Automatische Erkennung Nonverbaler Sprache

Holger Quast
Drittes Physikalisches Institut
Universität Göttingen
Bürgerstraße 42-44
D-37073 Göttingen
holcus@physik3.gwdg.de

und
Machine Perception Lab, Institute for Neural Computation
University of California, San Diego
http://mplab.ucsd.edu/~holcus

Ziel dieser Arbeit ist es, Muster in den akustischen Parametern von Sprache zu erkennen und daraus auf den nonverbalen, z.B. emotionalen, Bestandteil zu schließen. Dafür wurde eine Sprachdatenbank aufgenommen, deren para- und extralinguistischer Gehalt von amerikanischen und deutschen Hörern beurteilt wurde. Neuronale Netze lernen, anhand aus den Aufnahmen extrahierten Signalverarbeitungs-Größen das Beurteilungsverhalten der Hörer zu modellieren.

Einleitung

Gegenwärtige Sprachdialogsysteme beschränken sich darauf, den verbalen – linguistischen – Bestandteil eines Gesprächs zu erfassen, wodurch ein Großteil der gegebenen Information, wie zum Beispiel Emotionen, die Einstellung des Sprechers gegenüber dem Gesprächsstoff, körperliche Verfassung, etc. ignoriert wird. Wichtiger noch: auch die Bedeutung hängt in vielen Fällen von der Prosodie des Gesprochenen ab. Beschränkt sich zum Beispiel ein zukünftiges PC-Spracherkenner auf den verbalen Anteil eines verärgerten „Toll gemacht, Computer, Du hast gerade meine wichtigste Datei gelöscht!“, so wird der Lernerfolg im Rechner nicht im Sinne des Benutzers sein. Das häufige Auftauchen von Emoticons wie ;-) oder :(in Emails, also einem rein verbalen Medium, verdeutlicht, dass eine Nachricht nicht nur davon abhängt, was gesagt wird, sondern auch wie.

Gegenstand dieser Arbeit ist es, die Prosodie von Sprache geeignet zu repräsentieren und daran die nonverbalen Komponenten vokaler Kommunikation automatisch zu erkennen.

Psycholinguistische Bewertung

Als Grundlage für die Experimente wurde in Göttingen eine Datenbank mit 145 Sprachbeispielen von Schauspielern und nicht-Schauspielern aufgenommen. Alle sprachen den selben 8-Satz Monolog. Um möglichst natürliche Aufnahmen zu erhalten, wurden die Schauspieler gebeten, verschiedene *Situationen* zu interpretieren (nicht, wie in der Literatur üblich, möglichst starke einzelne *Emotionen* darzustellen). Standard-Szenen waren zum Beispiel Sprechen im Kreise der Familie, als Chef einer Firma, oder bei einer Rede vor dem Bundestag.

Zur Beurteilung der Sprachproben wurden in der Regel Testpersonen gebeten, in *forced choice* Experimenten jede Aufnahme genau einem nonverbalen Ausdruck aus einer begrenzten Anzahl zuzuordnen, z.B. Ekmans 6 Grundemotionen [Ekman 92]. Im Gegensatz dazu wurde der nonverbale Inhalt der Aufnahmen bei dieser Arbeit in jeder Kategorie bewertet. 20 Amerikanische Testpersonen, denen die verbale Information nicht zugänglich war, beurteilten ihren Eindruck beim Hören auf einer 5-Werte Skala von –2 bis +2 in einem *semantischen Differential* [Osgood/Snyder 69], das aus den Kategorien *angenehm, glücklich, erregt, physisch stark, selbstbewusst, ärgerlich* und *Führungsqualität* gebildet wird. Zum Vergleich nahmen deutsche Auswerter die gleiche Beurteilung über das Internet vor. Insgesamt wurden 150 Sätze bewertet.

Um unterschiedliche Bewertungsverhalten auszugleichen, wurden die Auswertungen jedes Hörers in jeder Kategorie auf einen Mittelwert von 0 und eine durchschnittliche absolute Abweichung von 0,5 normiert.

