftige Entwicklungen wird deshalb hier die Anwendbarkeit solcher Maße auf automatisch erstellte Transkripte theoretisch untersucht.

MOTIVATION

Bereits heute ist es möglich, akustische Sprachsignale automatischen Spracherkennungsprozessen zu unterziehen mit dem Ziel, eine segmentale phonetische (impressionistische) Transkription dieses Sprachsignals zu gewinnen. Dabei läßt jedoch meist ein einziger Blick auf das Resultat erkennen, wie gut oder besser wie schlecht diese automatische Transkription (aT) arbeitet. Wenn sich jedoch in Zukunft die Leistungsfähigkeit automatischer Erkennungsverfahren weiter verbessert, und daran scheint kein Zweifel, wird die Entscheidung "Transkribiert die Maschine richtig?" nicht mehr durch einfachen Augenschein zu treffen sein.
Bei der automatischen Spracherkennung läßt sich i.a. sehr leicht entscheiden, ob die automatische Erkennung gelingt oder nicht. Dazu wird das Resultat des automatischen Prozesses mit den Höreindrücken einer menschlichen Hörergruppe verglichen. Sehr oft besteht diese Hörergruppe nur aus einer einzigen Person, denn es darf hier vorausgesetzt werden, daß auch jeder andere (muttersprachliche) Hörer das gleiche wahrnehmen würde. Für die Prüfung eines automatisch erstellten Transkripts, das auch mit den Höreindrücken einer menschlichen Hörergruppe verglichen werden muß, ist gerade diese Voraussetzung nicht erfüllt. Denn bei hinreichend enger phonetischer Notation und hinreichend langem Sprachsignal wird unter zwei Transkribenten wohl niemals völlige Einigkeit über das Gehörte herrschen. Dies bedeutet aber, daß die Richtigkeit eines automatischen Transkripts, anders als im Falle der automatischen Spracherkennung, nicht an einem eindeutigen Muster geprüft werden kann. Die Überprüfbarkeit eines automatisch erstellten Transkripts hängt also von der Richtigkeit (Validität) manuell erstellter Vergleichstranskripte ab.
Vor der genaueren Untersuchung der Validitätsproblematik betrachten wir zunächst, wie ein phonetisches Transkript entsteht, und kennzeichnen damit zugleich den Begriff "Transkription", wie er hier verstanden werden soll. Unsere (wohl für die deutschsprachige Phonetik typische) Transkription geht von einem akustischen Sprachsignal aus, welches mithilfe der artikulatorisch definierten IPA-Symbole notiert wird. Der Transkribent verläßt sich dabei ausschließlich auf sein Gehör (auditive Transkription).
Diese Form der Transkription birgt nun eine Besonderheit in sich. Bei der auditiven Transkription mit IPA-Symbolen müssen mögliche artikulatorische Abweichungen des Sprechers kompensiert werden (so erscheint bei der Transkription der Äußerung eines Bauchredners die artikulatorisch nicht vorhandene Lippenrundung eines auditiv wahrgenommenen [y] aufgrund der artikulatorischen Definition dennoch im Transkript.) Notiert werden kann also immer nur ein vorgestellter, ideal artikulierender Sprecher. Umgekehrt bedeutet dies, daß selbst dann, wenn sich die Artikulation des Sprechers (z.B. mittels Röntgenfilm) optisch beobachten läßt, dies noch keine direkte Überprüfung der Notation gestattet.
Was ist also - bei auditiver Transkription mit IPA-Symbolen - eine richtige Notation? Die Antwort ergibt sich durch eine Grenzüberlegung: Wenn für einen Laut jeder Transkribent zu jeder Zeit das gleiche Symbol verwendet, so ist diese Notation richtig (valide)! Die Realität kann jedoch nur aus einer Stichprobe bestehen, die idealerweise verschiedene Transkribenten zu verschiedenen Zeiten von einer Sprachaufnahme durchführen. Und gleichfalls wird diese Stichprobe i.a. nicht für jeden Laut das gleiche Symbol enthalten. Den Grad der Übereinstimmung in einer solchen Stichprobe nennt man Reliabilität, den Grad der Verschiedenheit Variabilität. Reliabilität darf aber grundsätzlich nicht mit Validität verwechselt werden, denn eine maximal reliable auditive IPA-Notation eines von einem menschlichen Sprecher produzierten Sprachsignals ist nur deshalb valide, weil sie auf keine andere Weise als durch das Ohr direkt überprüfbar ist. Bei Notation in anderen Alphabeten, z.B. dem analphabetischen von Jespersen, kann jedoch eine auditive Notation aufgrund von optischen Informationen direkt verifiziert werden (dann nämlich, wenn die artikulatorischen Beschreibungsdimensionen des Alphabets besser den tatsächlichen Bewegungsdimensionen des Artikulationsapparats entsprechen, als dies beim IPA-Alphabet der Fall ist). Für eine eingehendere Diskussion dieser Fragen vgl. /1/. Prinzipiell ist also nur die Reliabilität eines Transkripts feststellbar. Ob damit auch seine Validität bestimmt ist, hängt offenbar vom gewählten Transkriptionsalphabet ab. Für die Bewertung eines automatisch erstellten Transkripts muß die Reliabilität nun zu einer quantifizierbaren Größe werden, zu einem Reliabiliätsmaß.

