ERKENNUNG VON SPRACHE AUS EINEM PARAMETRISIERTEN SIGNAL MIT HILFE DER DIGITALEN BILDVERARBEITUNG

Lukáš Džbánek, Andy Kuiper, Milan Sigmund

Technische Universität Brno, Institut für Radioelektronik
dzbanek@feec.vutbr.cz

Abstract: In diesem Beitrag wird ein Ergänzungsverfahren (EV) zur automatischen Worterkennung vorgestellt. Das entwickelte Verfahren und der Basis-Algorithmus, das sogenannte „Dynamic Time Warping“, wurden verwendet und erzielte Ergebnisse für ein reines Sprachsignal, sowie für ein verrauschtes Sprachsignal miteinander verglichen.

1 Einleitung

Heutzutage sind zwei Methoden zur automatischen Spracherkennung am weitesten verbreitet. Erste basiert auf „Hidden Markov Models“ (HMM) und wird hauptsächlich bei den Systemen eingesetzt, welche mit einem großen Wortschatz arbeiten (phonetisch basiert ~1000 Wörter). Die zweite Methode, das sogenannte „Dynamic Time Warping“ (DTW), ist für kleinere Systeme geeignet (~10 Wörter).

In diesem Beitrag wird eine weitere mögliche Variante beschrieben, die auf der digitalen Bildverarbeitung basiert. Dieses Verfahren wurde als Teil des gesamten Systems für eine Gerätesteuerung per Sprache in der Medizin erforscht. Sind die beiden erstgenannten Methoden für den medizinischen Bereich nicht hinreichend zuverlässig, besteht der Bedarf eine neue spezielle Alternative zu entwickeln. Einzelne Wörter werden durch konventionelle Methoden segmentiert und parametrisiert (Cepstrum, Mel-Cepstrum, LPC usw.). Die Resultatmatrix wird als ein Bild betrachtet und durch spezielle Algorithmen weiter verarbeitet, welches dann mit den Musterbildern verglichen wird.

Eine Einsatzmöglichkeit dieser Methode kann in Systemen mit kleinem Wortschatz (~10 Wörter) gesehen werden. Diese Tatsache führte also dazu, den entwickelten Algorithmus mit der DTW-Methode zu vergleichen.

2 Beschreibung der Methode

Die Methode besteht aus 4 Hauptblöcken (Abb.1).

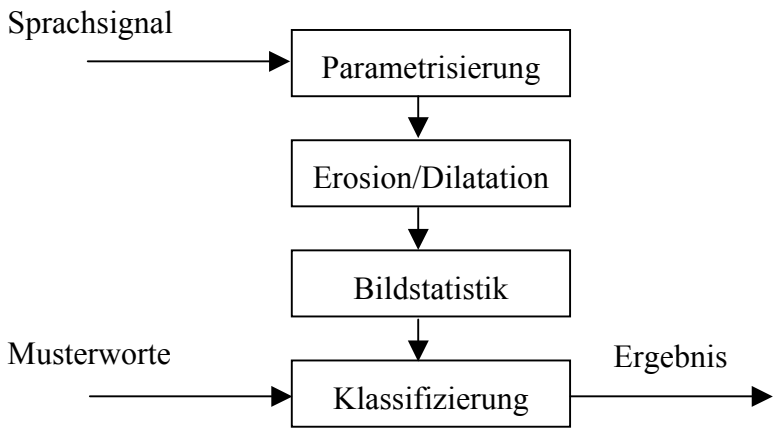


Abbildung 1 - Blockschaltbild der Methode

Das Sprachsignal wird parametrisiert und als 2D-Bild dargestellt. Dieses Bild wird weiter durch spezielle Algorithmen verarbeitet und alle nicht dominanten Teile des Bildes werden eliminiert. Es wird eine Statistik von diesem Bild erstellt und diese dann mit den Mustersignalen verglichen.

