Modellbasierte Sprachsegmentierung

Tobias Herbig, Oliver Gaupp, Franz Gerl

Harman/Becker Automotive Systems, Söflinger Str. 100, 89077 Ulm, Deutschland, {therbig,ogaupp,fgerl}@harmanbecker.com

Einleitung

Der Einsatz automatischer Spracherkennung in stark gestörter Umgebung erfordert eine robuste Segmentierung der Sprache des Benutzers gegenüber störenden Hintergrundgeräuschen. Ein typisches Szenario für Störung stellt die Bedienung eines Spracherkenners in einer verhallten Umgebung mit einem Benutzer im Vordergrund des Systems umgeben von interferierenden Sprechern im Hintergrund sowie zusätzlichem nicht sprachlichem Hintergrundgeräusch dar. Viele bestehende Verfahren basieren auf einer Bewertung des zeitlichen Verlaufs der Signalenergie sowie der Sprachgrundfrequenz. Mit diesen Verfahren können derartige Störungen bei geringem Signal-zu-Rausch Abstand (SNR) nicht mehr zuverlässig vom aktuellen Benutzer unterschieden werden, da diese nicht über a priori Wissen über den Benutzer beziehungsweise den akustischen Hintergrund verfügen. Es wird ein Algorithmus zur Segmentierung vorgestellt, der auf einer statistischen Modellierung von Sprache unterschiedlicher Benutzer sowie Störungen im Hintergrund basiert. Realisiert wird dieses Verfahren durch Gaußsche Mischverteilungen (GMMs), die für verschiedene Umgebungen trainiert und zur Laufzeit auf den akustischen Hintergrund adaptiert werden. Durch geeignete Normierung der Likelihood sowie einer Nachverarbeitung wird eine zuverlässige Segmentierung mit geringer Verzögerung erreicht.

Modellierung

Das menschliche Gehör ist auch bei geringem SNR noch in der Lage, Sprecher im Vordergrund von Sprechern im Hintergrund bzw. einer Überlagerung von Sprechern zu unterscheiden. Entscheidend sind hierfür die Energie des Sprachsignals, die Sprachgrundfrequenz sowie die Halligkeit durch die Charakteristik der Raumimpulsantwort.

In früheren Publikationen [1], [2] konnte gezeigt werden, dass ein modellgestützter Ansatz in Kombination mit einer Standard-Merkmalsextraktion aus der Spracherkennung in der Lage ist, die charakteristischen Verteilungen dieser Ereignisse zu erfassen und zuverlässig zu klassifizieren. Demonstriert wird das vorgestellte Verfahren anhand von MFCCs (Mel-Frequency-Cepstral-Coefficients). $\mathbf{x}_t$ bezeichnet den Merkmalsvektor zum Zeitpunkt t. Eine Geräuschreduktion des Signals in der akustischen Vorverarbeitung ist nicht erforderlich.

Als statistisches Modell werden im Folgenden GMMs verwendet, da sie gute Eigenschaften bezüglich der Approximation einer Verteilungsdichte besitzen. Im Gegensatz zu diskriminativen Modellen wird nicht die Trennbarkeit sondern die Modellierung zweier Verteilungen fokussiert. Der Vorteil von GMMs liegt im einfachen Trainingsverfahren durch den EM-Algorithmus oder das K-Means Verfahren sowie in der Möglichkeit, Modelle nachträglich in Abhängigkeit vom Segmentierungsergebnis unabhängig voneinander adaptieren zu können. Ein GMM besteht aus einer gewichteten Summe aus N multivariaten Gaußverteilungen, die im Folgenden auch als Klassen mit Index i bezeichnet werden. Jede Gaußverteilung wird durch $\mathbf{\Lambda}_i = \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}$ mit Erwartungswert $\boldsymbol{\mu}_i$ und Kovarianzmatrix $\mathbf{\Sigma}_i$ eindeutig definiert. Die Gewichte w_i der einzelnen Klassen können als a priori Wahrscheinlichkeit der Klassen interpretiert werden. $\mathbf{\Theta} = \{w_1, \mathbf{\Lambda}_1, \ldots, w_N, \mathbf{\Lambda}_N\}$ bezeichnet hierbei das Parameterset des GMMs.

$$p(\mathbf{x}_t|\mathbf{\Theta}) = \sum_{i=1}^{N} w_i \cdot \mathcal{N}_i\{\mathbf{x}_t|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i\} \tag{1}$$

Im Folgenden werden drei GMMs in separaten Trainingsphasen mit unterschiedlichem Material berechnet.

