Einführung

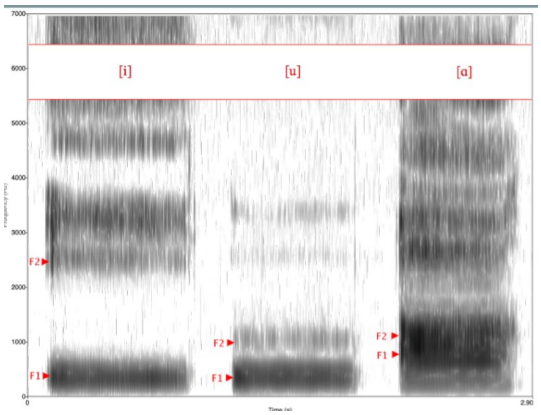
1. Nennen Sie einige Anwendungen für Automatische Spracherkennung!
    - Sprachassistenten
    - Werkerführung, Vocie Picking
    - Semi- Automatische Video Textunterschriften
    - Diktiersysteme
2. Warum ist Spracherkennung schwierig?
    - **Variabilität** auf verschiedenen Ebenen:
        - Das gleiche Wort klingt oft unterschiedlich (Sprecher, Alter, Dialekt,...)
        - Eine bestimmte Bedeutung kann verschieden ausgedrückt werden
    - **Ambiguität** (Mehrdeutigkeit) auf verschiedenen Ebenen:
        - Oft ist Hintergrund und Kontextwissen zum Verständnis erforderlich
3. Welche Gründe gibt es für den Einsatz von automatischer Spracherkennung?
    - Freihändig, keine Bewegungseinschränkungen
    - Geräte können sehr klein gebaut werden
    - Effizienter und schneller (in der richtigen Umgebung)
    - Kann im dunkeln benutzt werden
    - Für Menschen mit Behinderung
    - Geräte müssen nicht berührt werden (z.B. Klinischer Einsatz)
4. Warum ist gesprochene Sprache nicht immer das geeignetste Mittel, um mit Computern zu interagieren?
    - Manche Aktionen (z.B. Programm ausführen) sind mit wenigen Klicks schneller und effizienter lösbar
    - Können Privatsphäre der Nutzer verletzen, stören ggf. Kollegen oder andere Personen im Raum
    - Naive Menschen denken sie interagieren mit intelligenten Maschinen -> Fehlgeschlagene Interaktion und frustrierte Nutzer

Phonetische Grundlagen

5. Wie kann man sich die menschliche Sprachproduktion vorstellen?
    - Lunge generiert Luftstrom (Luftdruck)
    - Stimapparat bestehend aus Stimmbändern und Kehlkopf schwingen mit Grundfrequenz (Männl. <180Hz, Weibl. >180Hz)

- Nase- und Mundraum, sowie Zunge und Lippen modulieren Luftstrom zu gewünschtem Klang

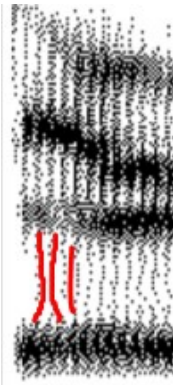
6. Wie funktioniert die Schallwahrnehmung im menschl. Ohr?
   - Aussenohr: Luftkanal, Trommelfell schwingt, wandelt Schalldruck in mechanische Schwingungen um
   - Mittelohr: Hammer, Amboss, Steigbügel: Übertragen und Verstärken Bewegung auf Membran (Ovales Fenster) in Gehörschnecke
   - Innenohr: Membran überträgt Schwingungen auf Flüssigkeit, verteilt Schwingungen in Gehörschnecke (Cochlea), Basilar Membran wird bei verschiedenen Frequenzen an verschiedenen Stellen angeregt zu schwingen. Haar Zellen in der Membran regen Nervensignale an, die vom Gehirn verarbeitet werden.
7. Welcher Frequenzbereich ist für Menschen hörbar?
   - Mensch kann im Bereich 20Hz - 20kHz hören.
   - Mensch kann 24 kritische Frequenzbänder unterscheiden (logarithmische Filterbank vgl. Mel bzw. Bark Skala)
8. Wodurch entsteht die Sprachgrundfrequenz?
   - Stimapparat bestehend aus Stimmbändern und Kehlkopf schwingen mit Grundfrequenz
9. In welchem Bereich liegt etwa die Sprachgrundfrequenz von Männern/Frauen?
   - Männder: <180Hz
   - Frauen: >180Hz
10. Nach welchen beiden messbaren Größen kann man die Vokale recht gut unterscheiden?
    - Nach den ersten beiden Formanten (die ersten beiden Maxima im gegletteten Spektrogramm)



11. Was ist ein Phonem?
    - Die kleinste Klang-Einheit einer sprache, die zwei Wörter unterscheidet.
    - Nicht alleine durch Klang, sondern auch durch Funktion definiert.

