MERKMALSEXTRAKTION FÜR DIE SPRACHERKENNUNG

Christian Lüke, Karl Schnell

Institut für Angewandte Physik, Goethe-Universität Frankfurt
Max-von-Laue-Straße 1, 60438 Frankfurt am Main
lueke@thecreation.de, schnell@iap.uni-frankfurt.de

Kurzfassung: Ein Problem für die Spracherkennung stellen die Variationen der Sprachaufnahmen dar, die einerseits auf die Variationen des Sprachsignals selbst und andererseits auf unterschiedliche Aufnahmeumgebungen zurückzuführen sind. Um die Spracherkennung gegenüber variierenden äußeren Einflüssen robuster zu gestalten, können diese in den Modellen der Mustererkennung berücksichtigt werden. Ein anderer Ansatz besteht darin, die Merkmalsextraktion der Spracherkennung so zu gestalten, dass sie weniger sensitiv bezüglich dieser Variationen ist. Eine ebenso einfache wie wirksame Methode, die Merkmalsextraktion robuster zu gestalten, stellen die so genannten Normierungsmethoden wie die CMN und CVM dar. In diesem Beitrag werden Erweiterungen der Standard-Normierungsmethoden vorgestellt. Damit ist es insbesondere möglich, stationäre und instationäre Bereiche unterschiedlich zu bewerten. In einer ersten Evaluierung bestätigen verbesserte Erkennungsraten diesen Ansatz.

1 Einleitung

Für die Spracherkennung wird das zu verarbeitende Sprachsignal üblicherweise in eine Folge von Merkmalsvektoren transformiert. Als Standard-Merkmale gelten die MFCCs (mel-frequency cepstral coefficients), die als eine parametrisierte Darstellung der Einhüllenden des Sprachspektrums angesehen werden können. Um die Merkmalsextraktion für die Erkennung robuster zu gestalten, können die einzelnen Merkmale und/oder die ganze Merkmalssequenz einer Äußerung nachverarbeitet werden. Als eine effektive Methode der Nachverarbeitung ganzer Merkmalsvektorsequenzen haben sich die so genannten Normierungsmethoden herausgestellt [1, 2]. In [3] wurde eine Variation der CMN (cepstral mean normalization) vorgestellt, die unterschiedlich gewichtete Mittellungen für die Berechnung ausnutzt. Aufbauend auf diesen Ergebnissen werden in diesem Beitrag Erweiterungen der Merkmalsextraktion präsentiert und evaluiert. Zunächst betreffen diese die einzelnen Merkmale selbst, auf die eine Limiter-Funktion angewandt wird. Weiterhin werden die Normierungsmethoden in der Weise erweitert, dass sie unterschiedliche Gewichtungen für stationäre und instationäre Bereiche des Sprachsignals aufweisen. Dies kann für die CMN wie für die CVM (cepstral variance normalization) realisiert werden. Damit soll insbesondere für kurze Sprachäußerungen die für die Normierung eigentlich zu geringe Datenmenge besser ausgenutzt werden. Die einzelnen Erweiterungen der Merkmalsextraktionsschritte enthalten Parameter, die bezüglich einer Einzelwort-Erkennung optimiert werden. Der für die Optimierung verwendete Korpus berücksichtigt dabei unterschiedliche Aufnahmebedingungen.

2 Merkmalsextraktion

Die Merkmalsextraktion beschreibt die gesamte Abbildung vom Sprachsignal zu einer Folge von Merkmalsvektoren, die für die Mustererkennung verwendet werden. Zunächst wird jede

aufgezeichnete Äußerung mittels einer automatischen Gain-Normierung vorverarbeitet, die sich an der maximalen Amplitude orientiert. Das amplitudennormierte Sprachsignal wird in Segmente von 46 ms Länge und 29 ms Überlapp aufgeteilt, die jeweils mit einem Hamming-Fenster gewichtet werden. Für die Merkmalsextraktion werden zunächst die Melfrequenz-Cepstralkoeffizienten (MFCCs) berechnet. Dabei werden die MFCCs konventionell mittels der DCT der Mel-Frequenzbänder berechnet, die eine Bandbreite von insgesamt 11 kHz umfassen. Anschließend wird die Norm eines jeden MFCC-Vektors $\vec{x}$ durch eine nichtlineare Funktion eines Limiters angepasst. Dies kann für die Mustererkennung Vorteile erbringen, da der Abstand zweier Merkmalsvektoren auch von ihrer Norm abhängig ist. Die Limiter-Funktion wird durch Gleichung (1) beschrieben und ist von den Parametern w_g und w_L abhängig

$$\hat{\vec{x}} = \begin{cases} \vec{x}\left(\dfrac{(1-w_g)}{w_L} + \dfrac{w_g}{\|\vec{x}\|}\right) & \|\vec{x}\| < w_L \\ \dfrac{\vec{x}}{\|\vec{x}\|} & \|\vec{x}\| \geq w_L . \end{cases} \tag{1}$$

