From H. Fastl et al, *31. Deutsche Jahrestagung für Akustik, DAGA '05*, pp. 315–316,
© Deutsche Gesellschaft für Akustik e.V., ISBN 3-9808659-1-6

Helfen „Fallen“ bei verrauschten Daten? - Spracherkennung mit TRAPs

Andreas Maier[1], Christian Hacker[1], Stefan Steidl[1], Elmar Nöth[1]
[1] *Lehrstuhl für Mustererkennung, 91058 Erlangen, Deutschland, Email: noeth@informatik.uni-erlangen.de*

Einleitung

In diesem Beitrag wird die Leistung von **T**empo**r**a**l** **P**atterns (TRAPs) auf klaren und verrauschten Daten untersucht. Ziel ist es zu zeigen, dass TRAP-Merkmale eine höhere Unabhängigkeit bezüglich der akustischen Gegebenheiten haben. Zu diesem Zweck wird ein Korpus in drei verschiedenen Hall- und Rausch-Stufen herangezogen: das AIBO Korpus [1]. Die Daten sind zum einen mit einem Nahbesprechungsmikrofon und zusätzlich mit einer Videokamera aufgezeichnet worden, d.h. es liegt eine zweite Version mit starkem Hall und Hintergrundgeräuschen vor. Durch Faltung der Nahbesprechungsdaten mit verschiedenen Impulsantworten konnte eine dritte Version des Korpus mit künstlichem Hall erstellt werden. Als Merkmale werden Mel-Cepstrum-Koeffizienten (MFCC), Temporal Patterns (TRAPs [2]) und im Modulationsspektrum gefilterte TRAPs herangezogen. Die TRAPs-Merkmale werden aus unterschiedlichen Bändern des Spektrums unter Berücksichtigung eines langen zeitlichen Kontexts berechnet. Mit jedem Merkmalsatz wird dann jeweils ein Erkenner trainiert. Bei der Evaluierung werden mehrere Erkenner kombiniert, um eine Verbesserung der Erkennungsrate zu erreichen. Dies geschieht durch Kombination der erkannten Ketten mit ROVER. Durch Training auf klaren bzw. künstlich verhallten Daten und Evaluierung auf stark verrauschten Daten kann gezeigt werden, dass TRAP-basierte Merkmale eine höhere Generalisierungsfähigkeit haben. Beim Training mit den klaren Daten kann ein relativer Zuwachs der Wortakkuratheit von 173 % (von 12.0 % auf 32.8 %) erreicht werden. Zunächst werden in einem Literaturüberblick TRAP-Merkmale motiviert. Es folgt eine Beschreibung der drei Versionen des Korpus. Danach werden wir auf den genauen Versuchsaufbau eingehen und die verwendeten Merkmale vorstellen. Im Anschluss werden die erzielten Ergebnisse präsentiert.

Literaturüberblick

Die Merkmale hier beruhen auf dem TRAP Ansatz von Hermansky und Sharma [2]. **T**empo**r**a**l** **P**atterns sind Muster mit langem zeitlichen Kontext von bis zu einer Sekunde in einem begrenzten Spektralband. Für jedes Frequenzband werden Neuronale Netze trainiert, um die Daten zu reduzieren; man erhält pro Frame und Frequenzband unter Berückdichtigung des Kontextes Wahrscheinlichkeiten für 29 Lautklassen. Die hochdimensionalen Merkmalvektoren aller Bänder (29 Merkmale pro Band) werden dann durch ein weiteres Neuronales Netz auf einen Ausgabevektor mit 29 Einträgen abgebildet. Mit diesen Merkmalen können bei Störungen, die nur in bestimmten Bändern auftreten, bessere Erkennungsraten erzielt werden als mit MFCC-Merkmalen. Das Modulationsspektrum wird durch eine Kurzzeit-Analyse in einem Frequenzband gewonnen. In [3] wird eine Filterung der Spektralbänder im Modulationsspektrum untersucht. Laut Hermansky sind nur die Änderungen zwischen 1 und 16 Hz in diesem Spektrum für die Erkennung wichtig. In [4] wird ROVER vorgestellt, ein Algorithmus, der Ausgaben von mehreren Spracherkennern verbindet. Die Ausgaben werden mit einer abgewandelten Version der dynamischen Programmierung auf Wortebene iterativ einander zugeordnet. Danach wird per Mehrheitsentscheidung eine neue Ausgabe erstellt um Erkennungsraten zu verbessern.