- Phoneme können unterschiedlich erzeugt werden z.B. Ra**ch**e, Ku**ch**en und Mil**ch** enthalten verschiedene Erzeugungen des Phonems /x/.

12. Was iste in Allophon?
    - Menge an gesprochenen Klängen, die ein Phonem bilden. Einzelnes Allophon kann mehrere Phoneme erzeugen
    - Bsp.: in Deutsch kann das allophon [x] benutzt werden um Rache (Phonem /x/) und Kragen (Phonem /r/) zu erzeugen.
13. Was versteht man unter Koartikulation?
    - Angrenzendes Phonem/Buchstabe klingt im aktuellen Laut mit. Aktueller Laut wird aufgrund angernzender Laute verändert (meist weil einfacher auszusprechen z.B.: **dt** wird zu **d**)
14. Wozu dient Prosodie bzw. Intonation?
    - Prosodie: Zusätliche Information durch Betonung
      - Kann Bedeutung eines Wortes festlegen (z.B. **um**fahren vs. um**fahren**)
    - Satzart festlegen:
      - Frage vs. Aussage
    - Intonation:
      - Pausen und Wortdehnung, Tonhöhe, Melodie und Rhytmus, Druck beim Reden
15. Was versteht man unter Formanten?
    - Die ersten beiden Maxima im gegletteten Spektrogramm (s. Frage 10)
16. Wie kann man aus dem Spektrum eines Signals die Grundfrequenz ermitteln
    - Den niedrigsten Peak im Spektrum
    - Den Abstand der Oberwellen im Spektrum
17. Wie kann man aus einem Breitbandspektrogramm die Grundfrequenz ermitteln?
    - Vertikale Linien zeigen Perioden in der Grundfrequenz an



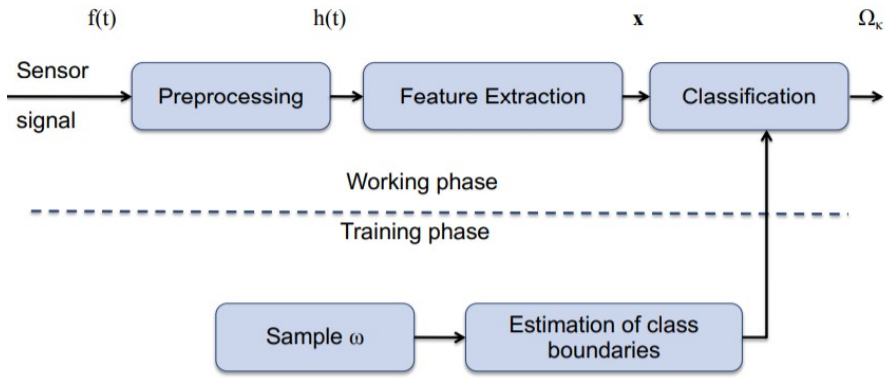
18. Wie kann man aus einem Schmalbandsprektrogramm die Grundfrequenz ermitteln?

- Frequenz der niedrigsten Horizontale Linie (nur wenn verfügbar)
- Besser: Abstand der Oberwellen (horizotale Linien)

19. Wie kann man direkt aus einem zeitsignal die Grundfrequenz eines Sprachsignals ermitteln?
    - Niedrigste Frequenz ermitteln z.B. mit Periodendauer zw. Nulldurchgängen

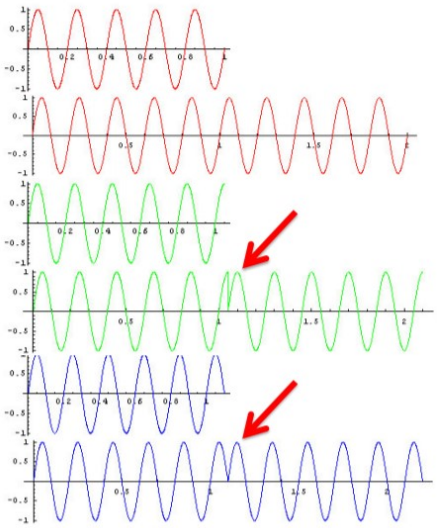
Mustererkennung

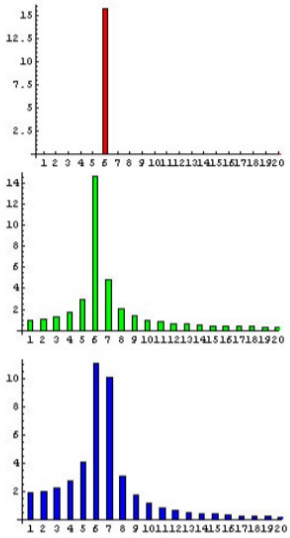
20. Womit beschäftigt sich das Gebiet der Mustererkennung
    - Der Prozess der automatischen Umwandlung eines Sensorsignals in eine aufgabenspezifische symbolische Beschreibung
21. Was verteht man unter der Klassifikation einfacher Muster (simple patterns)?
    - Das muster ist die Repräsentation eines Objekts der realen Welt
    - Jedes Objekt ist Mitglied genau einer Klassen
    - Die Anzahl der Klassen ist endlich
    - Beispiele:
        - Gesichtserkennung
        - Sprechererkennung
        - Buchstabenerkennung
22. Beschreiben Sie den grundsätzlichen Aufbau eines Klassifikationssystems!