ÄHNLICHKEITSMASSE

Reliabilitätsmaße oder allgemeiner Reliabilitätsmessungen basieren üblicherweise auf einer numerischen Bewertung der Differenzen zwischen den Symbolen des jeweiligen Transkriptionsalphabets, sog. Ähnlichkeitsmaßen (Ä-Maßen). Anders als den aus dem täglichen Leben vertrauten Maßen der physikalischen Umwelt fehlen den meisten dieser Maße jedoch die (mathematischen) Eigenschaften, die die Bewertung von physikalischen Meßergebnissen einfach gestalten. Welche Eigenschaften ein Maß besitzt, welchem Skalentyp es zuzuordnen ist (wie man in der psychologischen Testtheorie sagt), bestimmt in der Phonetik zuvorderst die Vorstellung, die der jeweilige Phonetiker von den Beziehungen der Symbole untereinander besitzt.
Bei den einfachsten Ä-Maßen sind alle Symbole des Alphabets gleichberechtigt. Sie stehen ohne erkennbare Ordnung nebeneinander, was letztlich bedeutet, daß die Differenz zwischen allen Symbolen gleich groß ist (Nominalskala). Kompliziertere Maße setzen eine Ordnung zwischen den Symbolen voraus, z.B., daß die Differenz zwischen [i] und [e] kleiner ist als die zwischen [i] und [a](Ordinalskala). Bei einer Intervallskala lassen sich darüberhinaus die Differenzen zwischen den Symbolen vergleichen. So ist z.B. bei den Kardinalvokalen der Unterschied zwischen Kardinal-[e] und -[ɛ] definitionsgemäß genauso groß wie der zwischen Kardinal-[ɛ] und - [a].
Um zu einem Zahlenwert zu gelangen, werden zunächst auf der Basis dieser Skalen (ab dem Ordinalskalenniveau) mehrdimensionale Räume konstruiert und die Symbole darin angeordnet. Bei der Konstruktion dieser Räume lassen sich zwei Hauptverfahren feststellen. Der eine Verfahrenstyp geht von den artikulatorischen Klassifikationsdimensionen des Alphabets aus und spannt den Raum entlang dieser Dimension zumeist orthogonal auf (nicht orthogonal z.B. als "Vokaldreieck"). Der andere, aufwendigere Verfahrenstyp erzeugt den Raum mittels auditiver Dimensionen, die nach Hörtests mit Versuchspersonen durch statistische Methoden wie z.B. MDS oder Faktorenanalyse gewonnen werden. Das Ä-Maß gibt dann den Abstand zweier Symbole in diesen Räumen an (mit einer meist heuristisch gewonnenen Abstandsfunktion).
Welcher der Skalentypen ist nun aber für die Transkription der richtige? Die Literatur dokumentiert hier verschiedene Meinungen, wobei die Befürworter des "transkriptorischen Messens" i.a. auf dem Intervallskalenniveau stehen, während sich seine Gegner (konsequenterweise) auf das Nominalskalenniveau zurückziehen (müssen). Alle bekannten Reliabilitätsuntersuchungen auf IPA-Basis (/2/,/3/,/4/,/5/,/6/,/7/,/8/) benutzen Ä-Maße zumindest auf Intervallskalenniveau.