Korpora

Wie schon in der Einleitung erwähnt, wird für die Experimente das Aibo-Korpus mit spontaner Kindersprache herangezogen. Die Versuchspersonen sollten den Aibo-Roboter der Firma Sony in natürlicher Sprache steuern und verschiedene Aufgaben erfüllen, z.B. durch einen Parcour laufen lassen. In Wirklichkeit wurde der Roboterhund von einem “Wizard-of-Oz” ferngesteuert um emotionale Sprache zu evozieren. Eine genaue Beschreibung ist in [1] zu finden. Die 8,5 h des Korpus liegen in drei Versionen vor: Zum einen wurde mit einem Nahbesprechungsmikrophon aufgezeichnet; diese Version wird im Folgenden als *Aibo nb* bezeichnet. Zu Dokumentationszwecken wurde ferner mit einer Videokamera (mit Mikrofon mit Kugelcharakteristik) die ganze Szene aufgenommen (*AIBO rm*). Diese Aufnahme enthält starke Rausch- und Hall-Effekte. Auch sind Geräusche vom Aibo, der zum Teil näher an der Kamera ist als die Versuchsperson, deutlich zu hören. Für *Aibo nbvh* wurden die *nb*-Daten künstlich verhallt. Der Hall wurde durch Faltung mit verschiedenen Impulsantworten erzeugt. Man beachte, dass die Impulsantworten aus keinem der beiden Klassenzimmer, in denen die Aibo-Aufnahmen stattfanden, stammen. Alle Daten wurden mit 16 kHz abgetastet und mit 16 bit quantisiert. Die Größe des Vokabulars beträgt 850 Wörter und 350 Wortabbrüche.

Versuchsaufbau

Im Folgenden werden die verwendeten Merkmale, TRAPs und gefilterte TRAPs, beschrieben. Danach werden Details zum Erkenner und deren Kombination gegeben.

Merkmale

Bei der Berechnung von TRAPs orientieren wir uns am Ansatz von Hermansky [2]; in einigen Punkten unter-

scheidet sich jedoch die Berechnung der von uns eingesetzten Merkmale. Die Baseline-Systeme basieren auf 11 Mel-Cepstrum-Merkmalen, der Gesamtenergie und 12 ersten Ableitungen. Die TRAP-Merkmale werden aus dem Spektrogramm erzeugt, indem für jedes der 18 logarithmierten Mel-Bänder ein zeitlicher Kontext von 2×150 ms betrachtet wird. Bei einer Fortschaltzeit von 10 ms entspricht das 31 Koeffizienten. Für jeden Erkenner wird nur ein Teil der Frequenzbänder verwendet, z. B. jedes dritte. Die Koeffizienten dieser Bänder werden konkateniert und dann mit der linearen Diskriminanzanalyse (LDA) auf 24 Dimensionen reduziert. Die für die LDA nötigen Klassenzugehörigkeiten wurden durch eine erzwungene Zuordnung der 46 Laute mit dem Baseline-System ermittelt. In einer zweiten Version werden die TRAPs geglättet, d. h. im Modulationsspektrum alle Frequenzanteile kleiner 1 Hz und größer 16 Hz entfernt. Die geglätteten TRAPs aus ausgewählten Bändern werden ebenfalls aneinander gehängt und ihre Dimension auf 24 reduziert.

Erkenner

Die Hidden-Markov-Modelle (HMM), die hier eingesetzt werden, werden mit der ISADORA-Umgebung des Lehrstuhls für Mustererkennung (LME) der Universität Erlangen-Nürnberg trainiert. Dieses System wird dort schon seit 1978 entwickelt und wurde schon auf vielen Gebieten auch jenseits der Sprachverarbeitung eingesetzt. Um verschiedene Erkenner zu verschmelzen, wird ROVER eingesetzt. Dabei werden die erkannten Wortfolgen und deren Konfidenz als Eingabe benutzt und verarbeitet. Leider unterstützt ROVER nur lineare Eingaben, also keine Wortgraphen und auch keine Sprachmodelle.