Digitalisierung und Merkmale

23. Welche grundlegenden Entscheidungen müssen getroffen werden, bevor ein analoges Signal digitalisiert wird?
    - Abtastrate (Shannon Theorem beachten!)
    - Quantisierung
24. Was besagt das Abtasttheorem? Beispiel?
    - Es muss mit der doppelten Frequenz der höchsten im Signal vorkommenden Frequenz abgetastet werden um Aliasing zu verhindern. Tiefpassfilter im Analogen Signal erforderlich.
    - Beispiel: $F_{max} = 8kHz \rightarrow F_0 = 2 \cdot F_{max} = 16kHz$
25. Welche Form hat i.d.R. die Kennlinie eines mit 8 Bit quantisierten Signals und warum?
    - Logarithmische Quantisierung (a-Law, μ-Law) (d. Amplitude, nicht verwechseln mit Mel bzw. Bark Filterbank)
    - Amplituden in Sprache ins etwa exponentialverteilt $\rightarrow$ kleine Amp lituden sind häufiger als große, hier ist eine höhere Auflösung also sinnvoll.
    - Logarithmische Quantisierung resultiert in etwa Gleichverteilung der Amplituden
    - Das Ergebnis ist eine signifikant bessere Audioqualität, jedoch immernoch schlechter als bei linearer 16-Bit Quantisierung
26. Worin liegt der vorteil eines mit 8-Bit quantisierten Signals gegenüber einem mit 16-Bit quantisierten Signal?
    - Die Datenmenge ist halb so groß
27. Wie entsteht der sogenannte Leck-Effekt (spectral leakage) und wie läasst er sich reduzieren? Beispiel?





- Entsteht beim Ausschneiden eines Fensters und Anwendung der DFT. DFT nimmt an, dass das Signal ausschließlich aus periodischen Signalen besteht, daher kommt es an den Rändern zu Sprüngen, da das Ausgangssignal nicht aus ausschließlich periodischen Signalen besteht.
- Wird durch Anwendung von Fensterfunktion (Hamming, Hannding)-Window reduziert. Diese Dämpfen die Periode des Signals an den Rändern des Fensters

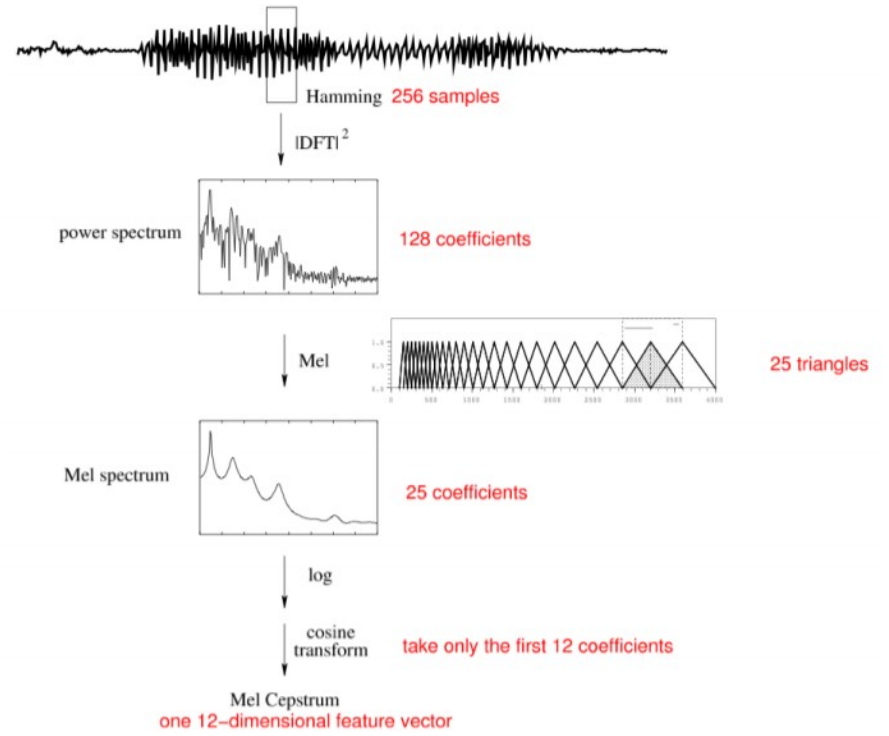
28. Welche typische Fenstergröße nutzt man in der automatischen Spracherkennung? Was wären die Vor- und Nachteile breiterer bzw. schmalerer Fenster?