RELIABILITÄTSMASSE

Während die erste quantitative Reliabilitätsuntersuchung zur Transkription bereits zu Anfang dieses Jahrhunderts stattfand /9/, verwenden erst die Arbeiten der '80er Jahre den Begriff "Messen der Reliabilität" bzw. "Reliabilitätsmaß" (R-Maß), der hier (wie in der psychologischen Testtheorie) als Maß für den Grad der Übereinstimmung je zweier Beobachter (Transkribenten) verstanden wird. Dieser "Korrelationskoeffizient" zweier Transkribenten ergibt sich als die gewichtete Summe aller Unterschiede zwischen den beiden (Stichproben-)Transkripten, gemessen mit dem jeweiligen symbolbezogenen Ä-Maß. Mit einer solchen Messung soll die Befähigung eines Transkribenten für eine transkriptorische Aufgabe festgestellt werden. Er gilt dann als befähigt, wenn seine "Reliabilitäts-Korrelation" zu einer größeren Gruppe /8/ oder zu einem "master-transcriber" /7/ einen bestimmten Grenzwert übersteigt.
Der Transkribent wird damit also zu einem Meßinstrument, dessen Reliabilität (Zuverlässigkeit) meßbar ist. Dieser "Entmenschlichung" des Wissenschaftlers mag es wohl hauptsächlich zuzuschreiben sein, daß Kritik an solchen Reliabilitätsuntersuchungen laut wird /10/.
Es scheint deshalb angeraten, den Begriff Reliabilität anders zu fassen, ihn nicht auf die messende Instanz, sondern auf das gemessene Resultat zu beziehen. So kann man bei einem hohen R-Maß zweier Transkribenten davon ausgehen, daß auch ihre Transkripte zuverlässig, also "reliabel" sind, und ihnen dieses Maß zuweisen (Text-Reliabilität [TR-Maß]). Umgekehrt darf allerdings bei einer geringen Text-Reliabilität nicht gefolgert werden, daß die Transkripte über den gesamten Text gleichmäßig weit voneinander abweichen. Tatsächlich haben die Re-

liabilitätsmessungen der '60er Jahre (/2/, /3/,/4/) gezeigt, daß z. B. die Reliabilität von hohen Vokalen wesentlich höher ist als die von tiefen. Die Meßgröße für die Reliabilität bei diesen Untersuchungen wurde geometrisch/heuristisch gewonnen und läßt sich als Abweichung von einem Mittelwert interpretieren (Lautklassenreliabilität [LR-Maß]). Gibt das LR-Maß quasi paradigmatisch die Reliabilität für jedes Symbol des Alphabets, so kann natürlich auch syntagmatisch jedem Laut des Textes ein solches Maß zugeordnet werden (Symbol-Realibilität [SR-Maß]).
Durch den Übergang vom TR- zum LR- und schließlich zum SR-Maß steigt der Rechen- und insbesondere der Darstellungsaufwand. Andererseits kommt man so der ja eigentlich angestrebten physikalischen Idealvorstellung immer näher, nämlich für jeden einzelnen Meßwert einen eigenen Reliabilitätswert zu bestimmen.
Messen R-Maße die Abweichungen der Notationen verschiedener Transkribenten für eine gegebene Aussprache, so messen Variabilitätsmaße die Abweichungen verschiedener Aussprachen bei ggfs verschiedenen Sprechern. Dafür lassen sich natürlich die gleichen Maße anwenden, die Werte sind nur anders zu interpretieren. Zur Konstruktion und Anwendung eines solchen Maßes auf Wortbasis (Wort-Variabilität [WV-Maß]) vgl./12/.

