

Spracherkennung

Einzelworterkennung

Anwendung 1: Word Spotting: Reagieren auf ein bestimmtes Wort

Anwendung 2: Erkennung segmentierter Sprache

Analyse des Sprachsignals

Erzeugung von Merkmalsvektoren

Phonemerkenkung

Fehlermöglichkeiten: Wortgrenzen

Erkennungsrate in Abhängigkeit von der Wortlänge

Buchstabieren

Sprecherabhängigkeit und Sprechererkennung

Skala der Sprecherunabhängigkeit

Sprecher-Modellierung

Anpassung an den Sprecher

Lambs and Goats (Lämmer und Böcke)

Erkennung von fließend gesprochenem Text

Stille in kontinuierlich gesprochener Sprache

Mensch-Maschine-Dialog

Word Spotting: Reagieren auf ein bestimmtes Wort ohne Berücksichtigung des Kontextes

Ziel

- Sichere Erkennen von (wenigen) Schlüsselwörtern;
- Erkennung unabhängig vom Kontext, bestehend aus Rauschen oder anderem Text.

Einsatzgebiete

- Computertelephonie (im Einsatz seit ca. 1994)
 - Übermittlung von Ziffern, *ja/nein* usw.
 - beschränkte Tonqualität
 - Störgeräusche durch Telefonleitung
- Befehlserkennung: Unterscheidung von Schlüsselwörtern von sonstigen Geräuschen;
- Erkennen relevanter Textstellen auf Grund von Stichwörtern.

Word Spotting II

Erkennungsmuster

- Stille - erkanntes Schlüsselwort - Stille

Probleme bei Computertelephonie

- Computertelephonie-Software wird vom gelegentlichen Nutzer nur schwer akzeptiert wegen
 - unnatürliche Dialogführung, der durchlaufene Entscheidungsbaum ist nicht transparent;
 - schlechte Erkennungsrate;
 - teilweise lange Antwortzeiten.
- Dialog-Dominanz durch ein scheinbar inkompetentes System.
- Unvorhergesehene Nutzerreaktionen aller Art müssen vorgesehen werden.

Erkennung segmentierter Sprache

Ziel

- relativ sicheres Erkennen von Einzelwörtern;
- Umfang von gegenwärtig ca. 50.000 Wortformen;
- Erschließung von unsicher erkannten Wörtern aus dem Kontext.

Einsatzgebiete

- Diktieren von Fachtexten
- Sprachsteuerung mit komplizierten Mehrwort-Befehlen

Anforderungen an die Eingabe

- Erfolgsrate besser bei einem Sprecher als bei Sprecherunabhängigkeit;
- Trainierte Sprecher sind besser als untrainierte;
- Umgebung ohne Nebengeräusche wünschenswert;
- Konstante technische Umgebung wünschenswert, d.h. gleiches Mikrofon, gleicher geringer Abstand Mund - Mikro.

Gegenwärtige Hardware-Anforderungen

- PC (mind. Pentium 133MHz / 32 MB), mehr ist besser, da Graphenmsuche in Echtzeit erfolgt und mit einem leistungsfähigen Rechner mehr durchsucht werden kann.
- nur Soundkarte für I/O, kein DSP-Chip

Analyse des Sprachsignals I

Frequenzspektrum

Ein konstanter Ton ist charakterisiert durch die auftretenden *Frequenzen* und ihre *Amplituden*, zusammengefaßt zum *Frequenzspektrum*.

Durch die zeitliche Veränderung wird das Frequenzspektrum zeitabhängig.

Graphische Darstellung: Spektrogramm (Potter, 1945)

- Horizontale Achse: Zeit
- Vertikale Achse: Frequenz (85 Hz ... 8000 Hz), Amplitudenhöhe als Graustufen