    - Breitband Spektrogramme:
        - 32 oder 64 Samples Fenster
        - Geringe Auflösung der Frequenzen, hohe Auflösung der Zeit
        - Vertikale Linien signalisieren Grundfrequenz
        - Formanten können abgelesen werden
    - Schmalband Spektrogramme:
        - 256 oder 512 Samples Fenster
        - Hohe Auflösung der Frequenzen, geringe Auflösung der Zeit
        - Horizontale Linien (Bänder) zeigen die Oberschwingungen
        - Grundfrequenz kann einfach berechnet werden

29. Was ist der Unteschied zwischen einem Breitband- und Schmalbandspektrogramm und wozu nutzt man diese?

    - Breitband:
        - Erkennung einzelner Phoneme bzw. deren Formanten
    - Schmalband:
        - Erkennung von Grundfrequenz bzw. deren Verlauf
    - Siehe Frage 28.

30. Welche Merkmale werden in der automatsichen Spracherkennung überwiegend eingesetzt? Wie errechnet man sie?
    - Koeffizienten des Mel-Frequenz-Cepstrums
    i. Spektrum berechnen
    ii. Mel-Filterbank anwenden
    iii. Auf logarithmisches Spektrum die diskrete Kosinus Transformation anwenden
    iv. Die ersten 12 oder 13 Koeffizienten betrachten



31. Welche Merkmale ergänz man, um den zeitlichen Verlauf der MFCCs besser zu erfassen? Wie errechnet man sie?
    - Ableitungen 1 und 2 der MFCCs hinzufügen. Ergibt 13 statische und 26 dynamische Merkmale
    - Dyn. Merkmale geben Entwicklung im Zeitverlauf an

Klassifikation

32. Welche Verfahren zur Klassifikation von Merkmalvektoren kennen Sie und wodurch zeichnen sich diese aus?
    - Parameterlose Klassifikation: Alle Trainingssamples werden zur Klassifikation verwendet (z.B. nearest neighbour)

- Verteilungsfreie Klassifikation: Die Parameter der Grenzfunktionen werden explizit bestimmt (z.B. linearer Klassifikator, support vector machine)
- Probabilistische Klassifikation: Probabilistische Dichtefunktionen werden zur Modellierung von Klassen genutzt (z.B. gaussian mixture model)
- Neurale Netzwerke (eigentlich verteilungsfreier Klassifikator): Regionen im Merkmalsraum werden durch Training von neuronalen Netzen bestimmt

33. Beschreiben Sie den Nächster-Nachbar-Klassifikator:

    - Für jedes Sample im Trainings-Datensatz wird der Abstand (z.B. euklidisch) zum Test-Sample berechnet. Die Klasse vom Trainings-Sample mit dem geringsten Abstand wird übernommen

    > Normieren bei Euklidschem-Abstand nicht vergessen, um einen gleichen Einfluss aller Merkmale zu garantieren

34. Welche Formel ist bei der statistischen Klassifikation von zentraler Bedeutung? Erläutern Sie diese!

    - Bayes Formel hier: Wahscheinlichkeit, dass X zur Klasse Ω gehört, berechnet sich aus a priori Wahrscheinlichkeit der Klasse Ω · Wahrscheinlichkeit von X in Abhängigkeit von Ω / Grundwahrscheinlichkeit von X über alle Klassen
    - a priori Wahrscheinlichkeit: $p(\Omega_k)$
    - Bedingte Dichtefunktion: $p(X|\Omega_k)$
    - a posteriori Wahrscheinlichkeit: $p(\Omega_k|X)$

$$p(\Omega_k|X) = \frac{p(\Omega_k) \cdot p(X|\Omega_k)}{p(X)}$$

- Beispiel für Bayes Formel
    - Einer in 250000 Menschen ist Terrorist: $p(\Omega_k) = p(terrorist) = 0.000004$
    - Ein System erkennt mit 90% Genauigkeit $p(positive|terrorist) = 0.9$ und einem Prozent false-positive Rate $p(positive|not\ terrorist) = 0.01$
    - Wie hoch ist die Wahrscheinlichkeit, dass eine inhaftierte Persion auch Terrorist ist?

