Robuste Erkennung gestörter Sprache im Automobil durch MMSE-Störgeräuschunterdrückung und Missing-Data-Spracherkennung

Dorothea Kolossa[1], Aleksander Klimas[2], Wolf Baumann[3], Reinhold Orglmeister[4]
Institut für Elektronik und medizinische Signalverarbeitung, TU Berlin, 10587 Berlin, Deutschland,
Email: [1] d.kolossa@ee.tu-berlin.de,[2] clif.cs.tu-berlin.de,[3] w.baumann@ee.tu-berlin.de,[4] reinhold.orglmeister@tu-berlin.de

Einleitung

Um zufriedenstellende Spracherkennungsergebnisse auch im Automobilbereich zu erhalten, ist eine adequate Störgeräuschunterdrückung von großer Bedeutung. Damit sind substanzielle SNR-Verbesserungen möglich, allerdings leidet die Güte der Spracherkennug oft unter dieser Verarbeitung. Hier wird eine Möglichkeit vorgeschlagen, die es erstmals erlaubt, statistische Signalverarbeitung im Frequenzbereich flexibel mit Missing-Data-Erkennung in einem für Spracherkennung günstigen Feature-Bereich zu verbinden. Und während die hier untersuchte Störgeräuschunterdrückung allein die Spracherkennungsrate um knapp 20% verbessern kann, bietet die vorgeschlagene Anbindung der Vorverarbeitung an die Missing-Feature-Erkennung einen zusätzlichen Erkennungsratengewinn von bis zu 36%, was die probabilistische Anbindung zwischen Vorverarbeitung und Spracherkennung für robuste Spracherkennung prädestiniert.

Überblick

In konventionellen Spracherkennungssystemen wird die Sprachsignalverarbeitung weitgehend unabhängig vom Erkennungsprozess behandelt. Auch wenn das Sprachsignal mit statistischen Methoden behandelt wird, die eine Schätzung nicht nur des Signals sondern auch von dessen Varianz erlauben, wird letztlich nur das geschätzte Signal selbst an den Erkenner weitergegeben. Stattdessen könnte man auch die Varianzinformationen im Erkennungsprozess nutzen, um daraus auf die Zuverlässigkeit einzelner Features zu schließen. Diese Vorgehensweise liegt der Missing-Data-Erkennung [1] und dem Uncertainty Decoding [2] zugrunde, die in den letzten Jahren vermehrte Aufmerksamkeit erfahren haben. Problematisch ist dabei, dass die Signalverarbeitung meist im Spektralbereich, die Erkennung dagegen meist im Mel-Cepstralbereich durchgeführt wird. Wenn nun, wie es zur Verbindung der beiden Teile nötig ist, die Erkennung in den Bereich der Vorverarbeitung verlegt wird, führt das zu suboptimalen Ergebnissen, die z.B. Raj et al. in [7] dazu veranlassen, wegen dieses Nachteils auf konfidenzbasierte Erkennung ganz zu verzichten. Hier schlagen wir eine Alternative vor, die darin besteht, die Sprach-Features gemeinsam mit ihren Varianzen aus dem Arbeitsbereich der Vorverarbeitung in den der Erkennung zu transformieren, wie es in Abb. 1 gezeigt ist. Für die nötige Transformation der Verteilungsdichten erweist sich die Unscented-Transformation als besonders flexibel. In den folgenden Abschnitten werden die Uncertain Feature Transformation und die darauf aufbauende Erkennung kurz beschrieben, gefolgt von den erzielten Ergebnissen und Schlußfolgerungen.

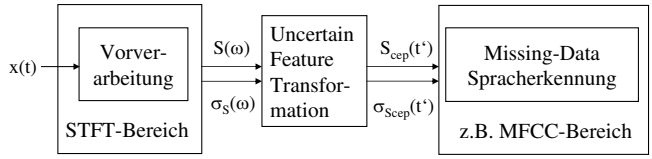

Abbildung 1: *Integration von Varianzinformationen in die Spracherkennung.*

Transformation der Varianzen

Um aus den Ergebnissen der Vorverarbeitung die Features und Varianzen im Cepstralbereich zu bestimmen, müssen die Verteilungsdichten am Ausgang der Vorverarbeitung in den Cepstralbereich transformiert werden. Für die linearen Blöcke der Feature-Extraction ist die Lösung gegeben durch

$$\mu_{f2} = \mathbf{T}\mu_{f1}, \quad \Sigma_{f2} = \mathbf{T}\Sigma_{f1}\mathbf{T}^T, \tag{1}$$

mit $\mathbf{T}$ als Transformationsmatrix sowie $\mu_{f1/f2}$ und $\Sigma_{f1/f2}$ als Mittelwert und Varianz vor bzw. nach der Transformation. Für die nichtlinearen Schritte der Transformation wäre jedoch zur analytischen Berechnung die Integration

$$\mu_{f2} = \int_{-\infty}^{\infty} T(f_1) p_{f1}(f_1) df_1 \quad \text{und} \tag{2}$$

$$\sigma_{f2}^2 = \int_{-\infty}^{\infty} (T(f_1) - \mu_{f2})^2 p_{f1}(f_1) df_1 \tag{3}$$

nötig. Im Fall von MFCC-Features existiert bereits eine analytische Lösung der auftretenden Integrale in [4]. Um aber eine allgemeingültige Methode zur Transformation zu erhalten, die unabhängig von den Features des Spracherkenners arbeitet, verwenden wir hier die Unscented Transformation [5]. Dadurch ist es nicht mehr nötig, für jede neue Spracherkennerparametrisierung analytische Integrationen durchzuführen. Stattdessen bietet die Unscented Transformation eine Lösung, die mit vertretbarem numerischen Aufwand die Momente der Ausgangsverteilungen bis zu der gewählten Ordnung approximiert.