Ergebnisse

Nun werden die vorgestellten Merkmale evaluiert. Im Folgenden stehen die Buchstaben T für TRAPs und F für gefilterte TRAPs. Zunächst werden Erkenner trainiert, die jeweils nur ein Drittel des Spektrums verwenden. Für T_1 etwa werden TRAPS aus den Frequenzbänder $1, 4, 7, \ldots$ berechnet, konkateniert und mit LDA reduziert. T_i bzw. F_i verwenden die Bänder $3n + i$ $(n = 0, 1, \ldots, 5)$. Die Ergebnisse für die Baseline-Erkenner mit MFCC-Merkmalen, den einzeln besten Erkennern mit gefilterten TRAPs (F_{best}) und für verschmolzene Erkenner können in Tab. 1 eingesehen werden. Die drei Versionen des Korpus werden erst unter gleichen Trainings- und Testbedingungen evaluiert. Anschließend wird untersucht, wie gut die verhallten Raummikro-Daten (*rm*) erkannt werden, wenn für das Training nur Daten vom Nahbesprechungsmikrofon (*nb*) bzw. künstlich verhallte Daten (*nbvh*) zur Verfügung stehen. Zum einen ist aus der Tabelle ersichtlich, dass ein Verschmelzen der TRAP-basierten Erkenner immer eine Verbesserung bringt. Meist steigert eine Kombination mit dem Baseline-System (MFCC) zusätzlich die Erkennungsrate. Beachtlich ist aber auf jeden Fall, dass die TRAPs-Erkenner bei ungleichen Trainings- und Testbedingungen deutlich besser abschneiden, als der entsprechende Baseline-Erkenner mit MFCC-Merkmalen.

Training	*nb*	*rm*	*nbvh*	*nb*	*nbvh*
Test	*nb*	*rm*	*nbvh*	*rm*	*rm*
MFCC	**77,2**	46,9	63,1	12.0	18,8
F_{best}	66,1	39,0	61,1	15,8	23,0
T_1,T_2,T_3	71,1	47,3	**65,9**	21,6	**34,3**
T_1 - T_3,MFCC	70,3	**48,7**	**66,7**	**32,8**	31,6
F_1,F_2,F_3	72,1	**48,9**	63,4	27,2	**34,3**
F_1 - F_3,MFCC	**73,0**	**48,7**	63,9	**31,8**	**34,9**

Tabelle 1: Baseline Erkenner (MFCC), einzeln optimale Erkenner und verschmolzene Erkenner mit 4-Gramm Sprachmodell; Wortakkuratheit in %;

Die Ergebnisse zeigen Vorteile der TRAP-Merkmale unter bestimmten akustischen Bedingungen. Zwar haben sie bei Training und Evaluierung auf der *nb* Version des Korpus – auch beim Kombinieren der Erkenner – stets eine schlechtere Leistung erbracht, bringen aber bei den anderen Versionen des Korpus Verbesserungen. Dieses Ergebnis ist ein Indiz, dass TRAP-basierte Merkmale besser generalisieren können; sie schneiden bei verschiedenen Trainings- und Test-Daten besser ab. Die Steigerung der Wortakkuratheit von bis zu 173 % (von 12.0 % auf 32.8 %) gegenüber dem Baseline-System zeigt dies.

Zusammenfassung

In dieser Arbeit wurde das AIBO Korpus in drei Versionen vorgestellt: mit Nahbesprechungsmikrophon, Raummikro und künstlich verhallt. An diesem Korpus wurden verschiedene Merkmale evaluiert: TRAPs und gefilterte TRAPs. Zum Vergleich wurden MFCC herangezogen. Die Ergebnisse verschiedener Erkenner, die jeweils nur Teile des Spektrums berücksichtigen, wurden mit ROVER kombiniert. Es konnte gezeigt werden, dass die TRAP-basierten Merkmale besser generalisieren können, da die Wortakkuratheit um bis zu 173 % gesteigert werden konnte, wenn für Training und Test verschiedene Versionen des Korpus verwendet werden.

Literatur

[1] A. Batliner, C. Hacker, S. Steidl, and E. Nöth. "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proc. of the 4th Int. Conf. of Language Resources and Evaluation '04*, pages 171–174, Lisbon, Portugal, 2004.

[2] H. Hermansky and S. Sharma. TRAPs - classifiers of temporal patterns. In *Proc. ICSLP '98*, volume 3, pages 1003–1006, Sydney, Australia, 1998.

[3] H. Hermansky. The modulation spectrum in automatic recognition of speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, USA, 1997.

[4] J. Fiscus. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction. In *Proc. IEEE ASRU Workshop*, pages 347–352, Santa Barbara, USA, 1997.