

Sprachverarbeitung: Übung 11

Sprachmerkmale zur Lautunterscheidung

Aufgabe 1: Cepstrale Glättung

Eine grobe Approximation des Leistungsdichtespektrums eines Sprachsignalabschnittes kann auf zwei Wegen erzeugt werden, mittels der LPC-Analyse oder durch cepstrale Glättung des Foursierspektrums. Es stellt sich somit die Frage: Wodurch unterscheiden sich diese Spektren?

Untersuchen Sie diese Frage anhand zweier Sprachsignalabschnitte des Lautes [i], die von einem Mann bzw. von einer Frau gesprochen worden sind. Gehen Sie dabei wie folgt vor: Multiplizieren Sie das Signalsegment mit einem leistungskompensierten Hamming-Fenster¹ der Länge 240 und wenden Sie die Fouriertransformation (Auflösung: 512) an, um das Leistungsdichtespektrum zu ermitteln. Um das Spektrum zu glätten, wird daraus das reelle Cepstrum berechnet, dieses mit einem cepstralen Fenster der Länge $L_c = 20$ multipliziert und schliesslich die DFT angewendet. Das cepstrale Fenster hat die Form:

$$w_c(n) = \begin{cases} 1 & \text{für } 0 \leq n \leq L_c \\ 0 & \text{für } L_c + 1 \leq n \leq N - L_c - 1 \\ 1 & \text{für } N - L_c \leq n \leq N - 1. \end{cases}$$

Ermitteln Sie nun noch das LPC-Spektrum. Verwenden Sie dafür die Matlab-Funktionen `lpc` (Prädiktorordnung 16) und `freqz`.

Um die drei Spektren gut miteinander vergleichen zu können, werden alle ins gleiche Koordinatensystem gezeichnet. Welche Unterschiede zwischen dem cepstral geglätteten Spektrum und dem LPC-Spektrum stellen Sie für die Männerstimme bzw. die Frauenstimme fest? Wie wirkt sich das Verändern der Prädiktorordnung aus?

Um weniger programmieren zu müssen, können Sie den Programmrahmen `ueb11_1_frame.m` im Directory **Gegebenes** verwenden und an den mit `%>>` markierten Stellen ergänzen.

Aufgabe 2: Mel-Spektrum

Für die Spracherkennung wird häufig das Mel-Cepstrum als Sprachmerkmal verwendet (vergl. Buch, Abschnitt 11.7.1). Die Matlab-Funktion `ueb11_2(filnam,dispfb)` lädt von der Datei

¹Ein leistungskompensiertes Hamming-Fenster ist $w_n = 1.588 w$, wobei w ein mit Gleichung (16) auf Seite 66 definiertes Hamming-Fenster ist (vergl. auch Übung 4, Aufgabe 2).

`filnam` ein Signalsegment, stellt es als Figur 1 dar, berechnet das Fourierspektrum, das Mel-Spektrum und die cepstral geglätteten Spektren und zeichnet sie als Figur 2. Mit dem Argument `filnam`, das die Form `'seg_phone_L_SN.wav'` hat, wird bestimmt, welches Lautsegment zu verwenden ist. Mit `L`, `S` und `N` wählen Sie den Laut `L = a|e|i|o|u|l|m|n|f|s`, den Sprecher `S = A|B` (Männer- oder Frauenstimme) und die Beispielnnummer `N = 1|2|...|9`.

Die Parameter der Funktion `ueb11_2`, insbesondere die Zahl der Kanäle der Mel-Filterbank `nfilt` und die Länge des cepstralen Fensters `cepwinlen`, sind so eingestellt, wie sie üblicherweise in der Spracherkennung verwendet werden. Selbstverständlich können Sie diese Parameter verändern und deren Wirkung beobachten. Um auch noch die Mel-Filterbank angezeigt zu erhalten, müssen Sie das Argument `dispfb` auf 1 setzen. Die Anzeige erfolgt dann als Figur 3.

Schauen Sie die geglätteten Spektren verschiedener Lautbeispiele an, und charakterisieren Sie wiederum die wesentlichen Unterschiede zwischen dem geglätteten Fourierspektrum, dem Mel-Spektrum und dem geglätteten Mel-Spektrum. Achten Sie insbesondere auf die Frequenzauflösung der Spektren und auf den Einfluss der Grundfrequenz bei der männlichen bzw. der weiblichen Stimme.

Selbstverständlich müssen Sie nicht alle 180 Lautbeispiele verwenden, sondern nur ein paar pro Sprecher, damit Sie die wesentlichen Unterschiede der dargestellten Spektren feststellen können.