2.1 Parametrisierung

Alle Sprachsignale wurden mit einer Abtastfrequenz von 8kHz und einer 16-Bit Auflösung aufgenommen. Gespeichert wurden diese dann als „wav"-Datei (Standard Windows-Format).

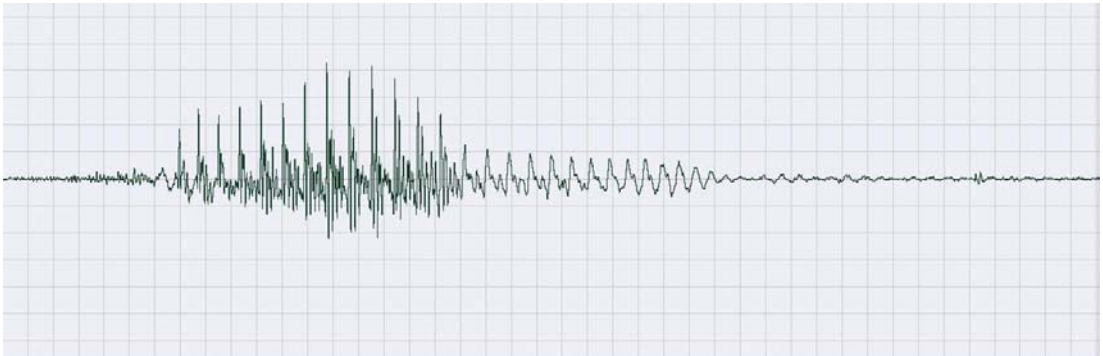


Abbildung 2 - Beispiel eines aufgenommenen Wortes („Hand"). Abtastfrequenz 8kHz, Auflösung 16-bit. Länge der Aufnahme: 420ms (Zeitverlauf).

Das Sprachsignal wird durch Cepstral-Koeffizienten (nach Mel-Scala) parametrisiert [4] und als 2D-Bild dargestellt (Abb.3).

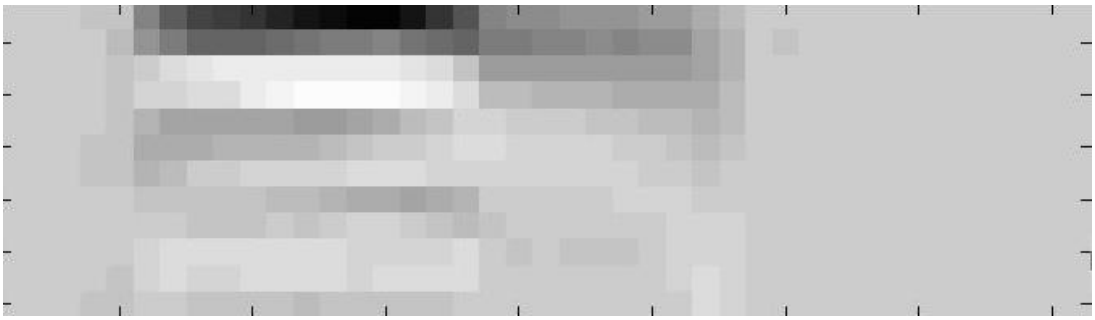


Abbildung 3 - Parametrisiertes Signal als 2D-Bild. Zeitverlauf Cepstral-Koeffizienten (von oben nach unten 1-12) – Segmentweise. Länge 42 Segmente. Segmentlänge: 20ms, Überlappung: 10ms. Mel-Cepstrum: 12 Koeffizienten – 15 Mel-Frequenz Fenstern. FFT Länge: 256 Abtastwerte.

Die Parametrisierung wird jeweils zur Wortgruppe passend gewählt, so dass der Abstand immer optimal wird [1]. Um die Zuverlässigkeit möglich hoch zu halten ist es nötig, eine optimale Parametrisierung zu jeder Wortgruppe zu finden.