Die Limiter-Funktion begrenzt die Norm $\|\vec{x}\|$ eines jeden MFCC-Vektors $\vec{x}$ auf w_L. Ist die Norm kleiner als w_L, so wird sie mit Hilfe des Parameters w_g skaliert. Der nächste Verarbeitungsschritt ist eine Sprachaktivitätserkennung bzw. voice activity detection (VAD), die schwellwertbasiert implementiert ist. Die Anfangs- bzw. Endmarken der VAD werden um zwei zusätzliche Segmente nach vorne bzw. hinten verschoben. Dieser einfache Algorithmus stellte sich als ausreichend für die Evaluation der Extraktionsmethoden heraus.

Die gesamten 42-dimensionalen Merkmalsvektoren $\vec{y}$ setzen sich zusammen aus

- der logarithmischen Segment-Energie
- den Deltas der logarithmischen Segment-Energie
- 20 MFCCs
- 20 Delta-MFCCs.

Auf die MFCCs wird, wie zuvor beschrieben, die Limiter-Funktion angewandt.

3 Cepstrale Normierungsverfahren

Im Folgenden werden zwei konventionelle und zwei modifizierte Normierungsverfahren behandelt. Die Normierungsverfahren werden jeweils auf eine gesamte Äußerung angewandt.

3.1 Konventionelle Normierungsverfahren

Die cepstrale Mittelwertnormierung bzw. cepstral mean normalization (CMN) berechnet das arithmetische Mittel μ_i jeder Folge $(y_{t,i})_{t=1\ldots T}$ cepstraler Vektorkomponenten für jede Äußerung und zieht es von den Folgengliedern ab, wie in Gl. (2) beschrieben

$$\mu_i = \frac{1}{T}\sum_{t=0}^{T-1} y_{t,i}$$
$$\hat{y}_{t,i} = y_{t,i} - \mu_i \,. \tag{2}$$

Dabei ist i der Index der Vektorkomponente, t der Segment-Index, und T die Segmentanzahl der Äußerung. In langen Äußerungen können mit Hilfe der CMN stationäre Störgeräusche und Aufnahmeeinflüsse eliminiert werden. Als Erweiterung der CMN normiert die cepstrale Varianznormierung bzw. cepstral variance normalization (CVN) zusätzlich auch die Varianz jeder Vektorkomponente. Dazu wird jede Vektorkomponente durch ihre Standardabweichung geteilt, wie in Gl. (3) dargestellt

$$\sigma_i^{\,2} = \frac{1}{T}\sum_{t=0}^{T-1} (y_{t,i} - \mu_i)^2$$
$$\hat{y}_{t,i} = \frac{1}{\sigma_i}(y_{t,i} - \mu_i). \tag{3}$$

Die CVN bewirkt eine Angleichung der Wertebereiche verschiedener Vektorkomponenten, was sich auf die anschließende Mustererkennung günstig auswirkt.

3.2 Gewichtete cepstrale Normierungsverfahren

Die oben dargestellten cepstralen Normierungsverfahren sind für lange Äußerungen konzipiert. Im Falle kurzer Äußerungen enthält der Mittelwert jedoch auch spektrale Charakteristika einzelner Laute. Um diesem Problem entgegenwirken zu können, wird den instationären Segmenten der Äußerung besondere Bedeutung zugemessen. Die Instationarität des Sprachsignals wird von artikulatorischen Bewegungen des Vokaltraktes und von Veränderungen der Anregung verursacht. Ein Maß für die Instationarität des t-ten Segments stellt die Norm Δy_t des Delta-Vektors dar. Im Folgenden wird die gewichtete cepstrale Mittelwertnormierung bzw. weighted cepstral mean normalization (WCMN) beschrieben, die in [3] erstmals vorgestellt wurde. Zunächst werden auf Grundlage des Maßes Δy_t Gewichte λ_t durch