Aufgabe 3: Güte von Sprachmerkmalen

Sie haben bisher etliche Sprachmerkmale kennen gelernt, z.B. die Autokorrelationsfunktion, die LPC-Koeffizienten, das Mel-Spektrum und zwei verschiedene Arten von Cepstren, nämlich das DFT-Cepstrum und das Mel-Cepstrum. Grundsätzlich lässt sich jedes dieser Merkmale für die Spracherkennung verwenden, wobei zu erwarten ist, dass sich nicht alle gleich gut dafür eignen.

Die Tauglichkeit eines Sprachmerkmals für die Spracherkennung kann grob abgeschätzt werden, indem man untersucht, wie gut das Merkmal Laute zu unterscheiden vermag. Dafür reicht nicht das Merkmal allein, sondern es ist noch ein Distanzmass nötig. Der Einfachheit halber wollen wir hier die euklidische Distanz verwenden, mit welcher die Distanz d zwischen zwei Merkmalsvektoren \mathbf{c}_1 und \mathbf{c}_2 mit je J Elementen berechnet wird als:

$$d = \sqrt{\sum_{j=0}^{J-1} [c_1(j) - c_2(j)]^2} \quad (1)$$

Eine Distanz zwischen zwei Merkmalsvektoren, die von verschiedenen Lauten stammen, wird nachfolgend als Kreuzdistanz bezeichnet. Distanzen zwischen Merkmalsvektoren gleicher Laute werden Eigendistanzen genannt. Anhand eines Sprachmerkmals können Laute umso besser unterschieden werden, je grösser die Kreuzdistanzen und je kleiner gleichzeitig die Eigendistanzen sind.

Man kann also die Tauglichkeit eines Sprachmerkmals abschätzen, indem man die Eigen- und die Kreuzdistanzen miteinander vergleicht. Dazu werden aus einer Sammlung von Lautsegmenten die Eigen- und Kreuzdistanzen ermittelt und je als Histogramm dargestellt, wie dies das Matlab-Skript `ueb11_3_frame.m` für das Merkmal Mel-Cepstrum macht. Es gilt: je mehr die Histogramme gegeneinander verschoben sind, desto besser ist das Sprachmerkmal.

Wenn Sie das Matlab-Skript `ueb11_3_frame.m` starten und durch Eingabe von 1 das Mel-Cepstrum als Merkmal auswählen, dann sehen Sie, dass sich die Bereiche der Eigen- und Kreuzdistanzen stark überlappen. Sie werden feststellen, dass dies auch für andere Sprachmerkmale der Fall ist. Ob ein Sprachmerkmal besser ist als ein anderes, ist deshalb aufgrund der Histogramme oft schwierig oder gar nicht zu beurteilen.

Besser lässt sich der Unterschied der Eigen- und Kreuzdistanzen anhand der Fisher-Distanz (wird in der Literatur oft als *Fisher ratio* bezeichnet)

$$d_f = (m_c - m_s)^2 / (\sigma_c^2 + \sigma_s^2) \quad (2)$$

beurteilen, wobei m_s und m_c die Mittelwerte der Eigen- und Kreuzdistanzen sind und σ_s^2 und σ_c^2 deren Varianzen. Die Fisher-Distanz ist umso grösser, je stärker sich die Eigen- und Kreuzdistanzen unterscheiden. Sie ist null, wenn das Sprachmerkmal gar nichts zur Unterscheidung von Lauten beiträgt.²

Nebst der Fisher-Distanz ermittelt das Matlab-Skript auch die Mittelwerte der Eigen- und Kreuzdistanzen für alle vorhandenen Laute und Lautpaare und gibt sie in einer Matrix-Darstellung aus. Daraus ist ersichtlich, welche Laute wie stark streuen (Werte auf der Diagonalen) und welche Laute sich besser bzw. schlechter unterscheiden lassen.

Sie können nun die oben genannten Sprachmerkmale auf Tauglichkeit prüfen bzw. mit dem Mel-Cepstrum vergleichen, indem Sie das Matlab-Skript `ueb11_3_frame.m` an der Stelle erweitern, wo die Merkmalsvektoren berechnet werden (beim `case`-Statement).

Zur Ermittlung der verschiedenen Sprachmerkmale verwenden Sie die Matlab-Funktionen `rceps`, `mel_spectrum`, `acf` und `lpc`. Nehmen Sie der Einfachheit halber bei allen Merkmalen dieselbe Vektorlänge (d.h. bei der Autokorrelation die ersten `lenvc` Werte und beim Mel-Spektrum eine Filterbank mit `lenvc` Filtern).

Vergleichen Sie die Fisher-Distanzen für die verschiedenen Sprachmerkmale. Überlegen Sie, welche Laute sich aufgrund der ausgegebenen Resultate mit dem jeweiligen Sprachmerkmal gut bzw. schlecht unterscheiden lassen.

²Mit dem Matlab-Skript `ueb11_3_frame.m` ist dies dann ersichtlich, wenn anstelle der Sprachmerkmale Zufallsvektoren verwendet werden, was bei Eingabe von 0 der Fall ist.