KONSTRUKTIONS- UND ANWENDUNGSPROBLEME

Diese kurze Übersicht dokumentiert, daß sich sehr leicht eine Vielzahl von Ä- und R-Maßen konstruieren läßt (für auch hier verwendbare Maße aus der psychologischen Testtheorie vgl. /11/). Für alle diese Masse bestehen jedoch gewisse gemeinsame Probleme, so daß man die Brauchbarkeit eines Maßes danach beurteilen kann, wie es diesen Problemen gegenübersteht. Es sollen hier einige der augenfälligsten Probleme genannt und zum Teil mit Anmerkungen versehen werden.
(1) Die meisten Ä-Maße (auf Intervallskalenbasis) sehen keine Vergleiche zwischen Vokalen und Konsonanten vor (vgl. /6/,/8/). Für die Ermittlung der Reliabilität muß jedoch manchmal der Abstand von einem Vokal zu einem Konsonanten bestimmt werden. (2) Bei der Konstruktion von solchen Maßen entsteht die Frage nach dem größtmöglichen Abstand im vokalischen und im konsonantischen Bereich. Müssen sie gleich groß sein oder nicht? Eine Reihe ähnlicher Überlegungen im Zusammenhang mit der Konstruktion von Ä-Maßen eskaliert dann in der Hauptfrage: Mit welcher Gewichtung gehen die einzelnen (Klassifikations-)Dimensionen in die Abstandsfunktion ein? (3) Der Übergang vom Ä-Maß zum R-Maß bringt ein neues Problem. Zwei Transkripte des gleiches Sprachsignals werden sicher sehr oft eine unterschiedliche Anzahl von Symbolen enthalten. Somit muß an einigen Textstellen der Abstand von "einem" Laut zu "keinem" Laut bestimmt werden. Die bekannten R-Maße (/6/,/8/) verwenden in diesem Fall den größtmöglichen Abstand. Ein Symbol wird aber doch wohl dann am ehesten im Text fehlen, wenn der jeweilige Laut im Signal undeutlich erscheint, wie etwa bei Reduktionen (z. B. [ra:tən]→[ra:t^{ə}n]→[ra:tn]). Der Unterschied zwischen den beiden letzten Notationen sollte deshalb nicht grundsätzlich genauso bewertet werden wie der zwischen zwei deutlich wahrnehmbaren Vokalen (etwa [y] und [a]). Es scheint also so, daß bei Reliabilitätsuntersuchungen zusätzlich zu einer Qualitätsmessung (mit dem Ä-Maß) auch eine Substanzmessung (betreffend Deutlichkeit, bei synthetischer Sprache: Natürlichkeit) vorgenommen werden und in das R-Maß einfließen sollte. (4) Dies führt aber auf eine grundsätzliche Überlegung: Eignen sich Ä-Maße übernaupt als Basis für R-Maße? Ganz deutlich stellt sich die Frage, wenn aufgrund von "semantischem Hören" (an undeutlichen Textstellen) ganz andere Wörter gehört werden (vgl. z.B. /13/). An solchen Stellen können Ä-Maße nicht viel aussagen, was bedeutet, daß sich Ä-Maße als Basis von R-Maßen nur dann eignen, wenn die jeweiligen Transkripte nicht zu sehr voneinander abweichen. (5) Eines der wesentlichen Probleme bei den bekannten R-Maßen betrifft die Nicht-Beachtung möglicher systematischer Abweichungen. So können gerade bei automatischer Anwendung der Meßvorschrift systematisch auftretende Abweichungen zu scheinbar geringen Reliabilitätswerten führen, obwohl sich vielleicht durch eine einfache Translation des Referenznetzes eines Transkribenten die Reliabilität entscheidend vergrößern würde. So könnte z.B. die Aspiration eines Plosivs, etwa [p], welche ein Transkribent (T2) als [p^{h}] (bzw. [p']) notiert, bei einem anderen (T3) durchweg als [p'] (bzw. [p]) erscheinen (vgl. /1/). Durch eine Transformation (Justierung) der Transkripte, also z.B. durch eine Anhebung des Aspirationsgrades bei T3 ließe sich die Reliabilität erhöhen. Wenn aber aus einer höheren Reliabilität auch eine höhere Validität folgt, stellt dieses Beispiel die deterministische Auffassung von Transkriptionsergebnissen infrage. Denn wenn nach eingehender Diskussion das [p'] des einen Transkribenten ein [p^{h}] beim zweiten bleibt, so ist das gemeinsame, justierte Transkript zwar reliabel, aber nicht mehr eindeutig. Diese Tatsache richtet den Blick direkt auf eine probabilistische Transkriptionsauffassung, die für einen Laut ggfs. mehrere Symbolalternativen, nach ihrer Wahrscheinlichkeit geordnet, zuläßt. Ein solches Transkript entspricht im übrigen auch dem "natürlichen Ergebnisverhalten" automatischer Prozesse, wo im Laufe der Berechnung immer mehrere Alternativlösungen vorhanden sind, und schließlich die wahrscheinlichste das Endergebnis bildet.

