Methoden zur robusten Spracherkennung unter Beobachtungsunsicherheit

Dorothea Kolossa

Arbeitsgruppe Kognitive Signalverarbeitung, Institut für Kommunikationsakustik, 44801 Bochum, Germany

Email: Dorothea.Kolossa@rub.de

Zusammenfassung

Automatische Spracherkennungssysteme gewinnen schnell an Verbreitung und sind unter günstigen akustischen Bedingungen inzwischen sehr zuverlässig. Sobald aber zeitvariante Störungen oder Nachhall auftreten, verlieren sie schnell an Erkennungsrate und sind daher meist nicht sinnvoll für Freisprechanwendungen in beliebigen Räumen nutzbar.

Zur Optimierung der Erkennerleistung in problematischen Umgebungen bieten allerdings sogenannte Uncertainty-of-Observation-Techniken ein großes Potential. Diese Methoden betrachten die Sprachmerkmale nicht als bekannt, sondern sehen in den extrahierten Merkmalsvektoren lediglich eine gestörte Version der sauberen Merkmale. Durch Schätzung und Verwendung der Merkmalsunsicherheit - der Schätzfehlervarianz nach Vorverarbeitung und Merkmalsextraktion - können diese Verfahren deutliche Gewinne in der Robustheit gegenüber Umgebungseinflüssen erzielen.

In diesem Übersichtsbeitrag werden aktuelle Techniken zur Spracherkennung unter Merkmalsunsicherheit vorgestellt, wobei neben bekannten Verfahren wie dem Uncertainty Decoding oder Significance Decoding auch eine unsicherheitsgesteuerte Dimensionsreduktion - die Noise Adaptive Linear Discriminant Analysis - betrachtet wird.

Einleitung

Abbildung 1 zeigt einen typischen Aufbau eines Spracherkennungssystems.

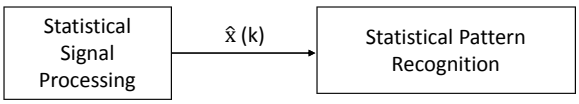

Abbildung 1: Blockdiagramm.

Hier werden als Ergebnis einer statistischen Sprachsignalvorverarbeitung oft nur Punktschätzungen an die nachfolgende Erkennerstufe weitergeleitet. Im Gegensatz dazu erfolgt die Anbindung zwischen Signalverarbeitung und Mustererkennung bei den sogenannten *Uncertainty-of-Observation-Techniken* über ein breiteres Interface: Dabei können, wie in Abb. 2 gezeigt, neben den Schätzwerten für die sauberen Sprachmerkmale auch die Varianzen dieser Schätzung übermittelt werden.

Es gibt prinzipiell zwei Arten, diese Merkmalsunsicherheiten im Decoder zu verwenden:

- zur Anpassung der Likelihood-Berechnung, die pro Frame die akustischen Beobachtungsvektoren mit den Modellen jedes Erkennerzustands abgleicht
- und zur Anpassung der Merkmalsberechnung.

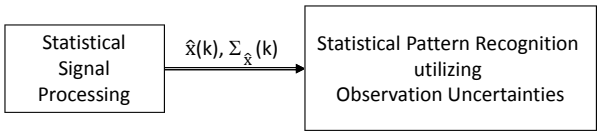

Abbildung 2: Blockdiagramm unter Verwendung von Merkmalsunsicherheiten.

Beide Varianten können das in Abb. 3 gezeigte Modell verwenden, um die Erzeugung der unsicheren Merkmale zu beschreiben. Dabei bezeichnet $\hat{x}$ den in der Vorverarbeitung (z.B. Störgeräuschreduktion) geschätzten Merkmalsvektor, von dem im folgenden angenommen wird, das er eine erwartungstreue Schätzung des sauberen Sprachsignals darstellt. $\Sigma_{\hat{x}}$ wird als Merkmalsunsicherheit bezeichnet und sollte selbst eine möglichst gute Schätzung der Schätzfehlervarianz von $\hat{x}$ darstellen.

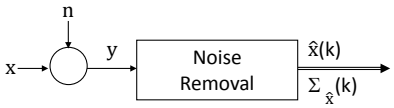

Abbildung 3: Modell der Beobachtungsunsicherheiten.

Modifizierte Likelihood-Berechnung

Zur modifizierten Likelihood-Berechnung kommen eine Reihe von Methoden in Frage, beispielsweise:

- Uncertainty Decoding (UD) [1] marginalisiert über alle möglichen Werte der sauberen Merkmale gegeben die gestörten Merkmale und deren Unsicherheit.
- Modified Imputation (MI) [2] liefert als zustandsabhängige Neuschätzung der Beobachtungsvektoren eine Maximum-Likelihood-Kombination aus Modellwissen und unsicherer Beobachtung. Diese Neuschätzung $\tilde{x}_q$ wird dann im Zustand q als Punktschätzung zur Berechnung der Likelihood verwendet.
- Significance Decoding (SD) [3] schätzt nicht nur $\tilde{x}_q$ neu, sondern bestimmt auch dessen Schätzfehlervarianz $\Sigma_{\tilde{x},q}$, und verwendet beides gemeinsam als zustandsabhängige, unsichere Beobachtung für eine Marginalisierung. So wird der Tatsache Rechnung getragen, dass auch die modellbasierte Schätzung nicht als fehlerfrei angenommen werden kann.