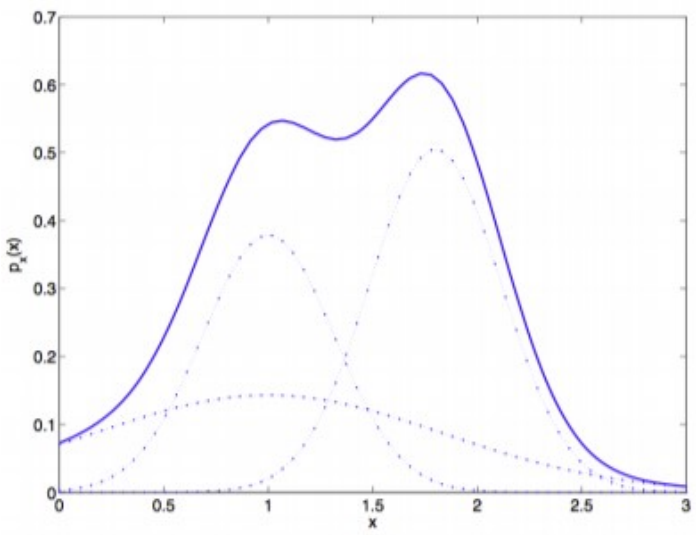
$$p(terrorist|positive) = \frac{p(terrorist) \cdot p(positive|terrorist)}{p(positive)}$$

$$p(terrorist|positive) = \frac{p(terrorist) \cdot p(positive|terrorist)}{p(terrorist) \cdot p(positive|terrorist) + p(not\ terrorist) \cdot p(positive|not\ terrorist)}$$

$$= \frac{0.000004 \cdot 0.9}{0.000004 \cdot 0.9 + 0.999996 \cdot 0.01} = 0.00036$$

- Bayes Regel
    - Der Nenner vom Bayes Theorem ist unabhängig der Klasse und kann daher für die Klassifikation ignoriert werden
    - $p(\Omega_k)p(X|\Omega_k) = max_\lambda\ [p(\Omega_\lambda)p(X|\Omega_\lambda)]$
    - Bayes Regel ist die Klassifikation mit der kleinsten Fehlerwahrscheinlichkeit

35. Welche Verfahren kennen Sie, mit denen man aus Sichproben von Merkmalsvektoren unüberwacht Kodebücher schätzen kann?
    - k-Means Algorithmus
    - EM-Algorithmus
36. Wie funktioniert der k-Means Algorithmus?
    - Wahl von k initialen Centroiden im Merkmalsraum (z.B. zufällig k Merkmalsvektoren selektieren)
    - Abstand zwischen allen Merkmalsvektoren und den Centroiden berechnen. Jeder Merkmalsvektor wird dem Cluster mit dem nähsten Centroiden zugeordnet
    - Neue Centroiden als Mittelpunkt (Mittelwerte) der Cluster berechnen
    - Wiederholen bis Kriterium erfüllt, z.B. Centroide****n sich nicht mehr ändern
37. Was versteht man unter einer Gaußschen Mischerteilung (GMM)
    - Alle Feature-Vektoren werden aus einer Kombination verschiedener Gauß-Verteilungen geneiert. (z.B. Feature F wurde zu 60% von G1, 30% von G2 und 10% von G1 produziert)



38. Wie funktioniert der EM-Algorithmus zur Kodebuchschätzung?
    i. Initialisiere Mittelpunkte und Standardabweichungen (zufällgi) aus Feature-Vektoren, ähnlich k-Means
    ii. Berechne für jeden Feature-Vektor für jede Klasse eine Gewichtung (die Wahrscheinlichkeit dar Klasse abhängig vom Feature-Vektor)
    iii. Mittelpunkte/Standardabweichungen, sowie Wahrscheinlichkeiten der Klassen aktualisieren. Darin gehen **ALLE** Feature-Vektoren, abhängig von der vorher berechneten Gewichtung ein
    iv. Wiederholen, bis Abbruchkriterium erfüllt

Deep Learning

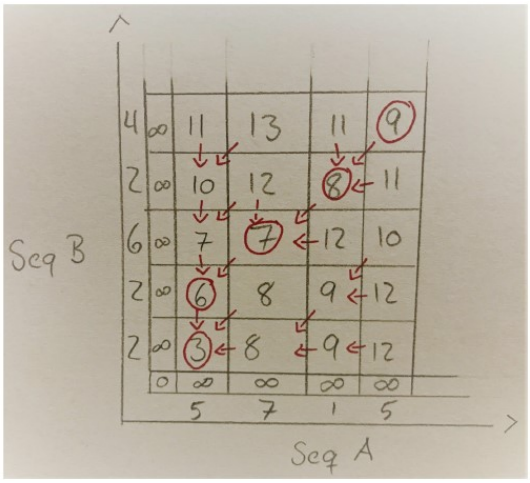
1. Was ist ein Perzeptron?
    - Vereinfachte Simulation eines biologischen Neurons
    - Besitzt gewichtete Eingänge, eine **nichtlineare** Aktivierungsfunktion und einen Bias

