Zeitschrift: Horizonte : Schweizer Forschungsmagazin
Herausgeber: Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung
Band: - (2003)
Heft: 58

Artikel: Dossier Künstliche Sinne : Maschinen, die zuhören
Autor: Dessibourg, Olivier
DOI: https://doi.org/10.5169/seals-552268

Maschinen, die zuhören

Spracherkennungssysteme versagen oft bei Hintergrundgeräuschen oder falsch betonten Wörtern. Nun will die Forschung vom natürlichen Ohr lernen.

VON OLIVIER DESSIBOURG

Wenn sie richtig hinhören, können Maschinen durchaus ein feines Gehör haben: Sie vermögen ihren Gesprächspartner zu erkennen und den Dialog sogar schriftlich festzuhalten – vorausgesetzt, das Gegenüber spricht deutlich, ist nicht gestresst oder emotional ergriffen, und keine Hintergrundgeräusche stören. In diesen Fällen kommen Spracherkennungssysteme an ihre Grenzen, haben Mühe, Wörter zu verstehen und Stimmen zu erkennen. Dem Ziel, diese Mängel zu beheben, haben sich zahlreiche Forschungsgruppen verschrieben, darunter auch eine Gruppe des Nationalen Forschungsschwerpunkts «Interaktives multimodales Informationsmanagement» am Dalle-Molle-Institut für perzeptive und künstliche Intelligenz (IDIAP) in Martigny.

