Spracherkennung im Automobile durch Verwendung von Missing Feature Techniken

D. Kolossa[1], R.F. Astudillo[2], R. Orglmeister[3]

Fachgebiet für Elektronik und medizinische Signalverarbeitung, TU Berlin, 10587 Berlin, Deutschland,

Email: [1] d.kolossa@ee.tu-berlin.de,[2] ramon@astudillo.com,[3] reinhold.orglmeister@tu-berlin.de

Einleitung

Um die Robustheit von automatischen Spracherkennungssystemen gegenüber den verschiedenen Störgeräuschen zu erhöhen, wie sie im Automobil zu finden sind, ist die Verwendung von Störgeräuschunterdrückungsmethoden notwendig. Diese Methoden bieten allerdings nur eine Schätzung des sauberen Signals an, die oft verbleibende Störungen und Artefakte enthält, welche durch Ungenauigkeiten in der Schätzung des Sprachspektrums entstehen. Dies ist besonders dann kritisch, wenn die Störungen instationär sind oder die Sprache gegenüber dem Rauschen eine niedrige Amplitude besitzt. Unter Umständen kann dann das verarbeitete Signal sogar zu schlechteren Erkennungsraten führen als das gestörte Mikrofonsignal. In diesem Artikel zeigen wir, dass eine Kombination von etablierter MMSE-Störgeräuschunterdrückung mit einem effektiven Rauschleistungsschätzer und mit Missing Feature Techniken die Spracherkennungsleistung in nichtstationären Störumgebungen signifikant erhöht, besonders, wenn sie mit nichtlinearen Mel-Cepstralkoeffizienten als Sprach-Features verwendet wird.

Überblick

Die meisten der Störgeräusche in Fahrzeugen, wie Motor-, Wind- und Störsprecherinterferenzen, sind stark nichtstationär. Klassische, auf Sprachpausenerkennung beruhende Methoden sind aber nicht in der Lage, Veränderungen im Störspektrum, die während der Sprachpräsenz auftreten, korrekt zu schätzen. Der *Improved Minima-Controlled Recursive Averaging*(IMCRA)-Schätzer [5], der hier verwendet wurde, umfasst zwei Iterationen von Glättung und Verfolgung der Minima [4] und liefert eine recht gute Schätzung der Energie auch nichtstationärer Störgeräusche. Aber auch eine gute Schätzung der Störgeräusche kann keine restlose Entfernung der Störgeräusche oder der Artefakte garantieren. Um zu vermeiden, dass sich die resultierenden Schätzfehler negativ auf die Spracherkennung auswirken, wird das geschätzte Signal $\hat{X}$ hier durch eine komplexe Gauß-Verteilung $\tilde{X}$ mit dem Durchschnittswert $\hat{X}$ und der Varianz $\lambda/2$ für jede Dimension ersetzt. Zur Spracherkennung wird dann anstelle von $\hat{X}$ diese komplexe Gaußverteilung $\tilde{X}$ durch die Feature-Extraktion an einen Missing-Feature Erkenner weitergeleitet, so dass die Unsicherheit im Erkennungsprozess berücksichtigt werden kann, was den Einfluss von Artefakten und falschen Schätzungen auf das Erkennungsergebnis reduziert [6].

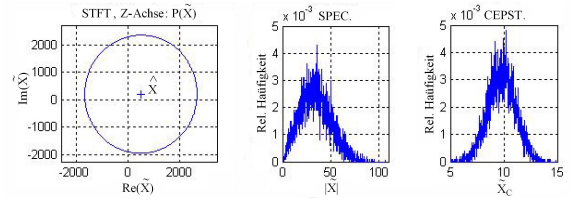

Abbildung 1: *Transformation von Unsicherheitsinformationen für Missing-Feature-Spracherkennung.*