$$f = \sum_{0}^{i}(w_i \cdot x_i) - b$$

2. Wie sehen die SChwellwert-Funktionen (Aktivierungsfunktionen) bei küntlichen neuronalen Netzen aus?
    - Ursprüunglich: Sprunkfunktion 0,1, heute Sigmoid, Tanh, Relu,...
    - Müssen differenzierbar sein, um Training via Backpropagation Algorithmus zu ermöglichen.
3. Was versteht man unter einem Feed-Forward-Netzwerk?
    - Ein NN ohne Rekursive Verbindungen, lediglich in eine Richtung.
4. Was versteht man uner einem MLP?
    - Multilayer Perzeptron: NN organisiert aus aufeinanderfolgenden Schichten.
5. In welcher Weise wirkt sich die Verwendung von neuronalen Netzen auf die Wahl von geeigneten Merkmalen für die Spracherkennung aus?
    - Aufbereitung der Merkmale wird minimiert, NN lernt besser und schneller als der Mensch.
    - Trend: immer "rohere" Daten als Input; Direkt Daten aus FFT oder rohes Audiosignal können verwendet werden.
    - Große Datenmengen erforderlich z.B. mehrere Tausend Stunden Sprache.

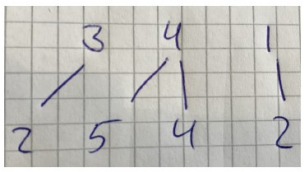
Dynamic Time Warping

6. Wozu dient der DTW-Algorithmus?
    - Erkennen von einzelnen Wörtern
7. Erläutern Sie den DTW-Algorithmus
    i. Seq A = x-Achse (Testwort Werte), Seq B = y-Achse (Trainingswort Werte)
    ii. Jeweils erste Zeile und Spalte mit ∞ initialisieren. Ursprung hat Wert 0
    iii. Distanz der Gegenüberstehenden Zahlen berechnen + Minimum von links, unten oder linksunten



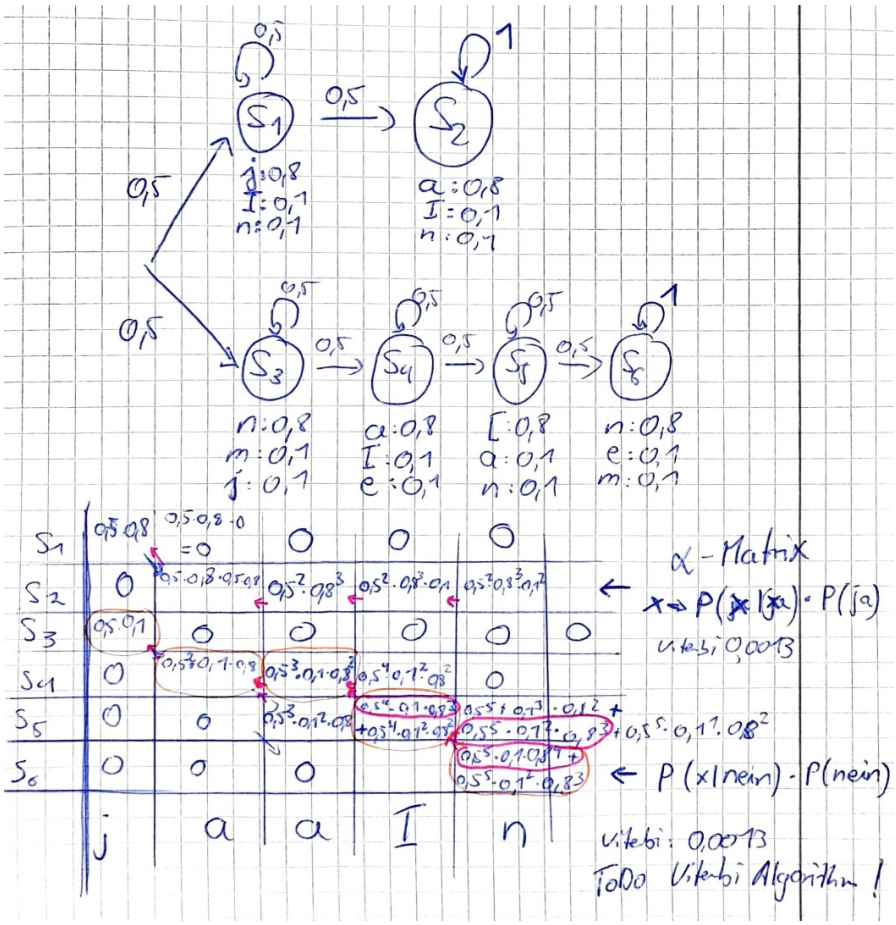
8. Wie erhält man die zeitliche Zuordnung zwischen dem Test- und dem Referenzsignal? Rechen Sie ein kurzes Beispiel durch, z.B. d(3-4-1,2-5-4-2)?