$$\Delta y_t = \left\| \vec{y}'_t - \vec{y}'_{t-1} \right\|$$
$$\lambda_t = 1 + w^{\text{norm}} \cdot \frac{\Delta y_t}{\max\{\Delta y_t\}} \tag{4}$$

berechnet. Die Vektoren $\vec{y}'_t$ können dabei entweder die vollständigen Merkmalsvektoren $\vec{y}_t$ sein, oder nur diejenigen Komponenten des Merkmalsvektors $\vec{y}_t$ enthalten, welche die MFCCs repräsentieren. Mittels des Parameters w^{norm} kann der Einfluss der Instationarität auf die Mittelwertsberechnung reguliert werden. Die Gewichte λ_t können Werte zwischen 1 und $1 + w^{\text{norm}}$ annehmen. Sie werden einerseits zur Bestimmung des gewichteten Mittelwerts $\tilde{\mu}_t$

und andererseits zur Skalierung der zugehörigen Vektorkomponenten genutzt. Die normierten Merkmalsvektoren der WCMN ergeben sich zu

$$\tilde{\mu}_i = \frac{\sum_t y_{t,i} \cdot \lambda_t}{\sum_t \lambda_t}$$
$$\hat{y}_{t,i} = y_{t,i} \cdot \lambda_t - \tilde{\mu}_i \,. \tag{5}$$

Wie bei der CVN findet auch hier eine Skalierung des Merkmalsvektors statt. Zur Skalierung wird in diesem Fall jedoch nur der Gewichtungsfaktor λ_t genutzt. Die Erkennungsrate des WCMN ist zwar höher als die der ungewichteten CMN, doch beide Verfahren werden von der CVN übertroffen. Um auch die Varianz für die gewichtete Normierung einzubeziehen, wird hier eine gewichtete cepstrale Varianznormierung bzw. weighted cepstral variance normalization (WCVN) eingeführt. Die im Folgenden vorgestellte WCVN verwendet zwei verschiedene Gewichte λ_t und φ_t zur Berechnung des Mittelwerts und der Varianz

$$\lambda_t = 1 + w_\lambda^{\text{norm}} \frac{\Delta y_t}{\max\{\Delta y_t\}}$$
$$\varphi_t = 1 + w_\varphi^{\text{norm}} \frac{\Delta y_t}{\max\{\Delta y_t\}} \,. \tag{6}$$

Die Gewichte λ_t werden analog zur WCMN für die Berechnung des gewichteten Mittelwertes $\tilde{\mu}_i$ verwendet. Die Gewichte φ_t werden mithilfe eines zweiten Parameters w_φ^{norm} bestimmt und gehen in die Berechnung der gewichteten Standardabweichung $\tilde{\sigma}_i$ ein. Die normierten Werte $\hat{y}_{t,i}$ werden berechnet, indem der gewichtete Mittelwert von den entsprechenden Komponenten des Merkmalsvektors abgezogen, und das Ergebnis durch die gewichtete Standardabweichung geteilt wird

$$\tilde{\mu}_i = \frac{\sum_t \lambda_t \cdot y_{t,i}}{\sum_t \lambda_t}$$
$$\tilde{\sigma}_i^{\,2} = \frac{\sum_t \varphi_t \cdot \left(y_{t,i} - \tilde{\mu}_i\right)^2}{\sum_t \varphi_t} \tag{7}$$
$$\hat{y}_{t,i} = \frac{\left(y_{t,i} - \tilde{\mu}_i\right)}{\tilde{\sigma}_i} \,.$$

Durch das Hervorheben instationärer Segmente bei der Berechnung des Mittelwerts und der Standardabweichung konnte die Erkennungsrate im Vergleich zur CVN bereits deutlich verbessert werden. Die Norm der Merkmalsvektoren beeinflusst die Merkmalserkennung. Führt man, wie in Gl. (8) beschrieben

$$\hat{y}_{t,i} = \frac{\left(y_{t,i} \cdot \lambda_t - \tilde{\mu}_i\right)}{\tilde{\sigma}_i}, \tag{8}$$

eine Skalierung der Merkmalsvektoren mit Hilfe des Gewichtungsfaktors λ_t durch, so haben die instationären Segmente in der Mustererkennung einen größeren Einfluss. Daher wurde bei der Bestimmung des cepstral normierten Vektors alternativ zu Gl. (7) zusätzlich der Faktor λ_t eingeführt. Für die durchgeführte Evaluierung konnte so eine weitere Verbesserung der Erkennungsrate erreicht werden.