Tabelle 1 zeigt einige beispielhafte Ergebnisse aus [3], die mit diesen Verfahren, sowie mit einem Standard-Decoder (ML für Maximum Likelihood) in verschiedenen Störsituationen erzielt worden sind, wobei in einer Kleinvokabular-Anwendung mit 52 Worten das ETSI-Advanced Front End zur Merkmalsextraktion verwendet wurde und die Unsicherheitsschätzung laut [4] erfolgt ist.

Tabelle 1: Modifizierte Likelihood-Berechnung, aus [3].

Stör -geräusch	SNR [dB]	ML	UD	MI	SD
Babble	15	83.48	84.82	87.87	**88.30**
	10	66.31	68.37	73.33	**75.74**
	05	45.32	47.09	52.91	**56.17**
	00	25.53	27.45	35.18	**35.46**
Weißes Rauschen	15	69.65	76.52	87.73	**88.09**
	10	55.04	64.26	78.30	**78.58**
	05	42.77	48.79	65.74	**66.45**
	00	30.28	32.41	**50.35**	48.94
Büro- lärm mit Nachhall	15	59.93	62.13	67.02	**67.59**
	10	45.32	49.43	57.80	**59.22**
	5	29.86	31.49	37.30	**38.01**
	0	22.77	23.12	24.54	**24.47**

Dabei wurden bei allen Modellen diagonale Kovarianzmatrizen verwendet. Die Störgeräusche wurden künstlich zu dem - im letzten Fall verhallten - Sprachsignal addiert.

Modifizierte Merkmalsberechnung

Außerdem kann statt der Likelihood-Berechnung auch die Merkmalsberechnung angepasst werden. Eine Lösung dafür bietet die Noise-Adaptive Linear Discriminant Analysis (NALDA) [5], die in Abhängigkeit von den Merkmalsunsicherheiten für jedes Frame den diskriminativsten Unterraum des Merkmalsvektors bestimmt, um unter Störeinflüssen eine maximal zuverlässige Unterscheidung zwischen den HMM-Zuständen zu ermöglichen. Tabelle 2 zeigt für die Daten des CHiME-Corpus [6] entsprechende Ergebnisse aus [5], wobei hier die wahren quadratischen Abweichungen als Oracle-Merkmalsunsicherheiten sowohl für UD als auch für NALDA verwendet wurden, so dass die gezeigten Ergebnisse das Potential der Methoden unter optimalen Bedingungen aufzeigen.

Tabelle 2: Vergleich zwischen Maximum-Likelihood-Decoding mit diagonalen oder vollen Kovarianzmatrizen, Uncertainty Decoding und Noise-Adaptive LDA, aus [5].

SNR	ML Diag.	ML volle Kov.	UD	NALDA
-6	79.9	82.6	84.6	88.6
-3	83.5	85.7	87.0	90.8
0	86.9	88.1	89.4	92.1
3	89.6	90.5	90.9	93.6
6	91.4	92.3	92.5	94.7
9	92.9	93.8	93.5	95.3

Zusammenfassung und Ausblick

Zur robusten Spracherkennung hat es sich bewährt, aus der Signalvorverarbeitung nicht nur eine Schätzung des sauberen Sprachsignals an den Erkenner zu übermitteln, sondern auch die Zuverlässigkeit dieser Schätzung. Hierdurch lassen sich große Gewinne in den Erkennungsraten erzielen, so dass die Verwendung einer automatischen Spracherkennung auch bei niedrigen Signal-Störabständen sinnvoll möglich wird. Es besteht allerdings weiteres Optimierungspotential, sowohl in Bezug auf die Schätzung als auch auf die Verwendung dieser sogenannten Merkmalsunsicherheiten. So zeigen aktuelle Untersuchungen, dass neben der bisher üblichen Anpassung der Likelihood-Berechnung auch eine unsicherheitsgesteuerte Dimensionsreduktion zu großen Verbesserungen der Erkennungsrate führen kann, was direkt auf die Frage einer diskriminativen Merkmalsextraktion unter Beobachtungsunsicherheiten führt. Eine weitere offene Frage ist die, wie Merkmalsunsicherheiten auch dann geschätzt und verwendet werden können, wenn die Spracherkennungsmodelle ein angepasstes Training oder eine Umgebungsadaption erfahren haben. Diese beiden Zielrichtungen erscheinen vielversprechend für zukünftige Entwicklungen hin auf eine hochgradig robuste und flexible Spracherkennung.

Danksagung

Einen herzlichen Dank möchte ich allen Kollegen aussprechen, die zu den hier besprochenen Verfahren und gezeigten Ergebnissen beigetragen haben, besonders Steffen Zeiler, Ahmed Hussen Abdelaziz, Ramón Fernandez Astudillo, Reinhold Haeb-Umbach, Rahim Saeidi und Volker Leutnant.

Literatur

[1] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, May 2005.

[2] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2005, pp. 82–85.

[3] A. H. Abdelaziz, S. Zeiler, D. Kolossa, V. Leutnant, and R. Haeb-Umbach, "Gmm-based significance decoding," in *Proc. ICASSP*, 2013.

[4] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement," in *Proc. Interspeech*, 2009, pp. 2491–2494.

[5] D. Kolossa, S. Zeiler, R. Saeidi, and R. Astudillo, "Noise-adaptive LDA: A new approach for speech recognition under observation uncertainty," *Signal Processing Letters, IEEE*, vol. 20, no. 11, pp. 1018–1021, 2013.

[6] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, 2010.