Phonemverwechslungen bei menschlicher und automatischer Spracherkennung

Bernd T. Meyer, Thomas Brand, Birger Kollmeier
Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, E-Mail: bernd.meyer@uni-oldenburg.de

Einleitung

Das Ziel dieser Studie war es, die Leistung von menschlicher Spracherkennung (Human Speech Recognition / HSR) und automatischer Spracherkennung (Automatic Speech Recognition / ASR) zu vergleichen. Die Leistung von ASR-Systemen unter schwierigen Bedingungen ist häufig so niedrig, dass ein sinnvoller Einsatz dieser Technologie nicht möglich ist. Andererseits ist dies Anreiz, die Prinzipien, die für HSR wichtig sind, zu analysieren und teilweise auf ASR zu übertragen, da der Mensch als Spracherkenner unübertroffen ist. Dazu sollen Fehlermuster in HSR und ASR miteinander verglichen werden, um Fehlerquellen zu identifizieren und ASR zu verbessern.
Um ähnliche experimentelle Bedingungen zu erreichen wurde dieselbe Sprachdatenbank verwendet, so dass Versuchspersonen nicht in der Lage waren, Kontextwissen auszunutzen. Der Effekt von Sprachmodellen konnte so ausgeblendet werden. Dies entkoppelt zwei der größten Fehlerquellen in ASR, nämlich das Frontend, das zur Merkmalsextraktion verwendet wird, und den Klassifikator. Für einen fairen Vergleich sollen Mensch und Maschine die gleiche Information zur Klassifikation erhalten. Dazu werden häufig eingesetzte ASR-Merkmale in hörbare Signale umgewandelt [2] und menschlichen Versuchspersonen dargeboten; der Schwerpunkt der Studie liegt daher bei der Merkmalsextraktion. Verschiedene Fehlermuster für die Phonemverwechslungen, die mit Hilfe von Verwechslungsmatrizen untersucht wurden, könnten Fehlerquellen aufzeigen und helfen, die Merkmalsextraktion in ASR zu verbessern.

Sprachdatenbank OLLO

HSR und ASR Experimente wurden mit dem Oldenburg Logatome Corpus (OLLO) durchgeführt [1]. Die Datenbank enthält 150 verschiedene nonsense-Äußerungen, die von insgesamt 50 Sprechern aufgenommen wurden. Diese Logatome bestehen aus einer Kombination von Vokal-Konsonant-Vokal (VCV) oder Konsonant-Vokal-Konsonant (CVC), wobei die äußeren Phoneme identisch sind. Um den Einfluss von sprachintrinsischen Variabilitäten untersuchen zu können, wurde jedes Logatom in den Varianten 'laut', 'leise', 'langsam', 'schnell' sowie 'fragend' und 'normal' aufgenommen. Zusätzlich zu zehn hochdeutschen Sprechern enthält OLLO Aufnahmen von je zehn Sprechern aus Ostfriesland, Bayern, Westfalen und dem französischsprachigen Teil von Belgien. Für ein verbessertes Training von akustischen Modellen bei ASR wurde jedes Logatom dreimal aufgenommen. Insgesamt enthält OLLO über 130.000 Logatome und Sätze; die Datenbank kann kostenlos unter der Adresse http://www.sirius.physik.uni-oldenburg.de heruntergeladen werden.