Das Modell $\mathbf{\Theta}_S$ wird auf einer Vielzahl von Sprechern trainiert, um die Charakteristik eines beliebigen Benutzers zu erfassen, der das sprachgesteuerte System aus einer geringen Distanz und aus unterschiedlichen Positionen zum Mikrophon bedient. Um die Robustheit der Segmentierung zu erhöhen, wird $\mathbf{\Theta}_S$ basierend auf gestörten Äußerungen berechnet, um eine bessere Übereinstimmung der Umgebungsbedingungen im Training und zur Laufzeit zu gewährleisten.

Das zweite Modell $\mathbf{\Theta}_I$ wird auf einer Überlagerung von interferierenden Sprechern aus dem Hintergrund trainiert, während das dritte Modell $\mathbf{\Theta}_N$ an nicht-sprachliche Störungen angepasst wird. Derartige Störungen können in einer KFZ-Anwendung beispielsweise aus Fahrgeräuschen bestehen.

Je nach Komplexität des Problems kann es sinnvoll sein, weitere Modelle für charakteristische Störungen zu entwickeln. Beispielsweise erhöht die getrennte Modellierung einer Überlagerung vieler Sprecher (babble noise) bzw. einzeln hervortretender Sprecher im Hintergrund die Robustheit des Verfahrens.

Klassifizierung

Durch die Modellierung der einzelnen Störquellen bzw. der Benutzer kann die Segmentierung durch einen Vergleich der Likelihood-Funktion der Modelle (1) entscheiden, ob eine Äußerung vom aktuellen Benutzer stammt.

Im Folgenden wird zu jedem Zeitpunkt die Likelihood für alle Modelle bestimmt und durch eine Normierung ein Score S_t für die Zuordnung von $\mathbf{x}_t$ zum Vordergrundmodell berechnet. Durch die Normierung von $p(\mathbf{x}_t|\mathbf{\Theta}_S)$ auf

die gewichtete Summe der Likelihood Werte aller Modelle wird einerseits die Echtzeitfähigkeit des Systems gewährleistet und andererseits gilt $0 \leq S_t \leq 1$, wodurch einzelne Scores in der Nachverarbeitung kein zu großes Gewicht erhalten. Ähnliche Ansätze sind bereits aus der Sprecheridentifikation in der Literatur bekannt [3]. α und β bezeichnen Gewichtungsfaktoren mit der Eigenschaft $\alpha, \beta > 0$.

$$S_t = \frac{p(\mathbf{x}_t|\mathbf{\Theta}_S)}{p(\mathbf{x}_t|\mathbf{\Theta}_S) + \alpha \cdot p(\mathbf{x}_t|\mathbf{\Theta}_I) + \beta \cdot p(\mathbf{x}_t|\mathbf{\Theta}_N)} \qquad (2)$$

Anschaulich kann die Normierung durch ein globales Modell $\mathbf{\Theta}_g$ erklärt werden, das die Klassen aller drei Modelle beinhaltet. S_t entspricht der Berechnung der posteriori Wahrscheinlichkeit $\sum_{i,\boldsymbol{\mu}_{g,i} \in \mathbf{\Theta_S}} p(i|\mathbf{x}_t, \mathbf{\Theta}_g)$, dass ein Vektor $\mathbf{x}_t$ den Klassen des Modells $\mathbf{\Theta}_g$ zugeordnet wird, die aus dem Modell $\mathbf{\Theta}_S$ hervorgegangen sind.

Zur Glättung wird in überlappenden Zeitfenstern der Länge T die Differenz aus Maximum und Minimum von S_t in Relation zum Mittelwert oder Median als Maß für die Unsicherheit der weichen Zuordnung der Merkmalsvektoren interpretiert und $\tilde{S}_t$ an den Schwellwert-Entscheider weitergereicht. Hierbei darf das Zeitfenster nur wenige Zeitpunkte umfassen, damit die Segmentierung nahe Echtzeit abläuft. Zusätzlich kann eine Mindestlänge für Äußerungen und Sprachpausen gefordert werden, um einzelne Fehler der Segmentierung zu unterdrücken.

$$\tilde{S}_t = \text{median}_n\{S_n\} - (\max_n\{S_n\} - \min_n\{S_n\}) \qquad (3)$$

$$t - \frac{T}{2} \leq n < t + \frac{T}{2} \qquad (4)$$

Anhand der Segmentierung können die Modelle basierend auf der aktuellen Äußerung an die Umgebung adaptiert werden, um Differenzen zwischen dem Training und Test zu reduzieren. Für die Adaption können alle bekannten Verfahren aus der Literatur angewendet werden wie beispielsweise dem Maximum A Posteriori (MAP) Verfahren.

Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Segmentierung des Referenzsystems basierend auf der Auswertung von Energie und Sprachgrundfrequenz sowie der vorgestellten modellbasierten Segmentierung verglichen.

Im Folgenden werden nur zwei Modelle betrachtet. Ein GMM modelliert beliebige Sprecher als Benutzer des Spracherkenners, während ein zweites GMM interferierende Sprecher (babble noise) im Hintergrund wiedergibt. Die Likelihood-Funktionen in (2) werden mit $\alpha = 1$ gleich gewichtet.

Beide GMMs bestehen aus jeweils 64 Klassen mit diagonalen Kovarianzmatrizen, die auf Merkmalsvektoren bestehend aus 15 MFCCs und normierter Energie trainiert werden. Die Abtastrate beträgt 16kHz. Als Korpus wird eine Sprachdatenbank mit ungestörten deutschsprachigen Äußerungen verwendet. Mit Aufnahmen von Hintergrundgeräuschen aus Umgebungen wie Bahnhof, Kantine und Cafe sowie Messungen der Raumimpulsantworten in unterschiedlichen Abständen zum Mikrophon wird ein Trainings- und Testdatensatz erzeugt. Im Test werden andere Umgebungen als im Training verwendet, um die Robustheit der Segmentierung in unbekannten Umgebungen zu untersuchen. Im Test wird das Hintergrundmodell einmalig auf einer kurzen Aufnahme $\leq 2sec$ mittels des MAP Verfahrens adaptiert. Das SNR im Test beträgt $7 - 12\,dB$. Tabelle 1 gibt den Vergleich zwischen

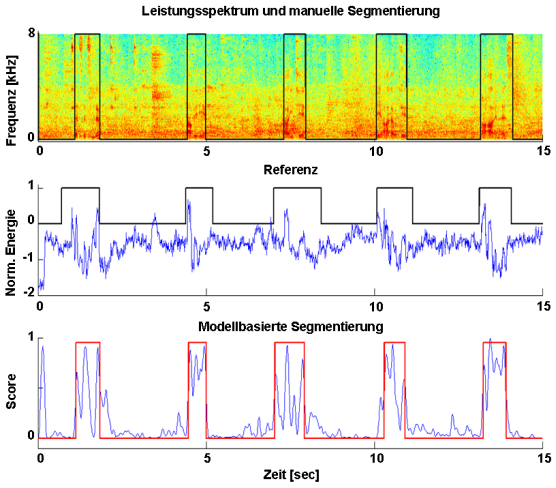

Abbildung 1: Segmentierung

	Referenz	Modellbasierte Segmentierung (ohne MAP)	Modellbasierte Segmentierung (mit MAP)
S_G [%]	41.55	70.19	91.38
R_{I} [%]	0.35	1.29	3.58
R_{II} [%]	81.63	41.30	10.09

Tabelle 1: Vergleich zwischen Referenz und modellbasierter Segmentierung

beiden Verfahren anhand der Genauigkeit S_G der Segmentierung wieder. R_{I} gibt die Fehlerrate an, bei der Hintergrundstörung erkannt wird, obwohl der Benutzer spricht. R_{II} stellt den umgekehrten Fehlerfall dar.

Literatur

[1] Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, Kiyohiro Shikano: *Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs*, Proc. INTERSPEECH-ICSLP, 2004.

[2] Wei-Ho Tsai, Hsin-Min Wang, Dwight Rodgers: *Automatic Singer Identification of Popular Music Recordings via Estimation and Modeling of Solo Vocal Signal*, EUROSPEECH-2003, 2993-2996.

[3] K. Markov and S. Nakagawa: *Frame Level Likelihood Normalization for Text-Independent Speaker Identification using GMMs*, Proc. ICSLP 96, pp 1764-1767, Philadelphia, USA, 1996.