# Spracherkennung bei akzentgefärbten Aussprachevarianten

*Stefan Schaden*

Institut für Kommunikationsakustik (IKA), Ruhr-Universität Bochum, D-44780 Bochum
schaden@ika.ruhr-uni-bochum.de

**ZUSAMMENFASSUNG**

Die vorgestellte Pilotstudie geht der Frage nach, welche lautlichen Besonderheiten, die bei der Behandlung fremdsprachlicher Akzente in der automatischen Spracherkennung zu berücksichtigen sind, in Aussprachevarianten von nicht-muttersprachlichen Sprechern auftreten. Zwei mögliche Adaptionsmethoden werden erörtert und anhand eines begrenzten ersten Experiments geprüft.

## 1 Einleitung

Ausgeprägte Abweichungen von der Sprachnorm werfen in der maschinellen Sprachverarbeitung stets besondere Probleme auf. Im Bereich der automatischen Spracherkennung ist dies z.B. der Fall, wenn als potentielle Nutzer eines Erkenners, der für die kanonische (d.h. Norm-) Aussprache optimiert wurde, auch Nicht-Muttersprachler vorgesehen sind, Sprecher also, die einen fremdsprachlichen ‚Akzent' aufweisen. Hier ist die phonetische Diskrepanz zur Normlautung z.T. so erheblich, daß die Leistung von Erkennern oftmals stark absinkt.

Gegenüber intra-lingualen Varianten (Dialekte, Spontansprache) weisen sprachenübergreifende Aussprachevarianten lautliche Besonderheiten auf, die eine maschinelle Behandlung erschweren. Während z.B. dialektale Varianten durch eine relativ stabile und systematische Verschiebung der normsprachlichen Lautdistinktionen gekennzeichnet sind, die sich mittels phonologischer Regeln beschreiben lassen (Fitt 1997), sind fremdsprachliche Akzentvarianten lautlich äußerst heterogen: Nicht-Muttersprachler bilden keine einheitliche Sprechergruppe; sie unterscheiden sich (a) in ihrem muttersprachlichen Hintergrund, (b) in ihren Kenntnissen und Fertigkeiten der zielsprachlichen Phonetik, und ziehen (c) in unterschiedlichem Maße Drittsprachenkenntnisse zur Aussprache des zielsprachlichen Wortschatzes heran.

Am IKA wird derzeit eine Sprachdatenbank mit Aufzeichnungen nicht-muttersprachlicher Sprecher verschiedener Herkunftssprachen aufgebaut. Im Mittelpunkt stehen dabei Deutsch, Englisch und Französisch, die jeweils als Ausgangssprache (**L1**) und Zielsprache (**L2**) herangezogen werden. Die Aufzeichnungen dienen primär einer phonetisch-phonologischen Analyse der auftretenden Aussprachevarianten; darüber hinaus können sie als Testmaterial für Spracherkennungssysteme verwendet werden.

Im folgenden wird exemplarisch anhand eines Ausschnitts des bisher erstellten Korpus erörtert, welche phonetischen Besonderheiten bei der Aussprache fremdsprachlichen Materials auftreten. Dabei scheinen einige der bei den Sprechern zu beobachtenden Fehlertypen spezifische Adaptionsmethoden nahezulegen. Es wurde daher anhand eines hinsichtlich Vokabulargröße (46 Wörter) und Erkennungsszenario (Einzelworterkennung) stark begrenzten Versuchs geprüft, ob und in wie weit sich hieraus Ansätze für Adaptionsmethoden ergeben könnten. Von zentralem Interesse war dabei nicht die absolute Performanz eines spezifischen Systems, sondern vielmehr die auftretenden Differenzen zwischen den einzelnen Sprechern mit ihren jeweils unterschiedlichen Kenntnissen der zielsprachlichen Aussprache.

## 2 Sprachmaterial und Sprecher

Als Sprachmaterial wurden je 46 Ortsnamen aus Frankreich und Großbritannien (England und Schottland) herangezogen. Eigennamen wurden gewählt, weil die Aussprache fremdsprachiger Namen auch für solche Sprecher ein realistisches Szenario darstellt, die über keinerlei Kenntnisse der Zielsprache verfügen (z.B. in einer Reisesituation). Je 6 Muttersprachlern des Englischen und Französischen wurde die Aufgabe gestellt, das in Form von Einzelwortlisten vorliegende Material der jeweiligen Zielsprache (Französisch bzw. Englisch) zu lesen. Die Sprachdaten wurden digital aufgezeichnet.