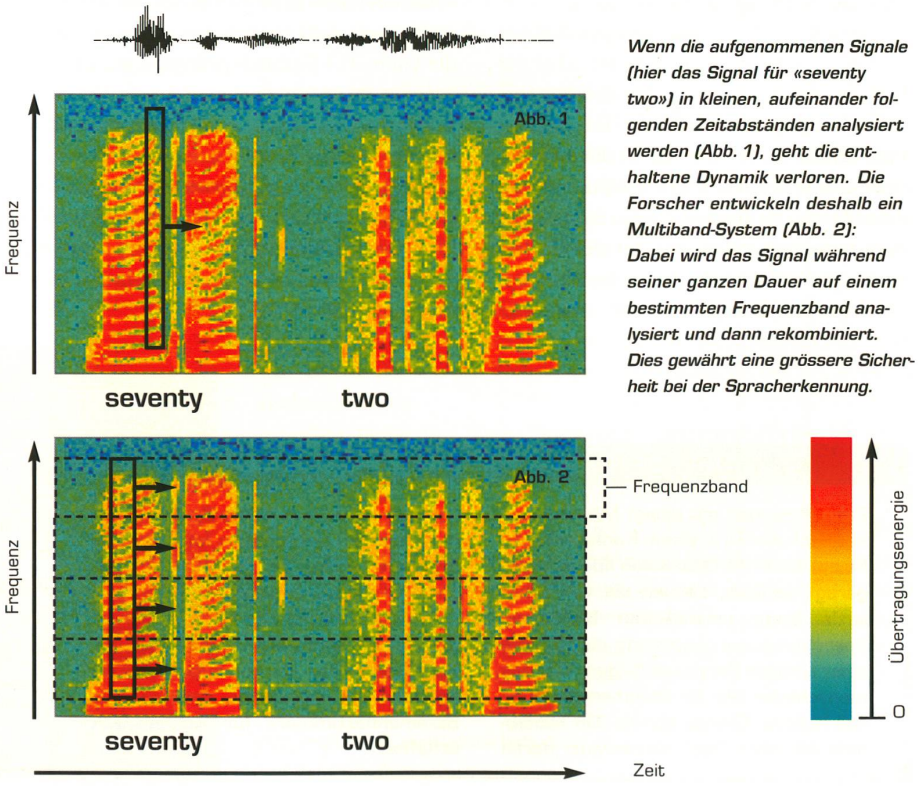
Frequenzspektren analysiert

«Diese Systeme analysieren zur Spracherkennung die Frequenzspektren», erklärt der Direktor des Instituts Hervé Bourlard. Jedes Tonsignal kann durch die Frequenzen (welche die Tonhöhen bestimmen) und die Übertragungsenergie (welche ungefähr der Lautstärke entspricht) charakterisiert werden. Spracherkennungssysteme analysieren die aufgenommenen Spektren schrittweise in kleinen Häppchen zu 25 bis 30 Millisekunden (Abb. 1). Der Inhalt dieser kurzen Ausschnitte wird dann über statistische Methoden mit einer Datenbank von Phonemen – den Grundelementen der Lautsprache – verglichen. Nach der Berücksichtigung lexikalischer und grammatikalischer Regeln werden die Daten an einen «Decoder» gesendet, welcher die wahrscheinlichsten Wortsequenzen auswählt. «Es handelt sich um eine statistische Methode, die nicht auf «Intelligenz», sondern auf einem simplen Vergleich mit einer Sammlung von Beispielen beruht», erläutert Bourlard. «Deshalb ist die Erkennung unzureichend, wenn ein Wort verkürzt ausgesprochen wird oder bei einer lebhaften Diskussion im Lärm unterzugehen droht.»

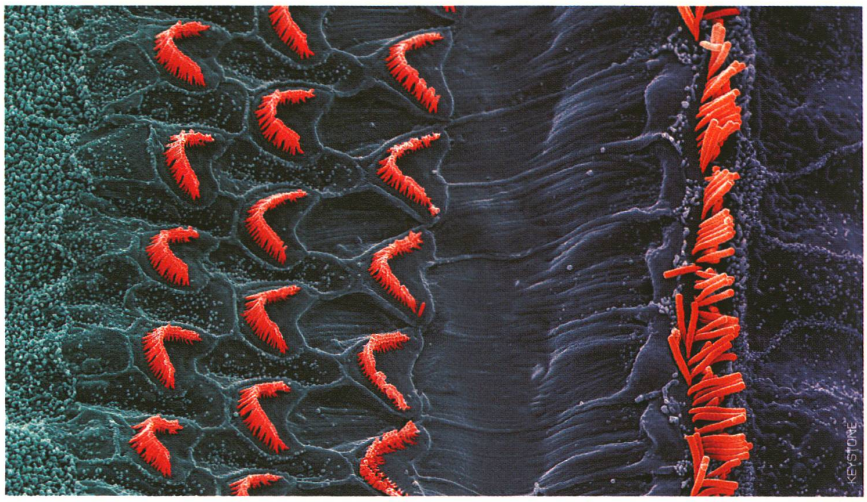
Eine Möglichkeit zur Verbesserung dieser Systeme besteht darin, dem Vorbild des menschlichen Ohrs nachzueifern. «Die Gehörgangsschnecke analysiert solche Spektren, bevor sie die akustischen Daten, die über Millionen vibrierender Flimmerhärchen eintreffen, an den Hörnerv weitergibt», führt er aus. Da das Ohr jedoch nicht für alle Tonfrequenzen gleich empfänglich ist, haben die Forschenden die verschiedenen Signale der Spektren in entsprechender Weise gewichtet.

Alles imitieren?

Es wäre aber unklug, alles unbesehen zu imitieren, ist Hervé Bourlard überzeugt: «Flugzeuge schlagen ja auch nicht mit den Flügeln und fliegen trotzdem. Der Mensch hat in diesem Fall von seinen Beobachtungen der Natur nur die entscheidenden Grundsätze der Aeronautik abgeleitet.» Ent-



Wenn die aufgenommenen Signale (hier das Signal für «seventy two») in kleinen, aufeinander folgenden Zeitabständen analysiert werden (Abb. 1), geht die enthaltene Dynamik verloren. Die Forscher entwickeln deshalb ein Multiband-System (Abb. 2): Dabei wird das Signal während seiner ganzen Dauer auf einem bestimmten Frequenzband analysiert und dann rekombiniert. Dies gewährt eine grössere Sicherheit bei der Spracherkennung.



Die Flimmerhärchen (hier rot eingefärbt) in der Gehörgangsschnecke wandeln Tonschwingungen in Nervenimpulse um.

sprechend müssen auch beim menschlichen Ohr die für solche Systeme entscheidenden Eigenschaften gefunden werden. «Sonst befinden wir uns quasi immer noch im Stadium flügelschlagender Flugobjekte...» Doch die Arbeit von Psychoakustikern kann durchaus beflügeln (siehe rechts).