Unscented Transformation

Wenn eine Zufallsvariable nichtlinear transformiert wird, stehen die ersten Momente am Ausgang über (2) und

(3) mit der Verteilungsdichte am Eingang in Beziehung. Diese Integration kann approximiert werden, indem ein Satz von sogenannten *Sigma-Punkten* generiert wird, der die Signalstatistiken am Eingang der Transformation bis zu einer gewählten Ordnung approximiert. Werden diese Punkte durch die nichtlineare Transformation propagiert, erhält man am Ausgang Punkte, die die Statistiken der Ausgangsverteilung bis zu der gewählten Ordnung exakt nachbilden [5].

Erkennung von varianzbehafteten Features

Um Varianzen in den Erkennungsprozess einzubeziehen, existieren einige Möglichkeiten, die teils als Missing-Data-Techniken (insbesondere Marginalisierung und Imputation), teils als Uncertainty Decoding bezeichnet werden, siehe z.B. [7, 2, 1]. Hier verwenden wir eine modifizierte Variante der Data-Imputation, die Featurevarianzen kontinuierlich auswertet [6]. Im Gegensatz dazu werden bei der konventionellen Imputation die Features nur als zuverlässig oder unzuverlässig eingestuft [1].

Experimente

Sprachdaten von der TIDigits-Datenbank für verbunden gesprochene Ziffern wurden in einem Fahrzeug im Stillstand und während der Fahrt aufgenommen. Das SNR während der Fahrt betrug -9,6 dB.

Signalvorverarbeitung

Für die Vorverarbeitung wird hier das Ephraim-Malah-Filter [3] verwendet. Dadurch erhält man eine Minimum Mean Square Error Schätzung des Sprachsignals $S(\omega, t)$ und der Rauschvarianz, die hier gleichzeitig als Varianz der Sprachsignalschätzung $\sigma_S(\omega, t)$ betrachtet wird.

Spracherkenner

Als Basis für den Spracherkenner wurde ein sprecherunabhängiges Modell der TIDigits trainiert und auf den Fahrzeugdaten bei Stillstand adaptiert. Als Features wurden die ersten 13 MFCCs und deren erste und zweite Ableitungen verwendet. Die Modelle waren Continuous-Density-HMMs mit sechs Gauß'schen Mischungen.

Ergebnisse

Die Erkennung der Fahrzeugdaten führte zu den in Tabelle 1 gezeigten Ergebnissen. Wie zu erkennen ist, wird die Erkennungsrate durch das Ephraim-Malah-Filter allein um 18% und durch zusätzliche Verwendung der transformierten Varianzen um weitere 36% gesteigert. Dabei ist erwartungsgemäß nicht relevant, ob die Varianzen durch Unscented Transformation oder analytisch berechnet werden.

Schlussfolgerungen

Es wurde eine Methode vorgestellt, die Signalvorverarbeitung und Spracherkennung stärker als bisher miteinander zu koppeln. Dazu werden Varianzinformationen mit Hilfe der Unscented Transformation in den Bereich der Spracherkenner-Features überführt. Durch dieses Verfahren werden deutliche Erkennungsratenverbesserungen erzielt, die denen einer analytischen Varianzberechnung entsprechen. Zugleich ist diese Implementierung flexibel bezüglich der Spracherkennungs-Parametrisierung. Dadurch können in zukünftigen Untersuchungen auch auditorisch motivierte Features eingesetzt werden, die für problematische Umgebungen wegen ihrer Robustheit besonders geeignet erscheinen.

	%Correct	%Accurcy
Referenzaufnahme	99.6	97.9
Fahrt (-9,5dB SNR)	35.2	27.0
Fahrt nach Ephraim-Malah-Filterung	41.6	31.4
Fahrt nach Filterung mit Konfidenzmaß (UT)	56.7	53.5
Fahrt nach Filterung mit Konfidenzmaß (A)	56.4	54.5

Tabelle 1: Erkennungsergebnisse der Fahrzeugdaten (A = Analytische, UT = Unscented Transformation)

Literatur

[1] Cooke M., Green P., Josifovski L. and Vizinho A. "Robust automatic speech recognition with missing and unreliable acoustic data", *Speech Communication*, pp. 267-285, 2001.

[2] Droppo J., Acero A., Deng L. "Uncertainty Decoding with Splice for noise robust speech recognition", *Proc. ICASSP 2002*, May2002.

[3] Ephraim Y. and Malah D. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Trans. Acoustics, Speech, and Signal Processing*, **32**, pp. 1109 - 1121, 1984.

[4] Gales M. J. F. "Model-Based Techniques for Noise Robust Speech Recognition", Ph.D. thesis, University of Cambridge, September 1995.

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

[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.