Durch eine dem Versuch vorausgehende Aufgabe, bei der 10 kurze Sätze in der Zielsprache vorgelesen wurden sowie durch eine Selbsteinschätzung der Sprecher über ihre Kenntnisse der Zielsprache wurde für jeden Sprecher eine sog. *Approximationsstufe* bestimmt. Dieser Wert bezeichnet die Annäherung des Sprechers an die Aussprachenorm der Zielsprache auf einer Skala von 0 bis 3 (0 = keine Annäherung; 3 = zielsprachliche Aussprache). Die Approximationsstufe ermöglicht es, eine Korrelation zwischen Kenntnisstufen der Sprecher und der Erkennungsrate bei verschiedenen Systemkonfigurationen zu ermitteln. In beiden Gruppen findet sich jeweils ein Sprecher mit sehr geringen Kenntnissen der Zielsprache (Approx. 0) sowie ein Sprecher mit relativ hoher Annäherung an die zielsprachliche Aussprachenorm. Die verbleibenden Sprecher verfügen über unterschiedlich ausgeprägte partielle Kenntnisse der zielsprachlichen Phonetik.

## 3 Typen von Aussprachevarianten

Bei den bisher durchgeführten Sprachaufzeichnungen zeigt sich insgesamt eine hohe lautliche Variabilität der von den Sprechern produzierten Aussprachevarianten. Eingangs wurden einige Einflußgrößen genannt – muttersprachliche Herkunft, Kenntnisse der Zielsprache, weitere Sprachkenntnisse –, die als Ursachen für diese Variabilität in Frage kommen.

Doch auch bei gleicher muttersprachlicher Herkunft und vergleichbarer Kenntnisstufe sind oftmals erhebliche Abweichungen zwischen einzelnen Sprechern zu beobachten. Eine linguistisch motivierte *Fehlertypologie* ermöglicht es, die Vielfalt der auftretenden Varianten zunächst auf einige prototypische Realisierungen zu reduzieren. Gleichzeitig ergeben sich hieraus möglicherweise Hinweise darauf, mittels welcher Methode bzw. in welcher Komponente eines Spracherkennungssystems die Aussprachevarianten zu modellieren sind (z.B. Adaption der akustischen Modelle; Einführung erweiterter oder alternativer phonetischer Lexika). An dieser Stelle seien exemplarisch zwei wichtige Fehlertypen genannt und illustriert, die in den bisherigen Versuchen eine wesentliche Rolle spielten.

**Übertragung der Graphem-Phonem-Beziehungen**. Wird, wie es im hier beschriebenen Versuch der Fall war, das fremdsprachliche Material gelesen, so treten Aussprachefehler insbesondere bei Sprechern mit geringen Kenntnissen durch die Übertragung der muttersprachlichen Schriftaussprache auf die Fremdsprache auf (Transfer der Graphem-nach-Phonem-Beziehungen (GTP-Regeln)). Dabei ergeben sich Varianten wie die folgenden:

| Sprache | Ortsname | Referenz | Variante |
|---|---|---|---|
| Frz. | *Mulhouse* | [myluz] | engl. [mʊɬhaʊz] |
| Frz. | *Tarbes* | [taʁb] | engl. [tʰɑːbz] |
| Engl. | *Stafford* | [stæfəd] | frz. [stafɔʁ] |
| Engl. | *Plymouth* | [plɪməθ] | frz. [plimut] |

Häufig treten auch Mischformen auf, bei denen die muttersprachlichen GTP-Regeln nur auf solche Teil-Graphemfolgen der zielsprachlichen Wörter angewendet werden, die der muttersprachlichen Graphotaktik entsprechen.

**Ersatz fremdsprachlicher Phoneme/Allophone**. Aus zahlreichen Studien der kontrastiven Phonetik ist bekannt, daß Sprecher (auch solche mit guten Kenntnissen einer Fremdsprache) die Tendenz aufweisen, fremdsprachliche Phone bzw. Phoneme, die in ihrer Muttersprache nicht vorhanden sind, durch ähnliche muttersprachliche Laute zu ersetzen (vgl. z.B. Wode 1980). Ein bekanntes Beispiel hierfür ist die Ersetzung des engl. Interdentals [θ] (*th*ick) durch [s] oder [f] bei deutschen Sprechern. Häufig sind die zu beobachtenden Substitutionen jedoch weniger stereotyp, so daß eine Prognose des von einem Sprecher (oder einer Sprechergruppe) gewählten Ersatzlautes problematisch ist. So ist z.B. phonetische Ähnlichkeit für die Wahl des bestmöglichen L1-Äquivalents nicht immer ein geeignetes Kriterium: Erstens kommen für einige L2-Phoneme mehrere äquivalente Ersatzlaute in

Frage (z.B. frz. [ɛ] oder [a] für engl. [æ] ); zweitens hat sich in den Versuchen gezeigt, daß die orthographische Repräsentation eines L2-Phonems einen deutlichen Einfluß auf den von den Sprechern gewählten muttersprachlichen Ersatzlaut ausübt.