4 Optimierung des DTW-Algorithmus

Zur Evaluierung der verschiedenen Verfahren der Merkmalsextraktion wurde der DTW-Algorithmus (Dynamic Time Warping) eingesetzt, der für spezielle Anwendungen noch Verwendung findet [4]. Der DTW-Algorithmus berechnet eine zweidimensionale Abstandskarte für ein Äußerungspaar und bestimmt den optimalen Weg durch diese Karte [5]. Besteht das Äußerungspaar aus zwei gleichen Wörtern, so weicht der optimale Weg in der Regel nur wenig von der Diagonalen ab. Werden Äußerungen von verschiedenen Wörtern verglichen, dann weicht der optimale Weg im Allgemeinen deutlich von der Diagonalen ab. Daher kann es sinnvoll sein, Beschränkungen bei der Berechnung des optimalen Weges zu definieren, um nur realistische Wege nahe der Diagonalen zuzulassen. Zum gleichen Zweck werden häufig Gewichte w^{diag} für diagonale Schritte auf der Abstandskarte verwendet, um lange horizontale oder vertikale Wege zu bestrafen. In Gleichung (9) ist die rekursive Berechnung des optimalen Weges $D_{t,s}$ zum Punkt (t,s) beschrieben

$$D_{t,s} = \min\begin{Bmatrix} D_{t-1,s} + d_{t,s} \\ D_{t,s-1} + d_{t,s} \\ D_{t-1,s-1} + d_{t,s} \cdot w^{\text{diag}} \end{Bmatrix}. \tag{9}$$

Dabei stellt $d_{t,s} = \left\| \vec{y}_t^{\,\alpha} - \vec{y}_s^{\,\beta} \right\|$ den euklidischen Abstand zwischen den Merkmalsvektoren $\vec{y}_t^{\,\alpha}$ und $\vec{y}_s^{\,\beta}$ der Äußerungen α und β dar. Die Erkennungsrate konnte durch die Einführung von Parametern w_i^{dist} verbessert werden, die in der Berechnung des Abstandes $d_{t,s}$ die Vektorkomponenten nach Gl. (10) unterschiedlich gewichten

$$d_{t,s}{}^2 = \sum_{i=1}^{42} w_i^{\text{dist}} \left(y_{t,i}^{\alpha} - y_{s,i}^{\beta}\right)^2. \tag{10}$$

Die 42 Gewichte w_i^{dist} sind konstant für alle Äußerungen und können bezüglich einer WER-Minimierung optimiert werden. Insgesamt ergeben sich je nach verwendetem Verfahren bis zu 47 Parameter, die optimiert werden können.

5 Auswertung

Der Korpus besteht aus 439 Äußerungen, die ein Vokabular von 58 Wörtern umfassen. Diese Äußerungen wurden aufgenommen in drei unterschiedlichen Umgebungen, mit verschiedenen Mikrophonen und mit verschiedenen Arten und Intensitäten von Störgeräuschen. Einige der Äußerungen wurden geflüstert gesprochen; die geflüsterten Äußerungen konnten nur mithilfe der cepstralen Normierungen erkannt werden. In Tabelle (1) sind die erreichten Erkennungsraten dargestellt. Es ist zu sehen, dass die gewichteten cepstralen Normierungsverfahren die zugehörigen Standardverfahren übertreffen.

Tabelle 1 - Wortfehlerraten

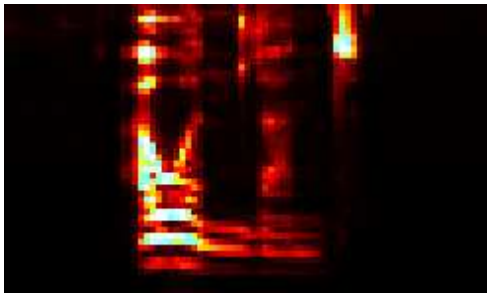
WER	w_i^{dist} fest	w_i^{dist} optimiert
ohne Normierung	9,11 %	8,66 %
CMN	8,20 %	6,38 %
WCMN	7,97 %	5,92 %
CVN	7,25 %	5,69 %
WCVN	5,47 %	5,24 %

Der Tabelle (1) kann entnommen werden, dass die Optimierung der w_i^{dist} teils zu signifikanten Verbesserungen der Erkennungsrate geführt hat. Es ist anzumerken, dass ohne den Einsatz der Limiter-Funktion die Erkennungsrate deutlich schlechter ausfiel. Wenn Äußerungen anderer Sprecher zum Korpus hinzugefügt werden, verschlechtert sich die Erkennungsrate. Es ist allerdings bekannt, dass der DTW-Algorithmus bei sprecherunabhängiger Spracherkennung schlechter abschneidet als beispielsweise Hidden-Markov-Modelle (HMM) [5].