Aus den 20 Bewertungen für jede Kategorie und jede Aufnahme wird ein Histogramm erstellt. Die Position der maximalen Mode auf der Auswertungs-Skala wird als der Bestwert des Datenpunktes gewählt. Zusätzlich wird die Höhe der Mode als ein Wert für die Übereinstimmung der Bewertungen notiert. Abbildung 1 zeigt, wieviele der 150 Histogramm-Maxima in jeder psycholinguistischen Kategorie eine gegebene Höhe erreichen.

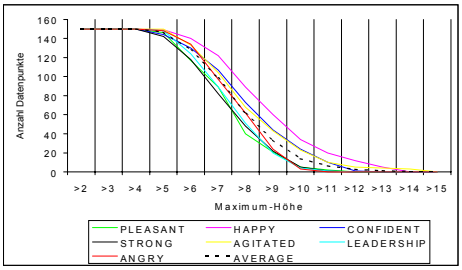

Abb. 1 Anzahl der Histogramm-Moden, die den auf der x-Achse notierten Wert überschreiten

Signalverarbeitung

Aus jedem Satz wurden 18 Parameter extrahiert, die die Aufnahme in den Kategorien Grundfrequenz (f_0), Lautheit, Durchschnittsspektrum und weitere prosodische Merkmale beschreiben. Diese sind Median der Grundfrequenz und der absoluten f_0-Veränderung pro Zeit; die Intensität in den Bändern 0-500Hz, 0,5-1kHz, 1-2kHz, 2-4kHz, 4-8kHz; das Lautheits-Maximum, der Median von Lautheit und Lautheitsveränderung; die Korrelation von Lautheit und f_0; Lautheit und f_0 am lautesten sowie am letzten stimmhaften Sample, das f_0-Interval zwischen dem lautesten und dem letzten Punkt; die Sprechgeschwindigkeit und das Geschlecht des Sprechers.

Hierfür wurde ein pitch tracker programmiert, der zwei verschiedene Methoden (Autokorrelation mit Centerclipping und Cepstrum, s. [Schroeder 99], [Hess 83]) kombiniert, und mit Berücksichtigung vorangegangener Werte einen robusten und präzisen Tonhöhen-Wert liefert. Beide Algorithmen liefern je 4 f_0-Kandidaten mit zugehörigen Wahrscheinlichkeitswerten (aus den peak-Höhen in Cepstrum und Autokorrelation). Diese Wahrscheinlichkeiten werden verringert, wenn sie außerhalb der normalen Tonhöhen-Änderung liegen.

Ein neues psycholinguistisch motiviertes Sprachlautheits-Modell [Quast 2000] erlaubt, einen Wert für die wahrgenommene absolute Lautheit eines Sprachbeispiels anzugeben, der unabhängig vom Verstärkungsgrad bei der Aufnahme oder beim Abspielen ist.

Basierend auf Werten aus [Paulus/Zwicker 1972] wird zunächst die Frequenzskala von Hz nach Bark transformiert. Für jeden Intensitätswert wird, abhängig von seiner Frequenz, eine *Kernlautheit* bestimmt. Um dem Verdeckungsverhalten unseres Gehörs Rechnung zu tragen, cf. [Zwicker 90], wird jede Kernlautheit entweder Ausgangspunkt einer Maskier-Fläche im spektralen Lautheits-Diagramm, oder aber sie wird selbst verdeckt, wie in Abbildung 2 dargestellt. Die Gesamtlautheit in einem Zeitfenster wird durch Integration aller Maskierflächen ermittelt.

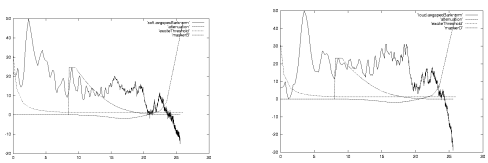

Abb. 2 Schematische Darstellung der psycho-akustischen Lautheitsberechnung für Durchschnitts-Spektren mit geringem (links) und hohem (rechts) vokalen Aufwand.