## 4 Erkennertest

Zur Adaption eines Spracherkenners an spezifische Sprechergruppen oder Sprechstile können Modifikationen prinzipiell an verschiedenen Systemkomponenten erfolgen: (i) auf der Ebene des Lexikons durch Anpassung der phonetischen Transkription, (ii) auf der Ebene der akustischen Modelle (z.B. Neutraining) oder (iii) – bei kontinuierlichen Erkennern – auf der Ebene des Sprachmodells (vgl. Strik & Cucchiarini 1999).

Im vorgestellten Versuch wurden Modifikationen des Erkenners (1) durch Manipulation des Lexikons und (2) durch einen Wechsel der sprachenspezifischen Phonem-Modelle vorgenommen. Ermittelt wurden die Erkennungsraten für die oben beschriebenen 6 Muttersprachler des Englischen und Französischen (für das jeweils fremdsprachliche Vokabular), wobei folgende drei Systemkonfigurationen miteinander verglichen wurden:

**L2/L2: Lexikon L2 und Phonem-Modelle L2**. Die Kombination eines Lexikons für die Standard-Aussprache der Zielsprache mit zielsprachlichen Phonem-Modellen entspricht der Standard-Konfiguration des Erkenners (*Baseline-System*). Es ist zu erwarten, daß mit dieser Konfiguration eine optimale Erkennungsrate nur bei Sprechern mit guten bis sehr guten Kenntnissen der zielsprachlichen Aussprache erzielt werden kann.

**L2/L1: Lexikon L2 und Phonem-Modelle L1**. Es wurde oben dargelegt, daß ein wichtiger Fehlertyp in einer Substitution von zielsprachlichen Phonemen/Allophonen durch muttersprachliche Laute besteht. Als Adaptionsmethode für diesen Typ akzentgefärbter Aussprache wurde ein zielsprachliches Referenzlexikon mit Phonem-Modellen der Ausgangssprache kombiniert. Da sich die Phonem-Inventare der beteiligten Sprachen unterscheiden, ist es zuvor erforderlich, einige zielsprachliche Phoneme auf das ausgangssprachliche Inventar abzubilden (Phonem-Mapping). Hierdurch wird die zugrundeliegende Transkription dem ausgangssprachlichen Lautinventar angeglichen.

**L1/L1: Lexikon L1 und Phonem-Modelle L1**. Zur Modellierung von Fehlern, denen ein Transfer der muttersprachlichen GTP-Beziehungen auf die Fremdsprache zugrundeliegt, wurden Lexika erstellt, in denen die zielsprachlichen Worteinheiten mittels eines Graphem-nach-Phonem-Umsetzers nach den Regeln der Ausgangssprache gezielt fehlerhaft phonemisiert wurden (Englisch nach französischen Regeln; Französisch nach englischen Regeln). Die so generierten Lexika wurden mit den ausgangssprachlichen Phonem-Modellen kombiniert. Diese Methode hat zunächst einen rein explorativen Charakter: Da die verwendeten GTP-Regelsätze für den Wortschatz von L1 entworfen wurden, entstehen, wenn man sie auf L2-Wortschatz anwendet, z.T. recht artifizielle Phonemfolgen, die bei realen Sprechern nicht auftreten. Denn es ist zu beobachten, daß auch Sprecher mit geringen L2-Kenntnissen die GTP-Regeln ihrer Muttersprache nur selten völlig unmodifiziert auf die Fremdsprache übertragen. Langfristig wäre hier der Aufbau eines sprachenübergreifenden erweiterten GTP-Umsetzers sinnvoll, mit dem solche Teilkenntnisse der Sprecher gezielt modelliert werden können.

In Tab. 1 und 2 sind die erzielten Erkennungsraten bei den oben beschriebenen Systemkonfigurationen dargestellt. Als Referenzwert wurde zuvor die Erkennungsrate bei einem Muttersprachler der Zielsprache ermittelt (*reference*). Gegenüber diesem Referenzwert ist bei beiden Sprachrichtungen ein deutliches Absinken der Erkennungsraten bei Nicht-Muttersprachlern zu beobachten. Es sind jedoch deutliche Unterschiede bei verschiedenen Sprechern und Konfigurationen festzustellen. Es zeigt sich die generelle Tendenz, daß bei niedriger Approximationsstufe des Sprechers die höchste Erkennungsleistung bei der Konfiguration **L1/L1** (ausgangssprachliche Phonem-Modelle und phonetische Umschrift nach den Regeln der Ausgangssprache) erzielt wird. Es handelt sich hierbei um Sprecher, die in weiten Teilen die muttersprachlichen GTP-Beziehungen auf die Zielsprache übertragen.