Testkonditionen

Die Sprachverständlichkeitstests mit Menschen umfassten zwei Konditionen:
Darbietung von resynthetisierten Signalen: Um einen fairen Vergleich durchzuführen, wurde untersucht, ob die für ASR üblicherweise eingesetzten Merkmale (Mel-Frequency Cepstral Coefficients / MFCCs) sämtliche Information enthalten, damit Menschen Sprache auf Phonemebene verstehen können. MFCCs stellen zwar eine kompakte Repräsentation von Sprachsignalen dar, bei der Berechnung wird jedoch Information über Phase und die Feinstruktur von Kurzzeitspektren verworfen sowie die spektrale Auflösung verringert. Im Störgeräusch könnte dies die Leistung von ASR Systemen vemindern, weil redundante Information, die von Menschen ausgenutzt wird, aus dem Signal entfernt wird. Um die Merkmale in Sprache umzuwandeln, wurde zunächst ein lineares neuronales Netz verwendet, um aus 13 cepstralen Koeffizienten die spektrale Einhüllende zu schätzen. Für eine qualitativ hochwertige Resynthese müssten Grundfrequenz sowie Information über die Anregung (stimmhaft oder stimmlos) verwendet werden. Diese Information ist jedoch in den ASR Merkmalen nicht enthalten und würde menschlichen Versuchspersonen darum einen unfairen Vorteil verschaffen. Daher wurde ein künstliches Anregungssignal mit fester Grundfrequenz verwendet. Durch Filterung wurde aus der spektralen Einhüllenden und dem Anregungssignal das resynthetisierte Signal berechnet. Da die Berechnung von MFCCs mit einem Informationsverlust behaftet ist, klingen diese Signale verzerrt und unnatürlich; trotzdem lagen bei Sprachmessungen in Ruhe HSR-Erkennungsraten sehr dicht bei 100%. Um statistisch signifikante Ergebnisse zu erhalten, wurde darum ein Rauschen mit sprachähnlichem Langzeitspektrum addiert (SNR: 0 dB).
2. Darbietung von Originalsignalen: Als Referenzkondition wurden unverarbeitete Signale für HSR Tests verwendet. Ein Vergleich mit Kondition 1 sollte zeigen, ob sich die Fehlermuster unterscheiden und ob Information, die zum Spracherkennen wichtig ist, bei der Berechnung von MFCCs verlorengeht. Vor der Präsentation von Signalen wurde ein Rauschen mit einem SNR von -10 dB addiert, was zu ähnlichen Gesamterkennungsraten wie bei Kondition 1 führt.

Menschliche Spracherkennung

Sechs normalhöhrende Versuchspersonen (VPs) ohne Dialekt nahmen an den Messungen teil. Die Signale wurden in einer schallisolierten Kabine über audiologische Kopfhörer (Sennheiser HDA200) dargeboten. Eine Online-Entzerrung und Randomisierung der Logatome wurde von der Messsoftware MessOL vorgenommen. Um

Fehler durch Unaufmerksamkeit zu vermeiden, wurden VPs angehalten, regelmäßig Pausen einzulegen. Den VPs wurde nach einer Trainingsphase eine Folge von Logatomen bei einem Pegel von 70 dB SPL dargeboten. Nach jeder Darbietung sollte das Logatom aus einer Liste von CVCs oder VCVs mit demselben äußeren Phonem und verschiedenen Mittelphonemen ausgewählt werden. In jeder der beiden Konditionen wurden den VPs je 3600 Logatome präsentiert. Die Messzeit für die 2 x 21600 Darbietungen lag bei etwa 96 Stunden inkl. Training, Instruktionen und Pausen.

Automatische Spracherkennung

Für die ASR Experimente wurde ein Hidden Markov Model mit drei Zuständen und acht Gaußschen Mixtures pro Zustand eingesetzt. Das System wurde so konfiguriert, dass wie bei HSR ein geschlossener Test verwendet wurde, d.h. dass Verwechslungen nur für das mittlere Phonem auftreten konnten. Dies wurde erreicht, indem je ein HMM für jedes äußere Phonem trainiert und getestet wurde. Zusätzlich wurden Delta- und Doppeldelta-Merkmale verwendet, so dass 39-dimensionale Merkmalsvektoren zur Klassifikation eingesetzt wurden. Die ASR-Testdaten enthielt sämtliche Äußerungen, die auch für HSR Experimente verwendet wurden (vier Sprecher mit je 900 Äußerungen); zusätzlich wurden die beiden Wiederholungen, die für jedes Logatom aufgenommen wurden, verwendet. hochdeutschen Sprecher aus OLLO wurden für die Trainingsprozedur gewählt, so dass das ASR System sprecherunabhängig war. Die Häufigkeit von Phonemen war in Trainings- und Testmenge gleichverteilt. ASR Resultate wurden für verschiedene SNRs ermittelt; für Training und Test wurde derselbe SNR verwendet.