Geht man nun davon aus, dass die Intensitäts-Anhebung höherer Frequenzen bei lauten Stimmen stärker ist als bei leisen, kann man durch Normieren des Mittelwertes der lautesten Frequenzen aus einer überschaubaren Anzahl stimmhafter Zeitfenster eine Basis für einen absoluten Lautheitswert gewinnen.

Maxima in der Lautheits/Zeit-Kontur korrelieren zu einem hohen Grad mit einzelnen Silben und bieten damit eine gute Möglichkeit zur Diskretisierung der Sprachaufnahme, ohne – wegen zeitlicher Verdeckungseffekte – viel Information zu verlieren. Gewinnt man die oben angeführten akustischen Parameter nur an Lautheits-Maxima, so kann die Datenmenge, die den nonverbalen Anteil repräsentiert, durchschnittlich um den Faktor 4000 reduziert werden (ausgehend von der ursprünglichen DAT Aufnahme).

Mustererkennung

Ein neuronales Netz (multilayer perceptron mit backpropagation, s.[Hecht-Nielsen 91]) wird trainiert, anhand der Signalverarbeitungs-Parameter die einzelnen psycholinguistischen Empfindungen zu lernen, so dass die Software nach dem Training zum Beispiel zu beurteilen vermag, wie erregt oder selbstbewußt eine neue, der Software noch unbekannte Sprachaufnahme klingt.

Beim Training wird ein Beispiel-Punkt um so stärker berücksichtigt, je größer die Übereinstimmung der Bewerter – gegeben durch die Histogramm-Moden-Höhe h – für eine Aufnahme war. Diese Abhängigkeit wird über die *Lernrate a* realisiert, die bestimmt, wie stark ein Gewicht der linearen Kombinierer in einem einzelnen Neuron bei einem Lernvorgang verändert werden soll:

$$\alpha(h) = \begin{cases} 0.01 & \text{for } h > 10 \\ 10^{\frac{h}{2}-7} & \text{sonst} \end{cases}$$

Resultate

Die neuronalen Netze vermögen die nonverbalen Empfindungen in den Kategorien selbstbewusst, stark, erregt, und Führungsqualität zu lernen. In den drei verbleibenden Kategorien ist dies hier nicht der Fall. Dies wird anhand der Fehler-Entwicklung, wie in Abb. 3 dargestellt, ersichtlich:

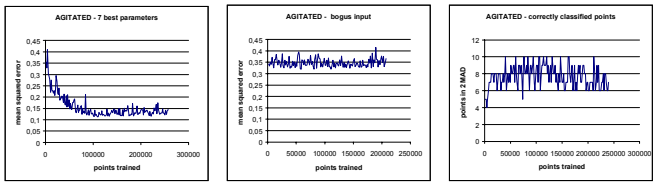

Abb. 3 Fehler im Verlauf des Lernvorgangs. Links der quadratische Fehler bei erfolgtem Lernen, mitte: ohne Lernerfolg. Rechts: Anzahl der 15 Testpunkte, die korrekt plaziert wurden.

Die Netzwerke in den vier gelernten Kategorien zeigen alle ein Lernverhalten wie im linken Graph von Abbildung 3 dokumentiert, die nichtgelernten eine Fehlerentwicklung wie im mittleren Diagramm.

Ein Punkt wird als richtige Einschätzung gewertet, wenn das Netzwerk ihn innerhalb eines Intervalls um die Moden-Position plaziert, das durch den 2-fachen Median aller Abweichungen um die Mode (2-MAD) definiert ist.

Abbildung 4 stellt dar, wie sich die Klassifizierungsleistung für eine verschiedene Anzahl von Eingabe-Parametern entwickelt (gemittelt über 20 Versuche). Dabei wird der Parameter gesucht, mit dem die höchste Genauigkeit erreicht werden kann. Dieser wird festgehalten und der zweiter Wert gesucht, mit dem die beste Leistung möglich ist, danach ein dritter etc. bis alle 18 festgelegt sind. Da die psycholinguistischen Daten nicht vollständig „weiß" sind, ist die Erkennungsleistung selbst bei „sinnlosem" akustischen pseudo-Input besser als zufällig, da die Verteilung der Hörer-Bewertungen gelernt wird.