Bourlards Forschungsgruppe befasst sich mit einem weiteren Problem: «Bei der Unterteilung des Signals in unabhängige kurze Zeitabschnitte geht die Information ihrer Dynamik verloren, die für das menschliche Ohr sehr wichtig ist», erklärt er. Denn sie ist für das Heraushören einer Stimme aus dem Umgebungslärm entscheidend. Nun wurde aber festgestellt, dass jedes Flimmerhärchen in seiner eigenen Frequenz schwingt und dass sich diese Härchen aufeinander abzustimmen scheinen, um die gewünschte akustische Information zeitlich nachzubilden. Diese Entdeckung liess die Forschenden aufhorchen. Und es gelang ihnen, diese flimmernde Parade zu simulieren: «Statt das Signal in zeitlichen Abschnitten zu analysieren, untersuchen wir Frequenzbänder (Abb. 2).» Dadurch gelingt es, im Lärm jene Kanäle zu bestimmen, die zuverlässige Informationen liefern, und entsprechend nur die nützlichsten Frequenzdaten zu rekombinieren. Diese sogenannten «Multiband-Systeme» sind der Stolz des IDIAP und haben bereits Anwendungen wie den «intelligenten Konferenzsaal» ermöglicht (siehe Kasten unten). ■

INTELLIGENTER KONFERENZSAAL

Zehn Personen mit einem Mikrofon vor sich diskutieren in einem Konferenzsaal. Selbst wenn die Diskussion äusserst angeregt verläuft, können die von Bourlards Team entwickelten Multiband-Spracherkennungssysteme die Stimme einer einzigen Person verfolgen. Wie das menschliche Ohr im Stimmengewirr eines Apéros. Ebenso könnte den Teilnehmenden einer Telefonkonferenz, deren Portrait auf einer mit dem Spracherkennungssystem gekoppelten Internetseite erscheint, immer die Identität der sprechenden Person angezeigt werden.

Die Ambitionen der IDIAP-Forschenden gehen aber noch weiter: Mit zusätzlicher Hilfe durch Bilderkennung möchten sie eigentliche «multimodale Stimmenbestimmer» für Konferenzen schaffen.

Emotionale Stimme

Spracherkennungssysteme werden bereits zur biometrischen Identifikation von Personen eingesetzt, beispielsweise in Banken. Sie arbeiten mit einer Zuverlässigkeit von etwa 95 Prozent. Doch verändert sich die Stimme des Sprechers leicht, weil er erregt ist, geraten die Systeme nicht selten in arge Schwierigkeiten. Diesem Mangel möchte ein Team an der Abteilung für Psychologie der Universität Genf abhelfen.

«Die Ingenieure setzen zur Verbesserung der Systeme auf Algorithmen und lassen dabei die Ursache der Veränderungen ausser Acht», bemerkt Klaus Scherer. «Mit dem Projekt «Emovox» versuchen wir, zuerst wirklich zu verstehen, was passiert, damit wir jene akustischen Parameter stärker gewichten können, die sich durch Stress oder Emotionen am wenigsten verändern.» Allerdings ist es schwierig, Personen zu einem bestimmten psychischen Zustand anzuhalten, um die akustischen Veränderungen ihrer Stimme zu messen. Deshalb setzte das Forschungsteam hundert Männer Stresssituationen aus, zeichnete ihre Stimme auf und analysierte 99 akustische Parameter wie die Grundfrequenz, die Standardabweichungen der Grundfrequenz, die Amplitude der Angriffsspitze des Signals und vieles mehr. «Zur Identitätsbestimmung sollten die bei einer Person stabilen Parameter herangezogen werden», erklärt Scherer. Damit das Verfahren aber wirklich zuverlässig ist, muss es sich dabei um Parameter handeln, die bei vielen verschiedenen Personen konstant bleiben.

Erste Ergebnisse, die kürzlich am Eurospeech-Kongress in Genf vorgestellt wurden, stützen die Erwartungen der Forschenden, die bereits solche Parameter bestimmen konnten. Die Daten müssen sich nun in weiteren Untersuchungen bestätigen lassen. «Wie die Stimme von einer Situation abhängt, ist ein Aspekt, der bei Spracherkennungssystemen oft verkannt wird. Unsere Ergebnisse könnten sich für die Ingenieure als nützlich erweisen», folgert Klaus Scherer. O. D.