1	∞	4	6	5	**3**
4	∞	3	**2**	**2**	4
3	∞	**1**	3	4	5
	0	∞	∞	∞	∞
		2	5	4	2



Hidden-Markov-Modelle+

9. Welche Parameter besitzt ein diskretes HMM?
    - π: die Startwahrscheinlichkeiten der Anfangspfade
    - A: die Übergangswahrscheinlichkeiten (zwischen allen Zuständen)
    - B: die Ausgabewahrscheinlichkeiten (für jeden Zustand und jedem generierten Symbol bzw. Phonem)



10. Welche Arten von HMMs haben wir noch kennengelernt und worin unterscheiden sich diese?
    - DNN-HMM: Ausgabewahrscheinlichkeiten von B werden duch neuronales Netzwerk bestimmt
    - GMM-HMM: Ausgabewahrscheinlichkeiten von B werden durch Gauß Mixture Model bestimmt
11. Welche HMM-Topologien sind für die Spracherkennung geeignet?
    - Left-to-right model: Von links nach rechts übergänge zu allen Zuständen möglich
    - Bakis model: von links noch rechts Übergänge jeweils zum nächsten und übernächsten Zustand möglich
    - Linear model: von links nach rechts Übergang nur zum nächsten Zustand möglich
12. Was versteht man unter der Produktionswahrscheinlichkeit und wie lässt sich diese naiv errechnen?
    - Die Wahrscheinlichkeit $P(X|\lambda)$, dass $X = x_1, ..., x_T$ vom HMM λ generiert wrude.
    - Berechnen: Summe über alle (Anfangs-Pfade mal das Produkt aller Übergangs- und Ausgabewahrscheinlichkeiten, die das Wort bilden)
    - Oder einfach: Berechnen aller Pfade, die das gesuchte wort abbilden können (brute Force!)
13. Wie ist die Grundidee eines effizienten Algorithmus zur Bestimmung der Produktionswahrscheinlichkeit?

- Forward-Berechnung mit α-Matrix
- Backward-Berechnung mit β-Matrix
- Funktioniert nur, wenn gesamte Wortinformation bereits vorhanden ist, also nicht live!

14. Wozu dient der Viterbi-algorithmus und worin besteht seine Grundidee?

    - Die Berechnung der Zustandssequenz $q* = q_1, ..., q_T$ mit maximalem $P(q_1, ..., q_T|X, \lambda)$
    - Abwandlung von Forward-Algorithmus, welcher **Maximierung** stattt Summierung **im Rekrustionsschritt** nutzt.

15. Beschreiben Sie die Grundidee der Schätzung von HMM-Parametern anhand einer Beobachtungsfolge (Stichprobe)?

    - Kann iterativ durch Abwandlung von EM-Algorithmus berechent werden.
    - Baum-Welch-Training (Kombination von Forward- und Backward-Algorithmus):
        - α-Matrix berechnen
        - β-Matrix berechnen
        - Iterativ vorgehen

16. Was versteht man unter einer Maximum-Liklihood-Schätzung?

    - z.B. bei Baum-Welch-Verfahren:
        - Versucht Parameter zu finden, welcher die max. Wahrscheinlichkeit bei den Trainingsdaten erhält
        - Versucht Parameter zu finden, der die Wahrscheinlichkeit maximiert, dass die Stichprobe vom Modell erzeugt wurde $\rightarrow$ den Parameter bestimmen, der die Trainingsstichprobe maximal wahrscheinlich macht
        - Gefahr: zu sehr auf Trainingsdaten angepasst

17. Welche Wortuntereinheit nimmt man i.d.R. für die Spracherkennung? Erläutern Sie die Idee dahinter!

    - Triphone:
        - Einzelner Laut (Phonem), beeinflusst von seinen beiden Nachbarn
        - Beispiel: "Sieben" klingt wie "Sie[bn]": Phonem *I* hat Vorgänger *s* und Nachfolger *e* $\rightarrow$ Triphon: s/I/e.
        - Jeweils HMM mit 3 Zuständen pro Phonem
        - Theoretisch $40 \cdot 40 \cdot 40 = 64000$ verschiedene triphone notwendig um 40 verschiedene Phoneme zu repräsentieren
        - Da viele Kombinationen praktisch nicht auftreten genügen etwa 10000

