Gedächtnisprotokoll zur Diplomprüfung im Vertiefungsfach SPRACHERKENNUNG

Fächer:	Spracherkennung (Ney, WS 01/02, Skript vom WS 99/00) Mustererkennung und neuronale Netze (Ney/Schlüter, SS 02) Digitale Signalverarbeitung für Sprache und Bilder (Ney/Schlüter, SS 2003, Kapitel 1-3)
Prüfer:	Ney
Datum:	19.9.2003
Note:	2.0

Spracherkennung

- Allgemein:
  - Wofür braucht man das?
  - Wie macht man das?
    ⇒ Diagramm aus dem Skript aufgemalt (Sprachsignal wird in Sequenz von Merkmalsvektoren überführt, Bayes'sche Entscheidungsregel, ...).

Übergang zur Mustererkennung

- Bayes'sche Entscheidungsregel:
  - Hinschreiben.
  - Wann ist sie optimal?
    ⇒ Bei minimaler Fehlerrate als Kriterium.
- Gauß:
  - Wenn als klassenbedingte Wahrscheinlichkeit die Gaußverteilung angenommen wird, wie sieht das dann aus?
    ⇒ Gaußverteilung hinschreiben (univariat, multivariat für stochastisch unabhängige und stochastisch abhängige Komponenten).
  - Warum nimmt man bei Spracherkennung das Bayes'sche Entscheidungskriterium (und nicht z.B. ein neuronales Netz)?
    ⇒ Wegen der Optimalität bei minimaler Fehlerrate habe ich gesagt, aber er wollte - glaube ich - noch mehr über Fehlerraten hören.
  - Wie kann man jetzt die freien Parameter schätzen?
    ⇒ Maximum-Likelihood- oder Momentenmethode.
- Maximum-Likelihood-Training:
  - Wie funktioniert das?
  - Maximum-Likelihood-Funktion für Gauß hinschreiben.
  - Was ist das Ergebnis für die Schätzwerte?
    ⇒ Ergebnis für μ und σ^2 hinschreiben.
- Mischverteilungen:
  - Was ist das und wie schreibt man das mathematisch formal auf?
  - Wie läuft das Training ab?
    ⇒ EM-Algorithmus (Ich sollte die grobe Funktionsweise erklären; er wollte auch noch den Zusammenhang zu Maximum-Likelihood hören; er hat mich gefragt, ob ich die Berechnungen der Parameter aufschreiben kann, konnte ich aber nicht).

Spracherkennung

- Nicht-lineare Zeitanpassung mit HMMs:
  - Wozu macht man das und wie funktioniert das?
  - Wie sieht die Gleichung für das Optimierungsproblem der Zeitanpassung aus?
  - Wie sieht nun das Training der Wortmodelle aus?
  - Was sind Emissions- und Transitionswahrscheinlichkeiten?
  - Was ändert sich bei der Erkennung von Wortketten?
    ⇒ Die unbekannten Wortübergänge müssen berücksichtigt werden.
- Wortkettenerkennung:
  - Wie funktioniert das?
  - Zeit- und Platzkomplexität für Unigramm und Bigramm.
- Großes Vokabular:
  - Motivation?
    ⇒ Wortmodelle wären bei großem Vokabular untertrainiert, ...
  - Was ändert sich jetzt?
    ⇒ Phonemmodelle statt Wortmodelle, Aussprachelexikon.
- Beam Search:
  - Motivation?
    ⇒ Geschwindigkeitsvorteil, da wenige Hypothesen aktiv (aber Suboptimalität).
  - Welche Hypothesen überleben?
    ⇒ Alle Hypothesen zum Zeitpukt t, deren Bewertung mindestens $f_{AC} \cdot Q_{max}(t)$ ist.
- Baumlexikon:
  - Was ändert sich jetzt bei der Suche?
    ⇒ Reorganisation des Aussprachelexikons, Geschwindigkeitsvorteil mit Beam-Search, Kopien für LM-Rekombination müssen jetzt schon bei Bigramm mitgeführt werden, da Wortidentitäten erst später bekannt sind.
- Sprachmodelle:
  - Training von Bigrammen (Zusammenhang zu Maximum-Likelihood).
  - Lineares Discounting.
    ⇒ Gleichung hinschreiben, Maximum-Likelihood- und Leaving-One-Out-Zusammenhang erklären.

Digitale Signalverarbeitung für Sprache und Bilder

- Fourier Transformation:
  - Definition für den kontinuierlichen Fall hinschreiben.
  - Wozu benutzt man sie?
    ⇒ Übergang vom Zeitbereich in den Frequenzbereich, Faltungssatz.
- Allgemein:
  - Wofür braucht man die digitale Signalverarbeitung in der Spracherkennung? Woran ist man interessiert?
    ⇒ Merkmalsvektor aus Signal erstellen, der die spektralen Eigenschaften des Signals gut wiederspiegelt.

- Was macht man konkret?
  ⇒ Short Time Analysis: Windowing, FT, Cepstrum-Berechnung (Transformation in Frequenzbereich → reeller Logarithmus → inverse Transformation zurück in den Zeitbereich; was das Cepstrum eigentlich ist, musste ich nicht erklären).

- DFT:
  - Warum benutzt man nun die DFT?
    ⇒ Computer arbeitet mit diskret abgetasteten Werten, ...

- Nyquist-Frequenz:
  - Welche Abtastfrequenz benutzt man?
    ⇒ Nyquist-Frequenz $\Omega_S \geq 2 \cdot \omega_B$.
  - Warum?
    ⇒ Rekonstruktion des kontinuierlichen Signals möglich, also kein Informationsverlust.

Fazit: Die Prüfung lief ziemlich entspannt ab. Ich hatte genug Zeit auf die Fragen zu antworten und wurde dabei auch nicht unterbrochen. Wenn ich mit meinen Ausführungen fertig war, guckte Herr Ney sich an, was ich auf die Blätter geschrieben hatte und fragte bei den Dingen nochmal detailierter nach, die ich etwas ungenau oder fehlerhaft hingeschrieben hatte. Prinzipiell fragt er immer erstmal sehr allgemein und geht dann ins Detail. Dabei will er oft die wichtigen Formeln wissen, aber nicht deren komplette Herleitung. Zu Beginn der Prüfung sagte er, dass sich jeder Prüfling die Reihenfolge der Themen aussuchen kann und ich entschied mich für die Spracherkennung. Das war ein sehr nettes Angebot, wurde allerdings etwas dadurch torpediert, dass die zweite Frage direkt zur Mustererkennung überging. Andererseits sind die Übergänge zwischen den zwei Themen ja auch recht fließend. Oft war ich etwas ungenau, wenn es ins Detail ging. Für eine 1.x sollte man sich daher möglichst wenig mit den Indizes in den Formeln vertuen und die Zusammenhänge zwischen den Vorlesungen auch gut erfaßt haben. Ich wußte z.B. auch oft nicht so genau, wo das Maximum-Likelihood-Kriterium überall reinspielt.