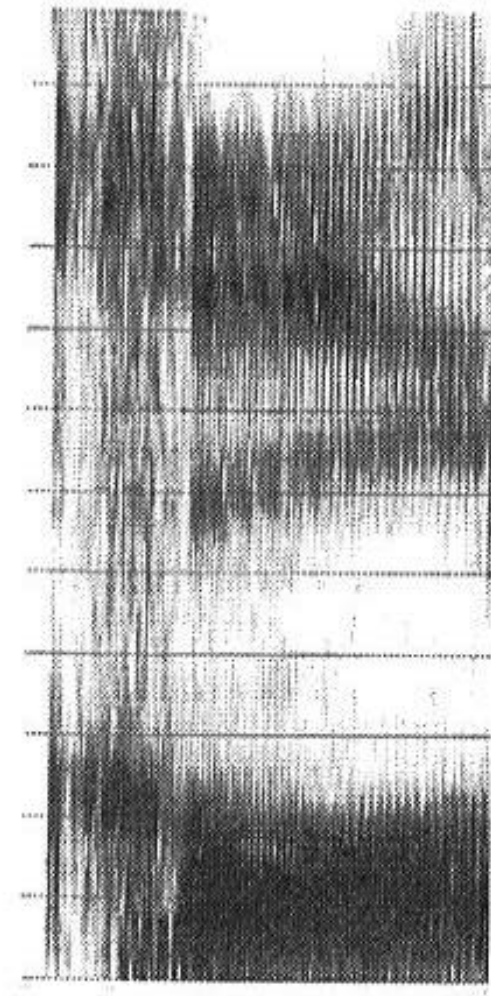


Figure 2.6 Spectrogram for the word "cool"

Analyse des Sprachsignals II

Unterschiede bei Wiederholungen

Auch bei demselben Sprecher weist bei Wiederholungen das Spektrogramm starke Unterschiede bei

- Zeitverlauf (Sprechgeschwindigkeit, Betonung)
- Tonhöhe (Satzmelodie, Betonung)
- Spektrum (Veränderung der Stimme im Tagesverlauf, wegen Heiserkeit oder über einen längeren Zeitraum)