2.2 Binarisierung, Erosion und Dilatation

Von dem parametrisierten, als Bild dargestelltem Signal (Abb.3), wird jetzt nur die wichtigste Information extrahiert. Dieses geschieht durch das sog. „EDGE Detection" [2], [3]. Es handelt sich um ein Set von verschiedenen Kantensuche-Funktionen. Die am weitest verbreiteten sind die sog. „Sobel" und „Prewitt" EDGE Detektoren. Letztgenannte wurde in diesem Algorithmus verwendet. Prinzipiell geht es um eine 2D-Faltung mit der folgenden Matrix

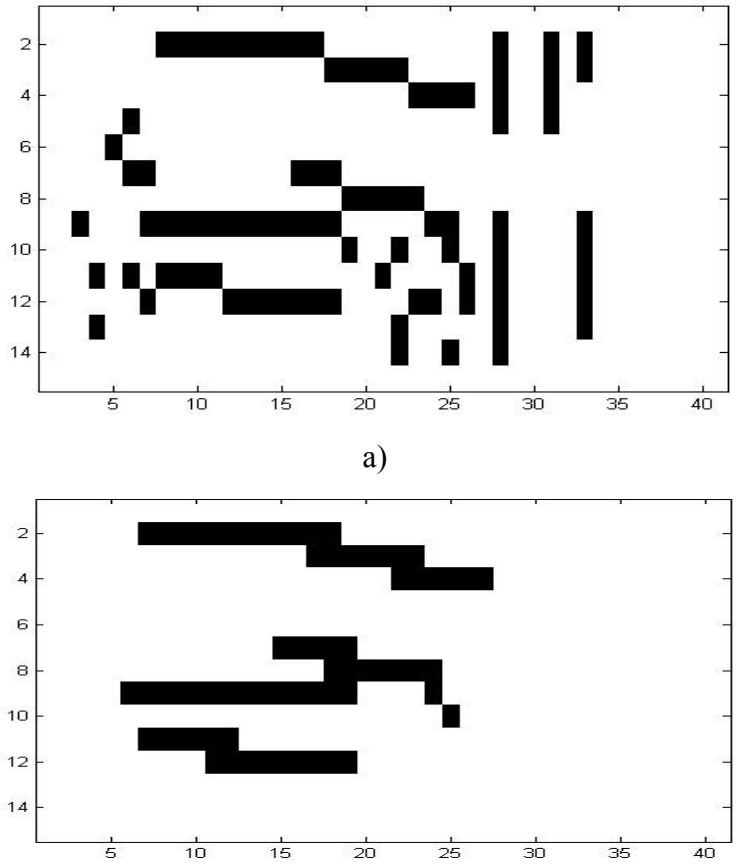
$$\begin{bmatrix} 1 & 0 & -1 \\ c & 0 & -c \\ 1 & 0 & -1 \end{bmatrix} \tag{1}$$

für die horizontale Richtung und für vertikale Richtung nimmt man

$$\begin{bmatrix} 1 & c & 1 \\ 0 & 0 & 0 \\ -1 & c & -1 \end{bmatrix} \tag{2}$$

wobei c=1 für Prewitt und c=2 für Sobel Verfahren.

Das Ergebnis der Faltung wird durch eine Schwellenfunktion binarisiert, so enthält man folgende Bilder, die man durch Algorithmen wie „Erosion“ und „Dilatation“ weiter verarbeitet so, dass von dem ganzem Bild (parametrisierten Sprachsignal) nur ein Muster übrig bleibt (Abb.4). Einstellungen der Funktionen „Erosion“ und „Dilatation“ hängen leider von der ausgewählten Parameter ab.



a)

b)

Abbildung 4 – Darstellung der Bildverarbeitung. a) nach der Binarisierung (Prewitt Operator, Schwelle = 0,001), b) nach den „Dilatations“ und „Erosions“ Methoden

2.3 Bildstatistik

Alle Sprachsignale, die zu bewerten sind, werden auf diese Weise (Kapitel 2.1-2.3) verarbeitet. Ergebnisse kann man in der Abb.5 sehen. Eine Musterinformation ist von den

Wörtern „Hand“ – Abb.5 a) und den Wörtern „Haut“ Abb.5 b) extrahiert. Diese werden dann in eine Statistik umgewandelt.