Um diesen Effekt zu eliminieren, wird diese Grundleistung auf 1 normiert.

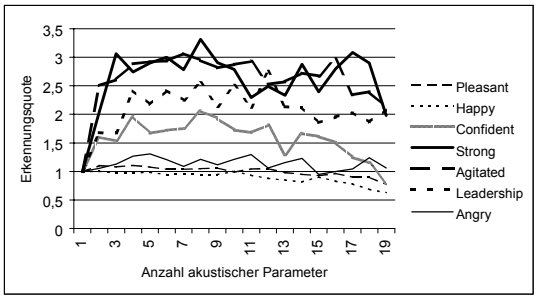

Abb 4 Die normierte Erkennungsleistung der Netzwerke in den einzelnen Kategorien.

Diskussion, Ausblick

Die Erkennungsleistung der neuronalen Netze liegen in allen 7 Kategorien im gleichen Bereich wie die der menschlichen Bewerter unter dem gleichen 2-MAD Auswertungs-Kriterium. Die 3 nicht erlernten Kategorien zeichnen sich dadurch aus, dass die Varianz der Bewertungen und somit vermutlich ebenfalls die nonverbale Information in ihnen am geringsten ist, was zur Hoffnung berechtigt, dass auch in diesen Klassen eine Mustererkennung grundsätzlich möglich ist. (Die Fähigkeit des Menschen, Emotionen in Sprache zu codieren und decodieren, wurde bereits in einer Reihe von Studien gezeigt, s. z.B. [Scherer et al 91])

Programme, die in der Lage sind, nonverbale Sprache zu erkennen und zu quantifizieren, können zu natürlichen Mensch-Maschine Dialogsystemen führen, da sie sowohl eine umfassendere Erkennung ermöglichen als auch eine der Situation angepasste Sprachsynthese. Tutorprogramme sind denkbar, mit denen ein Sprecher gezielt seinen Stimm-Ausdruck trainieren kann, so mag zum Beispiel ein Manager oder Politiker daran interessiert sein, auf einen überzeugenden, selbstbewussten Stil hinzuarbeiten, ein Lehrer auf einen interessanten, etc. Gehörlose Menschen, die mangels Feedback in der Regel mit einer unnatürlichen Prosodie sprechen, könnten eine ähnliche Software benutzen, die sie in der logopädischen Therapie unermüdlich und kostengünstig unterstützt. Ebenso könnten Autisten oder Menschen mit Asperger-Syndrom, die Probleme beim Erkennen und Äußern von Emotionen haben, dieses in stressfreiher Umgebung erlernen. Stress-monitoring Programme könnten bei Dialogen im Straßen- oder Flugverkehr warnen, wenn sich ein Fahrer oder Pilot in einer psychisch kritischen Stress-Situation befindet.

Mit den vorgestellten Mitteln, insbesondere durch die Datenreduktion und die robuste Echtzeit-Signalverarbeitung, erscheinen diese Anwendungen realisierbar.

Literatur

Ekman, P.: An argument for basic emotions. Cognition and Emotion **6**, 169–200 (1992)

Hecht-Nielsen, R.: *Neurocomputing* (Addison-Wesley 1991)

Hess, W.J.: *Pitch Determination of Speech Signals* (Springer, Heidelberg New York 1983)

Osgood, C.E., Snider, J.G: *Semantic Differential Technique: A Sourcebook* (Aldine Publishing Co., Chicago 1969)

Paulus, E., Zwicker, E.: Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln. Akustika **27**, 253–266 (19722)

Quast, H: Absolute Perceived Loudness of Speech. In *Proceedings of the 7th Joint Symposium on Neural Computation, USC* (INC; San Diego 2000)

Scherer, K.R. et al: Vocal Cues in Emotion Coding and Decoding. Motivation and Emotion **15**, 123–148 (1991)

Schroeder, M.R.: *Computer Speech* (Springer, Heidelberg New York 1999)

Zwicker, E., Fastl, H.: Psychoacoustics (Springer, Heidelberg New York 1990)