Die Kombination von zielsprachlicher Phonemisierung mit ausgangssprachlichen Phonem-Modellen (**L2/L1**) hat sich bei den untersuchten Sprachrichtungen nicht bewährt. Dieses Ergebnis hängt jedoch auch von den untersuchten Sprachenkombinationen und von der Möglichkeit eines adäquaten Phonem-Mappings ab: In einer weiteren durchgeführten Untersuchung in der Sprachrichtung L1 Deutsch → L2 Englisch profitierten einige deutsche Sprecher von einer Kombination von englischem Referenzlexikon mit deutschen Phonem-Modellen. Es ist denkbar, daß sich bestimmte Sprachenkombinationen besser für ein Phonem-Mapping eignen als andere.

Die Kombination **L2/L2** schließlich führte erwartungsgemäß nur bei Sprechern mit guten Kenntnissen zu den höchsten Erkennungsraten. Eine Annäherung an den Referenzwert (muttersprachlicher Sprecher) konnte jedoch nur bei einem Sprecher (*bpw*) erzielt werden.

| **Sprecher** | **Approx. (0-3)** | **Correct recognition %** | | |
|---|---|---|---|---|
| | | **L2/L2** | **L2/L1** | **L1/L1** |
| **reference** | (3) | **93.5** | | |
| cmi | 0 | 52.2 | 56.5 | **65.2** |
| jm | 1 | 69.6 | 65.2 | **71.7** |
| cro | 1 | 43.5 | 47.8 | **65.2** |
| km | 1 | **60.9** | 52.2 | 52.2 |
| jny | 2 | **69.6** | 45.7 | 45.7 |
| tdd | 2 | 58.7 | 69.6 | **76.1** |

*Tab. 1: Erkennungsrate L1 Englisch → L2 Französisch*

| **Sprecher** | **Approx. (0-3)** | **Correct recognition %** | | |
|---|---|---|---|---|
| | | **L2/L2** | **L2/L1** | **L1/L1** |
| **reference** | (3) | **91.3** | | |
| mb | 0 | 57.8 | 50.0 | **71.7** |
| pm | 1 | 32.6 | 34.8 | **39.1** |
| mep | 1 | 52.2 | 47.8 | **58.7** |
| asg | 2 | **50.0** | 32.6 | 34.8 |
| iw | 2 | **56.6** | 52.2 | 39.1 |
| bpw | 2 | **87.0** | 80.4 | 56.5 |

*Tab. 2: Erkennungsrate L1 Französisch → L2 Englisch*

## 5 Schlußfolgerung und Ausblick

Die hohe Streuung der erzielten Erkennungsraten bei verschiedenen Sprechern und Systemkonfigurationen reflektiert die hohe interindividuelle Variabilität bei der Aussprache fremdsprachigen Wortschatzes. Zwar ist die Grundtendenz zu beobachten, daß Sprecher mit sehr geringen Kenntnissen der Zielsprache, deren Aussprache durch eine Übertragung der muttersprachlichen Schriftaussprache-Regeln auf die Fremdsprache gekennzeichnet ist, von einem gezielt fehlerhaften, nach den GTP-Regeln von L1 generierten Lexikon profitieren können. Bei Sprechern mit durchschnittlichen Kenntnissen erweist es sich hingegen als schwierig, eindeutige Hinweise aus der Untersuchung abzuleiten.

Eine wichtige Vorarbeit für eine Adaption von Systemen an nichtmuttersprachliche Nutzer besteht zweifellos darin, die von den Sprechern herangezogenen Teilkenntnisse und -fertigkeiten hinreichend differenziert – und für jede Sprachrichtung gesondert – zu beschreiben. Dies ermöglicht es, spezifische Kenntnisstufen von Sprechern prototypisch zu erfassen und in sinnvoller Weise verschiedenen System-Komponenten zuzuordnen, in denen diese Teilkenntnisse zu modellieren sind.

## 6 Literatur


Fitt, S. (1997): "The generation of regional pronunciations of English for Speech Synthesis." Proc. *Eurospeech 1997*, pp. 2447-50.

Strik, H./Cucchiarini, C. (1999): "Modeling Pronunication Variation for ASR: Overview and Comparison of Methods." *Speech Communication* 29, pp. 225-246.

Wode, H. (1980): "Phonology in L2 acquisition." In: Felix, S.W., (ed.) *Second Language Development. Trends and issues*. Tübingen: Narr, pp. 123-36.



*Diese Studie entstand am Institut für Kommunikationsakustik der Ruhr-Universität Bochum (Prof. Jens Blauert) unter der Betreuung von PD Dr. Ute Jekosch.*