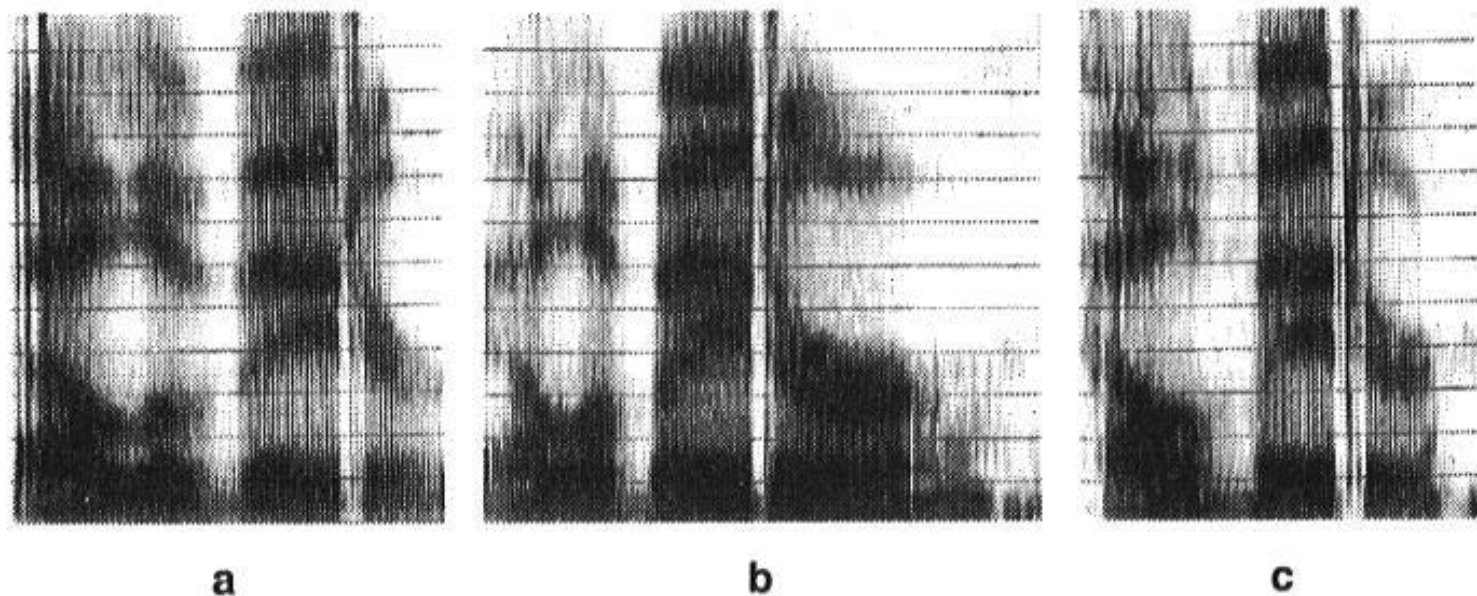


Figure 3.1 Spectrograms of three utterances of the word “elevator” by the author

Erzeugung von Merkmalsvektoren

Datenreduzierung

Die Auswertung von der aufgenommenen Tonfolge zum erkannten Wort macht eine drastische Datenreduzierung nötig.

- 1 sec. Ton gesampelt mit 22 kHz mono mit 8 bit liefert 22.000 Byte Daten.
- Für ein gesprochenes längeres Wort ergibt sich ein Reduktionsfaktor von ca. 1000.
- Als Zwischenschritt wird ein Merkmalsvektor erzeugt.

Reduktionsschritte

- Der Frequenzbereich wird in 8-20 Intervalle geteilt. Innerhalb der Intervalle erfolgt keine Unterscheidung mehr.
- Der Energiebereich wird (möglicherweise in Abhängigkeit vom Frequenzbereich) in wenige Intervalle geteilt.
- Insgesamt wird meist mit 256 Möglichkeiten gearbeitet.
- Die Abtastung erfolgt alle 20-30 ms.

Der Merkmalsvektor

- Zwischen zwei Ruhephasen wird so ein Merkmalsvektor mit Einträgen im Bereich 0-255 entsprechend der Wortlänge gebildet.
- In unserem Beispiel von 1 Sekunde hat der Vektor eine Länge zwischen 30 und 50.

Phonemerkennung

Ziel

Dem Merkmalsvektor soll eine Folge von Phonemen zugeordnet werden, die insgesamt ein Wort aus dem Lexikon ergeben.

Problem

- Diese Zuordnung ist häufig nicht eindeutig möglich.
- Deshalb gewichtete Liste der wahrscheinlichsten Kandidaten.
- Auswahl in Echtzeit verlangt schnelle Verfahren, z.B.
 - Übergangswahrscheinlichkeiten (oder n-Gramme) von Phonemen oder
 - Grammatik, die Wortbildung aus Phonemen beschreibt.
 - Zusammenfassung als **Triphone-Modell**: Zu einem Phonem werden Kontext-Informationen über Paare von linken und rechten Nachbarn gespeichert.

Fehlermöglichkeiten

Fehler bei der Reduktion

- Abtastrate von 20-30 ms kann kurze "t"-Laute verschlucken.

Bei segmentiertem Text: Fehlerhafte Wortgrenzen

- Keine Erkennung einer Wortgrenze wegen zu kurzer oder fehlender Sprechpause.
- Zusätzliche Wortgrenze z.B. wegen Pause in Kompositum (z.B. *Bilder-Rahmen*).

Bei Word Spotting: Verwechseln von Störgeräuschen mit Wörtern

- Auslassungen: Verstanden wird 1-2-4 statt 1-2-3-4;
- Substitution: Verstanden wird 3 statt 2;
- Einfügung: Verstanden wird 1-2-3-8-4 statt 1-2-3-4.

Buchstabieren

Mensch-Mensch-Dialog

Buchstabieren ist übliches Mittel zur Erklärung bei Mehrdeutigkeiten (z.B. bei Namen) oder "schwierigen" Wörtern.