Stufenweise Unsicherheitstransformation

Feature-Extraktionsmethoden für die Spracherkennung sind in der Regel nichtlinear, daher sind komplexe Berechnungen nötig, um ihren Einfluss auf die ersten und zweiten Momente der angenommenen Verteilungsfunktionen zu bestimmen. Außerdem sind diese Berechnungen spezifisch für jeden Feature-Extraktions-Typ. Die Nutzung von Pseudo-Montecarlo-Methoden, wie der Unscented Transformation (UT) [3, 6] bietet eine generelle Lösung für das Problem, ist aber besonders bei hochdimensionalen Zufallsvariablen, wie im anfangs angenommenen STFT-Bereich, langsamer und ungenauer als eine analytische Lösung. Die Alternative, die bessere Ergebnisse zeigte, ist eine Kombination aus sowohl analytischen als auch Unscented Transformationen in einer stufenweisen Methode. Für die hier verwendete Mel-Cepstral Feature Extraktion wurden die folgenden analytischen Berechnungen bzw. UT-Approximationen verwendet: Wenn die Unsicherheit des k-ten Frequenzbandes der STFT mit einer komplexen Gauß-Verteilung des Mittelwertes $\hat{X}_k$ und der Varianz der realen Dimension $\lambda_k/2$ modelliert wird, so folgt der Betrag dieses Frequenzbandes einer Rice-Verteilung mit den Parametern $|\hat{X}_k|$ und $\sqrt{\lambda_k/2}$. Die Varianz nach der Betrags-Transformation, $\Sigma_{|\tilde{X}|}$, kann dann mit der folgenden Formel berechnet werden:

$$\Sigma_{|\tilde{X}|}(k,k) = \lambda_k + |\hat{X}_k|^2 - \frac{\pi\lambda_k}{4} \cdot L_2^1\left(-\frac{|\hat{X}_k|^2}{\lambda_k}\right)^2 . \quad (1)$$

Hierbei ist $L_2^1(x)$ das Laguerre-Polynom, welches in Termen der modifizierten Bessel-Funktionen ausgedrückt werden kann. Die Mel-Filterbank und die diskrete Kosinus-Transformation (DCT) sind lineare Transformationen und die Kovarianz Σ' am Ausgang dieser Blöcke kann leicht berechnet werden aus der Kovarianz-Matrix Σ an dem jeweiligen Eingang mittels $\Sigma' = A\Sigma A^T$. Hier ist A entweder die Mel-Frequenz-Matrix oder die DCT-Matrix. Die logarithmische Transformation, die nach der Mel-Filterbank erfolgt, wird durch die UT recht genau abgeschätzt, da diese Zufallsvariable eine günstigere Größe hat als im Fall der STFT. Ein weiterer Vorteil der Nut-

zung der UT an dieser Stelle ist die leichtere Berücksichtigung der nicht-diagonalen Kovarianz-Matrix, die aus der Mel-Filterbank resultiert, so dass die Unsicherheit im Cepstralbereich auf diese Weise genauer berechnet werden kann.

Unsicherheits-Schätzung und Missing-Feature-Erkennung

Die Unsicherheiten nach der MMSE-Störgeräuschunterdrückungsfilterung wurden aus internen MMSE-Schätzparametern und der Störsignalamplitude abgeschätzt. Sobald diese Information in den Feature-Bereich transformiert ist, können verschiedene Techniken zum Einsatz kommen, um gestörte Features robust zu erkennen. Missing-Feature-Techniken wie die Marginalisierung, die Imputation [7], die modifizierte Imputation [6] oder das Uncertainty Decoding [1].

Experimente

Die Testdaten beinhalten 200 Dateien von 10 verschiedenen Sprechern der TIDIGITS-Datenbank. Zwei nichtstationäre Störgeräusche, Fahrtwind bei offenem Fenster und in einer Fußgängerzone aufgenommene Gesprächsfetzen, wurden mit definierten Segment-SNRs zu den Sprachsignalen addiert. Der MMSE-LSA-Algorithmus [2] wurde in Kombination mit dem IMCRA-Schätzer [5] zur Störgeräuschunterdrückung verwendet. Die Spracherkennungsmodelle waren Phonem-HMMs mit 6 Gauß'schen Mischungskomponenten, die mit der gesamten TI-DIGITS-Datenbank auf ungestörten Daten trainiert wurden. Um die Exaktheit der stufenweisen Unsicherheitstransformation zu verifizieren, wurden die errechneten Varianzen im Cepstral-Bereich verglichen mit denen, die mit einer Montecarlo-Simulation mit 1000 Punkten errechnet wurden. Schließlich wurden Erkennungsversuche durchgeführt, die die Unsicherheiten im Frequenzbereich einerseits mit (*Ideal*) und einerseits ohne (*Approx.*) Kenntnis des realen Spektrums errechneten. Die Ergebnisse wurden verglichen mit den besten Ergebnissen die im Spektralbereich mit Ephraim-Malah-Filterung und Missing Feature Techniken (*Ideal*) erzielt werden konnten.