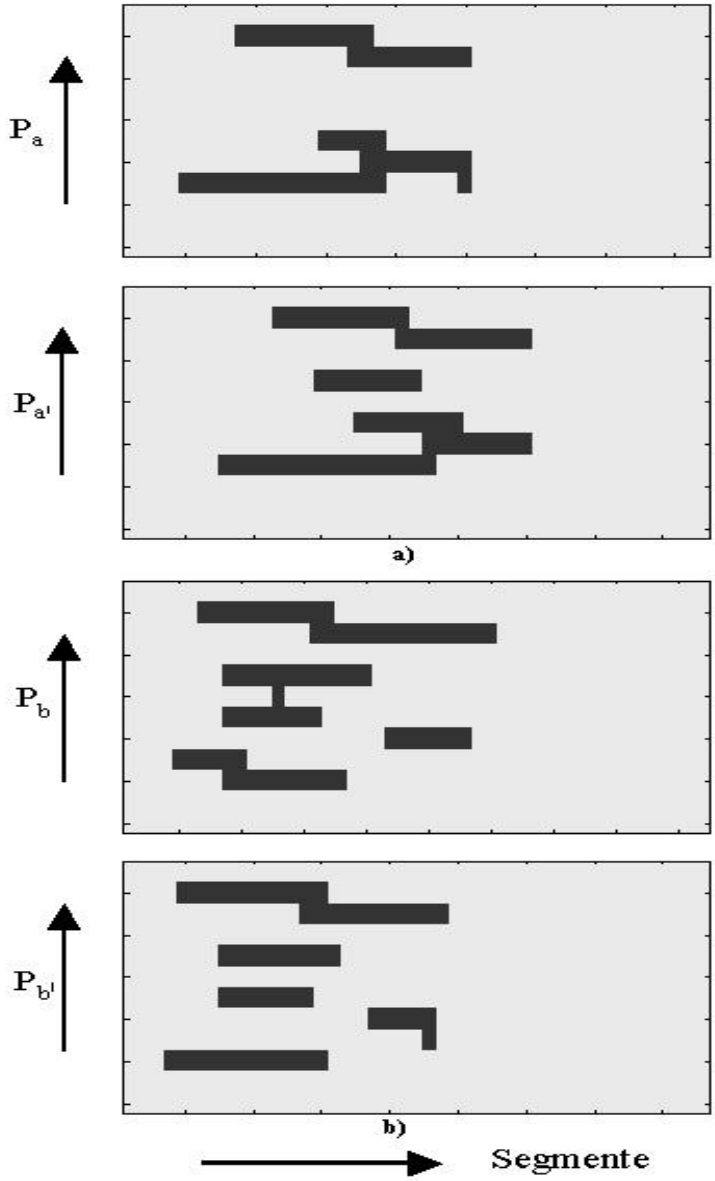


Abbildung 5 – Beispiel von extrahierten Merkmalen der Wörter (ausgewählte Beispiele) von oben nach unten 2 Wiederholungen des Wortes „Hand“ a) und des Wortes „Haut“ b), das Zeichen „P“ steht für Parameter)

Grundidee der Statistik der Bilder basiert auf der sog. Grauwertlauflängeanalyse. In dieser Methode wurde die Anzahl Pixel, je nach Länge, Richtung und Intensität berechnet. Im binarisierten Bild gibt es nur eine „Farbe“, deswegen arbeitet die entwickelte Methode nur mit den Längen und Richtungen der Linien. Durch die Statistik wird der Zeitverlauf aufgelöst und es wird keine weitere Zeitanpassung benötigt (wie z.B. bei DTW).