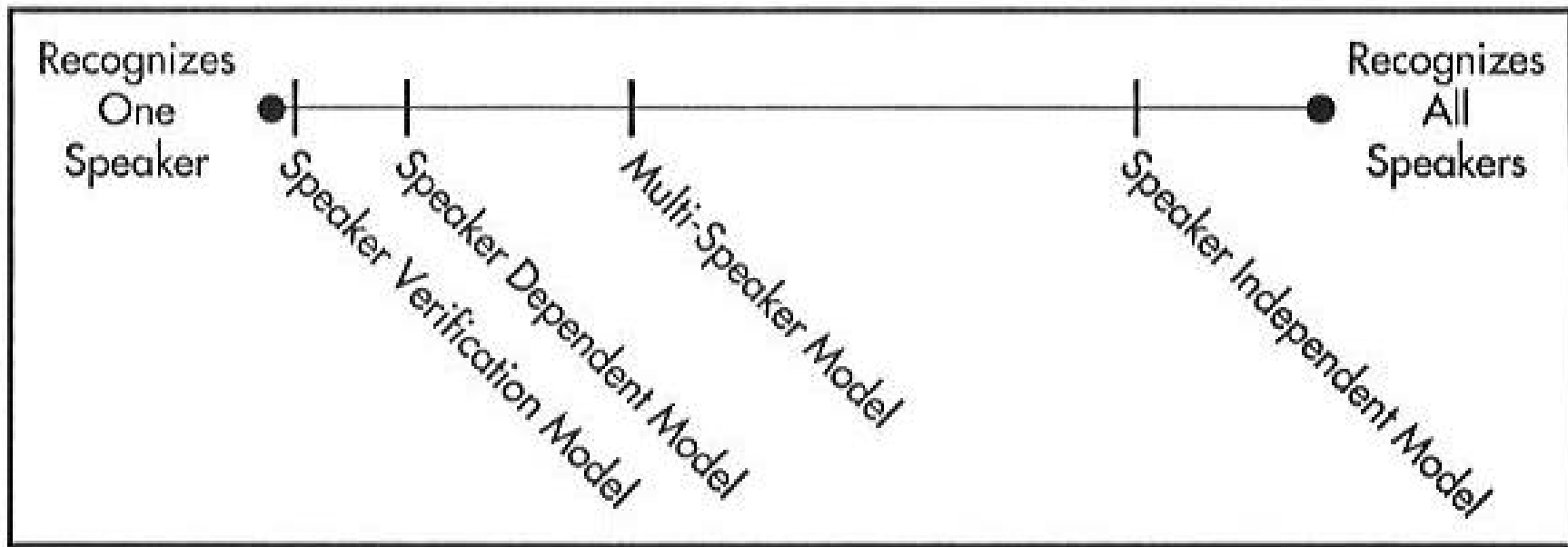
Mensch-Maschine-Dialog

Im typischen Mensch-Maschine-Dialog steht Keyboard zum Buchstabieren zur Verfügung. Trotzdem soll auch akustisch buchstabiert werden können.

Besonderheiten beim Buchstabieren

- Nur wenige Zeichen (d.h. Wörter für die einzelnen Zeichen), also sollte gute Erkennung möglich sein.
- Trotzdem sind einige Zeichen leicht zu verwechseln (b und p, im Deutschen 2 und 3, im Englischen 5 und 9).
- Deshalb besondere Ausspracheformen ("zwo" für zwei, "nigh-in" bzw. "niner" für 9).
- Spezielle Buchstabieralphabete (Anton, Berta, Cäsar, Dora, ...) (nicht einheitlich und nicht immer bekannt).
- Spezielle Markierungen für Beginn und Ende des Buchstabierens nötig.

Skala der Sprecherunabhängigkeit



Bessere Erkennungsrate

Größeres Vokabular

Schlechtere
Erkennungsrate

Kleineres Vokabular

Ein sprecherabhängiges Sprachmodell kann auch von einem anderen Sprecher benutzt werden; dann sinkt aber die Erkennungsrate. Dieser Effekt kann gemindert werden, falls die beiden Sprecher stark unterschiedliches Vokabular benutzen.

Erkennungsrate in Abhängigkeit von der Wortlänge

- Merkmalsvektoren für Wörter sind in ihrer Länge proportional zur Sprechdauer.
- Lange Wörter haben also lange Vektoren, die sich (auch mit leichten Abweichungen) noch sicher identifizieren lassen.
- Schwierigkeiten gibt es mit kurzen Wörtern (Dauer ca. 1/5 sec., d.h. Merkmalsvektor der Länge 5-10), das sind aber gerade häufige Artikel und Präpositionen (z.B. ein, an, am, um, hin, in, im, ...)