VERWENDBARKEIT FÜR DIE AUTOMATISCHE TRANSKRIPTION

Grundsätzlich lassen sich natürlich alle erwähnten Ä- und R-Maße auch zur Beurteilung automatischer Meßprozesse verwenden. Welche sind jedoch zu bevorzugen? Typischerweise variieren akustische Meßgrößen, die die Ausgangsbasis für jeden aT-Prozeß bilden, kontinuierlich. Diesem entspricht am ehesten ein R-Maß mit einem zugrundeliegenden Ä-Maß auf Intervallskalenniveau. Gleichfalls wird man ein Ä-Maß bevorzugen, das auf akustischer bzw. auditiver Ähnlichkeit beruht. Denn die akustischen Meßgrößen (Parameter) lassen sich oft nur schwerlich mithilfe der artikulatorischen Klassifikationsdimensionen interpretieren. So kann auf der Basis akustischer Meßgrößen kaum erklärt werden, warum etwa der Abstand zwischen [p] und [t] kleiner ist als der zwischen [p] und [k], was jedes Ä-Maß auf artikulatorischer Basis grundsätzlich vorsieht. Andererseits besteht bei den auditiven Ä-Maßen aufgrund ihrer Konstruktion mithilfe von Hörtests die Gefahr einer sprachspezifischen Färbung.
Wenn das TR-Maß eines automatisch erstellten Transkripts nicht zu sehr von dem einer menschlichen Vergleichsgruppe abweicht, so weist dies auf die Brauchbarkeit des aT-Verfahrens hin. Diese Aussage läßt sich beim Übergang von den globalen TR-Maßen zu den LR-Maßen weiter verfeinern. Es ist durchaus vorstellbar, daß ein aT-Algorithmus für gewisse Lautklassen bereits akzeptable Ergebnisse liefert, für andere dagegen noch nicht. Nach Anwendung der aT müßte dann der Text abschließend nur noch einmal für kritische Lautklassen manuell überprüft werden. Der Übergang zu einem SR-Maß erlaubt schließlich die differenzierteste Beurteilung. Die Prüfung des automatischen Transkripts könnte beispielsweise nur an den maximal reliablen Textstellen stattfinden oder, wenn der Algorithmus mehrere gewichtete Alternativen ausgibt, auch diese einbeziehen.

AUSBLICK

Sind auf der einen Seite R-Maße die Voraussetzung zur Güteprüfung von rechnergestützten Analyseprozeduren, so gestatten rechnergestützte Syntheseprozeduren in Zukunft vielleicht sogar eine Validitätsprüfung. Bereits heute ist es nämlich möglich, ausgehend von einer phonetischen Symbolfolge durch Simulation des menschlichen Spracherzeugungsmechanismus synthetische Sprache zu erzeugen, die jedoch noch nicht die Qualität natürlicher menschlicher Sprache erreicht. Sollte es aber in Zukunft gelingen, den Sprechvorgang so gut nachzubilden, daß sich das Resultat weder auditiv (Hörtests!) noch artikulatorisch (Sehtests!) von einem natürlichsprachlichen unterscheidet, so wäre damit die Grundlage für eine echte Validitätsprüfung geschaffen. Die richtige Symbolfolge als Basis für die Synthese ist dann unabhängig von der Analyse (Transkription) bekannt!

LITERATUR

/1/ R. Greisbach: Grundlagen der Automatisierbarkeit phonetischer Transkription. Diss. Köln 1986 (im Druck).
/2/ P. Ladefoged: The nature of vowel quality. Rev. Lab.Fonet.Exper., Coimbra, 5 (1960) 73-162.
/3/ J. Laver: Variability in vowel perception. Lang.&Speech 8 (1965) 95-121.
/4/ G. Heike: Auditive und akustische Beschreibung lautlicher Äußerungen mit Hilfe eines lautlichen Bezugssystems. Z.f. Mundartforschung, Beih., N.F. 3 u. 4 (1967) 356-362.
/5/ W.H. Vieregge et al.: A distinctive feature based system for the evaluation of segmental description in Dutch. Proc. 10th Int. Congr. Phon.Scie. Dordrecht 1984, 654-659.
/6/ W.H. Vieregge: Ein Maß zur Reliabilitätsbestimmung phonetisch-segmenteller Transkriptionen. Z.f. Dialektol.u. Ling. 52 (1985) 167-180.
/7/ W.H. Vieregge: The problem of validity of segmental transkriptions. Proc. Inst. Phonetics, Cath. Univ. Nijmegen 10 (1986) 23-26.
/8/ A. Almeida, A. Braun: "Richtig" und "Falsch" in der phonetischen Transkription. Z.F. Dialektol.u. Ling. 53(1986) 158-172.
/9/ B. Schädel: Über Schwankungen und Fehlergrenzen beim phonetischen Notieren. Bull. de Dial.Rom. 2 (1910) 1-29.
/10/ M. Bürkle: Zur Validität eines Maßes zur Reliabilitätsbestimmung phonetisch-segmenteller Transkription. Z.f. Dialektol. u. Ling. 53 (1986) 173-181.
/11/ J. Asendorf, H.G. Wallbot: Maße der Beobachterüberienstimmung. Z.f. Sozialpsychol. 10 (1979) 243-252.
/12/ S. Geršić: Mathematisch-statistische Untersuchungen zur phonetischen Variabilität am Beispiel von Mundartaufnahmen aus der Batschka. Göppingen 1971.
/13/ P. Winkler: Anwendungen phonetischer Methoden für die Analyse von Face-to-Face-Situationen. In: P. Winkler (Hrsg.): Methoden der Analyse von Face-to-Face-Situationen. Stuttgart 1981. 9-46.