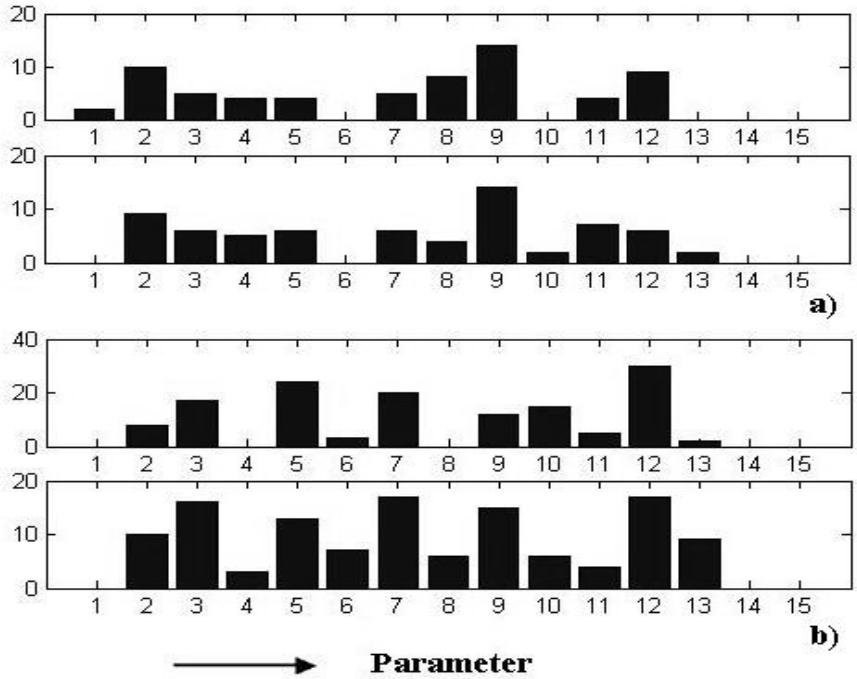


Abbildung 6 – Statistik der Wörter aus Abb.5. Von oben nach unten 2 Wiederholungen des Wortes „Hand“ a) und des Wortes „Haut“ b). Die Graphen sind Zeitlich unabhängig, auf der x-Achse sind einzelne Parameter der Wörter.

2.4 Euklidischer Abstand

In Abb.6 sieht man die Ähnlichkeiten der Graphen a) und b), bzw. der Unterschiede zwischen a) und b). Um diese numerisch zu bestimmen, wird ein euklidischer Abstand benutzt

$$D = \sum_i \left(\mathbf{P}_\mathrm{T}(i) - \mathbf{P}_\mathrm{M}(i)\right)^2 \tag{3}$$

wobei D ein euklidischer Abstand ist, $\mathbf{P}_\mathrm{T}$ ein Vektor von getesteten Parametern und $\mathbf{P}_\mathrm{M}$ ein Vektor der Musterparameter ist.

3 Ergebnisse

Es wurden die Abstände zwischen 10 Wörtern berechnet, jeweils 5 Repräsentationen der Worte „Hand“ und „Haut“. Diese werden für die entwickelten Methoden, sowie für die DTW Methode berechnet. So enthält man für jedes Wort 9 Werte. Vier Abstände zu demselben Wort und 5 zu dem anderem. Die Ergebnisse lassen sich durch eine Normalverteilung approximieren [5]. Danach enthält man die Werte

$\mu_\mathrm{P}, \mu_\mathrm{N}$ Mittelwerte der Abstände

$\sigma_\mathrm{P}, \sigma_\mathrm{N}$ Standard-Abweichung (STD) - Varianz

wobei der Buchstabe „P“ für den Fall derselben Wörter (P-positiv) steht und der Buchstabe „N“ für die anderen (N-negativ).