Sprecher-Modellierung I

Sprecherabhängiges Modell

- Grund-Training: Fester Text wird vorgelesen, daraus wird ein Modell errechnet, wie Triphone gesprochen werden.
- Andere Wörter im Lexikon werden erkannt, falls deren Aussprache dem Modell entspricht.
- Entspricht die Aussprache nicht dem Modell, müssen einzelne Wörter weiter geübt werden.
- Derzeit ist es möglich, daß ein Wort praktisch niemals akzeptiert wird, falls die Aussprache nicht den phonetischen Erwartungen des Systems entspricht.

Multi-Speaker-Modell

- Training erfolgt durch mehrere Personen.
- Das Sprachmodell wird über die Benutzer gemittelt.
- Deshalb etwas schlechtere Erkennungsrate.
- Lexikon-Erweiterung schwierig, da von mehreren Personen gesprochen werden muß.
- Zusammenfassung: Größerer administrativer Aufwand bei schlechterer Erkennung.
- Einsatzgebiete: Fester Personenkreis mit (sehr) beschränktem Vokabular.

Sprecher-Modellierung II

Sprecherunabhängiges Modell

- - Sprachmodell ist mitgeliefert.
- - Kein Training nötig, deshalb minimaler administrativer Aufwand.
- - Dafür schlechtere Erkennungsrate.
- - Anpassung des Nutzers an das System in gewissen Grenzen möglich (d.h. Aussprache so, daß es verstanden wird).
- - Lexikon-Erweiterung in gewissem Umfang möglich ohne Sprechertraining, wenn phonologische Zerlegung korrekt möglich ist. Beschränkung in der Gesamtzahl durch Unterscheidbarkeit der Merkmalsvektoren.

Unterschied zur menschlichen Spracherkennung

- Menschliche Spracherkennung ist nicht wirklich sprecherunabhängig (Experiment: N Personen lesen einen Text so, daß abwechselnd jeder ein Wort liest. Das ist äußerst schwer verständlich.)
- Statt dessen erfolgt eine Anpassung an den Sprecher.

Automatische Anpassung an den Sprecher

- Zunächst enthält das System ein benutzerunabhängiges Modell.
- Wird ein Wort korrekt erkannt, wicht aber (wiederholt auf die gleiche Art) vom Modell ab, so wird das Modell angepaßt.
- Ergebnis: Typische Aussprache einzelner Phoneme oder Triphone wird sukzessive ins Modell integriert.
- Administrative Voraussetzung: Wichtig ist, daß jeweils derselbe Sprecher spricht, damit nicht "umgelernt" wird.
- Sprechererkennung ist nützlich.

Lambs and Goats (Lämmer und Böcke)

Die Zusammenarbeit Sprecher - Diktiersystem klappt unterschiedlich gut.

Lambs

Für die meisten Personen funktioniert der Trainingsprozeß gut, danach läßt sich ein Diktiersystem gut nutzen.

Goats

Es gibt Menschen, für die Diktiersysteme ungeeignet sind. Folgende Gründe können vorliegen:

- Akustische Muster in der Aussprache, die eine Erkennung deutlich erschweren.
- Mangelnde Kooperation mit dem System (Mikrophon-Probleme, mangelnde Segmentierung, ...)

Stille in kontinuierlich gesprochener Sprache

Rolle der Pausen

- Pausen werden als Trenner genutzt, um die Stücke dazwischen zu identifizieren.
- Die zu identifizierenden Stücke sind i.a. keine Wörter.
- Vorteil: Diese Stücke sind i.a. länger als die schwer zu erkennenden kurzen Wörter.

Position der Pausen

- Vor Explosiv-Lauten (t, p, k)
- Als Trenner in einem Kompositum (*Tee-Ei*)
- Bei vielen Satzzeichen
- ...

Prosodie: Satzmelodie

Dialog 1

A: Ist der Termin bei Ihnen auch gegen Ende der Woche möglich, zum Beispiel am Freitag?

B: Ja (,) zur Not paßt auch der Freitag.

Dialog 2

A: Ist der Termin bei Ihnen auch gegen Ende der Woche möglich, zum Beispiel am Donnerstag?

B: Ja (,) zur Not paßt auch der Freitag.

Dialog 3

A: Ist der Termin bei Ihnen auch gegen Ende der Woche möglich, zum Beispiel am Sonnabend?

B: Ja, zur Not. Paßt auch der Freitag?