Die Abbildungen 1 bis 4 zeigen eine Äußerung des Wortes „Auenland“ in spektraler, cepstraler und merkmals-basierter Darstellung über die Zeit; positive Werte erscheinen rötlich, negative Werte bläulich. Während in Abb. 2 die zentralen Segmente spektral flach erscheinen, hebt die Limiter-Funktion die spektrale Struktur hervor, wie in Abb. 3 zu sehen ist. In Abb. 4 sind die Merkmalsvektoren zusammen mit der Norm Δy_t (gelber Graph) dargestellt. Der Graph korreliert direkt mit den Delta-MFCCs in der oberen Hälfte der Abbildung.

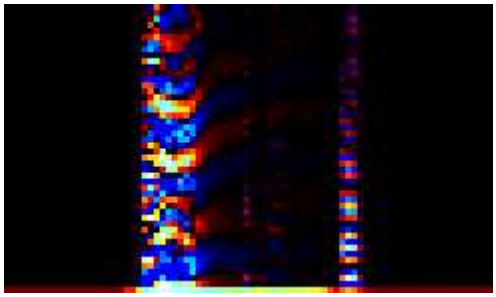
6 Zusammenfassung

Im vorliegenden Beitrag wurde ein neues Normierungsverfahren vorgestellt, das für die Merkmalsextraktion in der Spracherkennung Verwendung findet. Für diesen Zweck wurden verschiedene Normierungsverfahren mit Hilfe eines einfachen DTW-basierten Spracherkenners evaluiert. Um die Wortfehlerrate zu minimieren, wurden Parameter der Merkmalsextraktion sowie des DTW-Algorithmus optimiert. Eine erste Evaluierung mit dem einfachen DTW-Spracherkenner auf Basis eines kleinen Wortschatzes zeigte, dass die hier vorgeschlagene weighted cepstral variance normalization (WCVN) signifikante Verbesserungen im Vergleich zu den Standard-Normierungsverfahren erzielen kann.



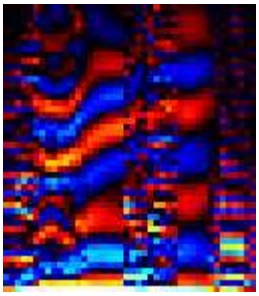
Zeit

Abbildung 1 – Mel-skaliertes Spektrogramm der Äußerung “Auenland”.



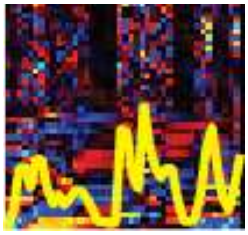
Zeit

Abbildung 2 – Cepstrogramm der Äußerung “Auenland”.



Zeit

Abbildung 3 – Cepstrogramm der Äußerung “Auenland” nach der Nachverarbeitung durch Limiter und VAD.



Zeit

Abbildung 4 – Merkmalsvektoren der Äußerung “Auenland” mit Delta-MFCCs (obere Hälfte) und normierten MFCCs (untere Hälfte); Graph der Norm Δy_t der Delta-Vektoren (gelber Graph).

Literatur

[1] Atal, B.: ‘Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification’, Journal of the Acoustical Society of America, vol. 55, pp. 1304-1312, June 1974.

[2] Droppo, J.; Acero, A.: ‘Environmental Robustness’, in: Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (Eds.) 'Handbook of Speech Processing', Springer-Verlag, 2008.

[3] Lüke, C. and Schnell, K.: 'Feature Extraction for Speech Recognition', Proc. Int. Conf. on Acoustics in Rotterdam - Joint NAG/DAGA, Rotterdam Netherlands, 2009.

[4] Abdulla, W.H.; Chow, D.; Sin, G.: ‘Cross-words reference template for DTW based speech recognition systems’, in Proc. IEEE TENCON 2003, Bangalore India, pp. 1576-1579, 2003.

[5] Rabiner, L.; Juang, B. H.: ‘Fundamentals of Speech Recognition’, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.