Sprachmodelle

18. Was versteht man unter einem N-Gramm?

- Tupel aufeinanderfolgender Worte mit N Elementen

19. Wie erhält man ML-Schätzwerte für N-Gramme?
    - Zähler: Anzal der vorgekommenen Kombinationen: #("to Chicago") = 2
    - Nenner: Anzahl des Vorgängerwortes: #("to") = 4
    - Ergebnis: $P(”Boston”|”to”) = \frac{2}{4} = 0.5 = 50\%$
20. Warum sind diese in der Praxis von Nachteil?
    - Nicht existierende Kombinationen werden nie erkannt
    - Durch limitierte Textlängen ist die Anzal der N-Gramme limitiert. Viele N-Gramme kommen niemals in einem Trainingstext vor.
    - N-Gramme sind extrem ungleichmäßig verteilt. Zipf's Law.
21. Wie lassen sich die Schätzwerte glätten?
    - Laplace-Smooting:
        - Zähler von ML-Schätzwert + 1
        - Nenner von ML-Schätzwert + L (Gesamtanzahl d. untersch. Wörter im Text)
    - Jeffreq(-Perks)-Smoothing:
        - Wie Laplace, nur mit $\frac{1}{2}$ und $\frac{L}{2}$
    - Backoff-Smoothing: Länge der N-Gramme kürzen, wenn kein Vorkommen
22. Welche weitere Möglichkeit gibt es die Parameterzahl zu reduzieren?
    - Wortkategorien (z.B. Städtenamen, Personen-namen, Wochentage, ...)
    - Word-Embedding (Mit neuronalem Netz nach Aussage)
23. Wie errechnet sich die sog. Test-Set-Perplexität einer Stichprobe, gegeben ein Sprachmodell?
    - 10-Ziffern Vokabular:
    - m ist Länge des Test-Samples
    - w = "eins fünf drei zwei fünf"; $P(w) = \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{10^5}$
    - $PP(w) = \frac{1}{10^5}^{-\frac{1}{5}} = \frac{1}{\sqrt[m]{P(w)}}$
24. Wie kamm man diese interpretieren?
    - Je kleiner die Perplexität, desto einfacher ist die Spracherkennung.
    - Sehr gute Perplexität 10, Siri ca. 250
25. Wie kann man mit Sprachmodellen Themen (topics) klassifizieren?
    - Themaabhängige N-Gramm Sprachmodelle mit entsprechenden Texten (News, Sport, Business,...) trainieren.
    - Perplexitäten zu unbekanntem Test-Text berechnen.
    - Für das thema mit der niedrigsten Perplexität entscheiden.

26. Wie kann man mit Sprachmodellen Sprachen klassifizieren?
    - Sprachspezifische Sonderzeichen ersetzen (*ue* statt *ü*)
    - Dann wie Frage 25.
27. Wie lassen sich rekurrente neuronale Netze auf Sprachmdodelle einsetzen?
    - RNNs können die Wahrscheinlichkeit von Wortsequenzen schätzen.
    - Davor: Word-Embedding, also Wort wird in Merkmalsraum nahe ähnlichen Wörtern gruppiert.
    - RNNs haben den Vorteil, dass die Kontextgröße unbegrenzt ist.

Dekodierung

28. Was ist ein konfluenter Zustand (confluent state)?
    - Zustand ohne Symbol, wird verwendet um HMM neu zu starten für Wortketten
    - Reduzieren die KOmplexität
29. Was versteht man unter Beam Search (Strahlsuche)?
    - Unwahrscheinliche Pfade werden bei jedem Schritt verworfen
    - Gefahr: Ausschluss von ggf. passenden Pfaden schon zum Beginn, weil erstes Teilwort nicht passt
30. Wie kann man bei sehr großen Wortschätzen den Wortschatz sinnvoll organisieren?
    - Prefix Pronounciation Tree:
        - Baumstruktur nach aufeinanderfolgenden Phonemen, logarithmische Suchzeit
31. Welche weitere Methode zur Beschleunigung des Dekodiervorgangs haben wir kennengelernt?
    - Mehrphasen Dekodierung:
        - bi-Gramm Graph: Gitterstruktur (Word Lattice), wenige Pfade
        - tri-Gramm Sprachmodelle
32. Welche beiden heuristischen Parameter fürht man ein, die von der reinen Lehre der Bayes-Formel abweichen?
    - Insertion penalty ρ: Faktor, der die Wahrscheinlichkeit von Wortgrenzen verringern soll, um längere Wörter zu erzwingen. Typischer Wert: $\rho = 10^{-6}$
    - Sprachmodell Gewichte LW: Erhöht den Einfluss des Sprachmodells auf das Ergebnis. Typischer Wert: $LW = 4$

End-to-End Deep Learning

33. Erläutern Sie die Grundidee von DeepSpeech!

- Ein RNN bekommt Sequenzen von Merkmalen (FFT Koeffizienten) als Eingabe und liefert Sequenzen von Buchstaben als Ausgabe
- RNN besitzt verbindungen "in die Vergangenheit und Zukunft" und kann daher erst am Ende der Gesamten Eingabesequenz ein Ergebnis liefern

34. In welcher Weise können Language Models bei diesem Ansatz in den Erkennungsvorgang einbezogen werden?

    - Zur Fehlerkorrektur der Ausgaben bzw. um das Ausgabeergebnis zu verbessern.