Spracherkennungs-Ergebnisse

Tabelle 1 zeigt die relevantesten Ergebnisse. Die korrekt identifizierten Wörter in Prozent (Word Correctness "WC") und die korrekt identifizierten Wörter minus der Anzahl der Einfügungen (Word Accuracy "WA") wurden als Gütemaße genutzt. Die modifizierte Imputation (MI) [6], getestet mit idealen und approximierten Unsicherheits-Schätzungen, zeigte die besten Ergebnisse aller verglichenen Missing-Feature-Techniken im Cepstralbereich.

Schlussfolgerungen

Die stufenweise Transformation liefert einen schnellen und effizienten Weg, um unsichere Sprachfeatures aus dem Frequenzbereich zum Zweck der Missing-Feature-Erkennung in den Bereich der Spracherkennungs-Features zu transformieren. So können statistische Informationen, berechnet im STFT-Bereich, zum Beispiel im Mel-Cepstral-Bereich verwendet werden. Die hier vorgeschlagene Lösung läßt sich leicht auch für andere nichtlineare Front-Ends einsetzen. Die Nutzung der Missing-Feature-Methoden im Cepstral-Bereich kann die Effizienz der Erkennung um bis zu 20% erhöhen. Die Tests zeigen aber auch noch die Notwendigkeit einer robusten Schätzung der Unsicherheit nach der MMSE-Störgeräuscheunterdrückung, da bisherige Methoden abhängig von der Art der Störgeräusche sind.

	WC[%]	WA[%]	WC[%]	WA[%]
Referenz	98.76	98.76	98.76	98.76
Windgeräusche	-15dB SNR		5dB SNR	
Noisy	59.66	28.44	94.74	87.94
MMSE-LSA	69.86	34.78	96.45	75.27
MI (Approx.)	72.64	46.68	91.65	88.72
MI (Ideal)	79.60	51.93	97.53	94.28
Optimal Spekt.	59.81	44.98	84.08	44.05
Störsprecher	-15dB SNR		5dB SNR	
Noisy	52.40	22.87	95.83	92.43
MMSE-LSA	61.51	36.63	97.06	92.43
MI (Approx.)	69.71	22.72	96.14	94.90
MI (Ideal)	80.99	48.53	97.06	96.45
Optimal Spekt.	52.55	19.23	88.24	58.89

Tabelle 1: Relevante Testergebnisse

Literatur

[1] Droppo J., Acero A. and Deng L. "Uncertainty Decoding with Splice for noise robust speech recognition", *Proc. ICASSP 2002*, May 2002.

[2] Ephraim Y. and Malah D. "Speech enhancement using a minimum-mean square error short-time Log-spectral amplitude estimator", *IEEE Trans. ASSP*, **11**(5), pp. 443 - 445, April 1985.

[3] Julier S.J. and Uhlmann J.K. "A General Method for Approximating Nonlinear Transformations of Probability Distributions", Technical Report, Dept. Eng. Science, University of Oxford, UK, 1996.

[4] Martin R., "Spectral subtraction based on minimum statistics", Proc. Eur. Signal Processing Conf., 1994.

[5] Cohen I. "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", *IEEE Trans. Speech and Audio Processing*, **11**(5), Sept. 2003

[6] Kolossa D., Klimas A. and Orglmeister R. "Separation and Robust Recognition of Noisy, Convolutive Speech Mixtures using Time-Frequency Masking and Missing Data Techniques", *Proc. WASPAA 2005*, October 2005.

[7] Raj B., Seltzer M. and Stern R. "Reconstruction of Missing Features for Robust Speech Recognition", *Speech Communication,* **43**, S. 275-296, 2004.