Das getestete Wort wird richtig erkannt, falls folgende Formel gilt:

$$0 > D_\mathrm{min} - D_\mathrm{eigen} \tag{4}$$

wobei D_eigen der Abstand zum Referenzmuster der wahren Klasse und D_min der kleinste Abstand der übrigen Klassen, ist.

Mit der Kenntnis der Formel für die Normal-Dichte lässt sich folgende Formel ableiten:

$$u = \frac{\mu_N - \mu_P}{\sqrt{\sigma_P^{\ 2} + \sigma_N^{\ 2}}} \tag{5}$$

wobei u ein Parameter der $f(u)$ tabellierte Verteilungsdichtefunktion für Normal-Dichte ist und diese dann der geschätzten Erkennungsrate (Tab.1) entspricht [6].

Typ des Sprachsignals	**Erkennungsrate $f(u)$ [%]**	
	DTW	**EV**
reines Signal	99,7	81,2
verrauschtes Signal (100ms)	98,0 ± 1,1	81,2 ± 0,2
verrauschtes Signal (200ms)	92,8 ± 3,5	81,4 ± 0,4

Tabelle 1 – Ergebnisse und Vergleich der klassischen DTW-Methode mit dem Ergänzungsverfahren (EV). Dargestellte Werte sind Erkennungsraten in Prozent mit Varianzen (berechnet durch 7 Messungen). Mit dem Begriff „verrauschtes Signal“ ist die falsche Wortgrenzenbestimmung gemeint – Zufallsgenerator der Falschwortgrenzenbestimmung – der Anfang der Wörter wird um <-100,0>ms (bzw. Ende um <0,100>ms) falsch erkannt (dasselbe für 200ms).

4 Zusammenfassung und Ausblick

Aus den Ergebnissen von Tab.1 ist ersichtlich, dass die EV-Methode wesentlich schlechtere Ergebnisse als die DTW-Methode bringt. Es gibt jedoch im Vergleich zur DTW-Methode einen wesentlichen Vorteil. Die EV-Methode ist (siehe Ergebnisse) sehr robust gegen eine schlechtere Wortgrenzenbestimmung (DTW-Methode hängt sehr stark von der Wortgrenzenbestimmung ab).

Die entwickelte Methode ist zur Zeit noch nicht anwendungsbereit (die Erkennungsrate ist nicht hinreichend), zeigt aber einen möglichen Weg für Erkennungssysteme mit kleinerem Wortschatz (~10 Wörter).

Literatur

[1] Dzbanek, L.: Programm zur Darstellung der Parametrisierung eines Sprachsignals. In: KONVENS 2004. Wien 2004 (angenommen).

[2] Pitas, I.: Digital image processing algorithms and applications. New York: A Wiley-interscience publication 2000 (ISBN 0-471-37739-2), pp. 242 – 249.

[3] Bose, T.: Digital signal and image processing. New York: Addison-Wesley Publishing company 2004 (ISBN 0-471-32727-1), pp. 570 – 586.

[4] O´Shaughnessy, D.: Speech Communication. Human and Machine. New York: A Wiley-interscience publication 1987 (ISBN 0-201-16520-1), pp. 420 – 423.

[5] Vary, P., Heute, U., Hess, W.: Digitale Sprachsignal-Verarbeitung. Karlsruhe: B.G. Teubner Stuttgart 1998 (ISBN 3-519-06165-1), pp. 147 – 152.

[6] Lamos, F., Potocky, R.: Pravdepodobnost a matematicka statistika. Bratislava: Alfa 1989 (ISBN 80-05-00115-0), pp. 316 – 317.

Dieses Projekt wurde vom Ministerium für Bildung, Jugend und Sport (FRVS 1619/2004) und der GACR (Czech Science Foundation) No. 102/03/H109 "Methods, Structures and Components of Electronic Wireless Communication" gefördert.