Ergebnisse

Erkennungsraten aus HSR-Tests (beide Konditionen) sowie für ASR Experimente bei verschiedenen SNRs sind in Tab. 1 dargestellt. Die grau schattierten Felder zeigen, dass bei einem SNR von -10 dB die ASR-Fehlerrate 167% über der von HSR liegt. Die durchschnittlichen Erkennungsraten sind sich für beide HSR-Konditionen bei einer SNR-Differenz von 10 dB sehr ähnlich (Rahmen mit durchgezogener Linie). Ein Vergleich von HSR-Vokal- und Konsonanterkennung zeigt, dass die Erkennungsrate für Vokalphoneme bei *ungünstigerem* SNR fast 10% *höher* liegt (gestrichelter Rahmen). Dies zeigt, dass Information über Vokalphoneme in MFCCs nur unzureichend kodiert ist. Gleichzeitig ist die Konsonanterkennungsrate für Originalsignale niedriger als für resynthetisierte Logatome. Dies kann auf den ungünstigeren SNR zurückgeführt werden. Für diese Schlussfolgerungen gehen wir davon aus, dass durch den Resyntheseprozess die komplette Information aus den MFCCs wieder hörbar gemacht wird.
Eine Analyse auf Basis von Verwechlungsmatrizen zeigt, dass es Fälle in der Phonemklassifikation gibt, bei denen die Information zur Vermeidung von Fehlern zwar in den Features kodiert ist, diese vom Erkenner jedoch nicht optimal genutzt wird: Bei der Verwechslung der Phoneme /e/ und /i/ treten bspw. für beide HSR-Konditionen relativ geringe Fehleraten auf (15 bzw. 17 %), während die ASR-Fehlerrate mit 27 % deutlich höher ist. In anderen Fällen werden resynthetisierte Phoneme schlechter erkannt als die unveränderten Signale: Bei Originalsignalen wird das Phonem /ʃ/ in 99 % der Fälle richtig klassifiziert, während die Erkennungsraten für die zweite HSR-Kondition und ASR bei nur 94 % liegen. Dies belegt auf mikroskopischer Ebene, dass MFCCs nicht ausreichen, um Sprache im Störgeräusch zu kodieren.

	Kondition	Avg.	VCV	CVC
HSR	Resynth. (0 dB)	72.4	74.5	70.7
	Original (-10 dB)	74.5	67.7	80.5
ASR	clean	80.4	85.2	76.3
	10 dB	76.0	76.4	75.6
	0 dB	64.5	56.4	71.5
	-10 dB	31.8	22.1	40.2

Tabelle 1: HSR- und ASR-Erkennungsraten in Prozent. Die durchschnittlichen Erkennungsraten sind nach Konsonant- und Vokalphonemen aufgeschlüsselt (VCV und CVC).

Schlussfolgerungen

1. Die Lücke zwischen ASR und HSR wird durch überlegene Sprachmodelle des Menschen größer, je komplizierter die Erkennungsaufgabe ist. Für die verhältnismäßig einfache Aufgabe der Phonemerkennung ist die Wortfehlerrate von ASR fast dreimal so hoch wie bei HSR.
2. Erkennungsraten von resynthetisierten und Originalsignalen liegen bei einer SNR-Differenz von 10 dB sehr dicht zusammen. MFCCs enthalten also nicht sämliche Informationen, damit Menschen im Störgeräusch die gleiche Sprachverständlichkeit erreichen wie mit unveränderten Signalen. Der Informationsverlust, der bei der Berechnung zustande kommt, kann also durch eine SNR-Differenz von 10 dB ausgedrückt werden.
3. Die Information, die zur Erkennung von Vokalphonemen benötigt wird, wird durch MFCCs besonders schlecht kodiert. Ursachen hierfür könnte die fehlende Phaseninformation oder die reduzierte spektrale Auflösung sein, die eine Diskrimination zwischen Formanten erschwert.
4. Fehlermuster in Verwechlungsmatrizen können helfen, Fehler in der Klassifikation auf Merkmalsextraktion oder den Klassifizierer selbst einzugrenzen.

Literatur

[1] Wesker T. et al.: Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines. Interspeech 2005.

[2] Demuynck, K. and Garcia, O. and Dirk Van Compernolle: Synthesizing Speech from Speech Recognition Parameters, ICSLP 2004.