Fehlerdiagnose und Feedback in einem Aussprachetrainer für den Fremdsprachenerwerb

Wolfgang Menzel
Fachbereich Informatik, Universität Hamburg, Vogt-Kölln-Straße 30, D-22527 Hamburg

Zur Verbesserung der Feedback-Qualität von Computersystemen für das Aussprachetraining sind detaillierte und präzise Kenntnisse über die jeweiligen Aussprachefehler eines Schülers erforderlich. Um derartige Informationen verfügbar zu machen, wurde eine Reihe von Komponenten zur phonetischen Diagnose von gesprochenen Äußerungen realisiert. Ziel des Projekts war die Integration von Standardtechnologie zur Spracherkennung in eine Umgebung zum Aussprachetraining für erwachsene Lerner auf mittlerem Sprachniveau. Obwohl die hier beschriebenen Verfahren prinzipiell auf beliebige Sprachpaare übertragbar sind, ist das derzeitige System auf italienische und deutsche Muttersprachler mit der Zielsprache Englisch optimiert.

Einleitung

Dank der rasanten Fortschritte auf dem Gebiet der automatischen Spracherkennung sind heutzutage in zunehmendem Maße computer-gestützte Lösungen für das Aussprachetraining beim Fremdsprachenlernen verfügbar (Dalby, Kewley–Port, and Sillings 1998; Herron, Menzel, Atwell, Bisiani, Daneluzzi, Morton, and Schmidt, 1999; Eskenazi and Hansma, 1998, Witt and Young, 1997, 1998). Aus fremdsprachendidaktischer Sicht sind dabei zwei Entwurfskriterien von besonderer Bedeutung:

1. Das Aussprachetraining sollte nicht primär als rein reproduktive Aktivität, sondern so weit dies möglich ist, als Teil eines komplexen, aber kreativen Sprachproduktionsprozesses betrachtet werden. Daher ist es erforderlich, derartige Übungen in ein Szenario zu integrieren, das den Schüler in erster Linie dazu ermutigt, seine kommunikativen Absichten zu verfolgen. Aspekte der Aussprache und andere formbezogenen Gesichtspunkte stellen in solch einem Kontext eigentlich nur Hilfsmittel zur erfolgreichen Realisierung dieser Ziele dar.
2. Das Trainingssystem sollte ausreichend Feedback-Information zur Verfügung stellen, die es dem Studenten erlaubt, mögliche Unzulänglichkeiten zu erkennen, und die ihn deutlich auf Wege zur Verbesserung der Artikulation hinweist.

Elementare Voraussetzung, um in diesem Zusammenhang überhaupt von kommunikativem Handeln sprechen zu können, ist jedoch eine wenigstens minimale Auswahlmöglichkeit bei der Formulierung der Schülereingaben, wobei das System in der Lage sein muss festzustellen, welche der Optionen vom Schüler tatsächlich gewählt wurde. Hierfür ist eine Komponente zur Spracherkennung unerlässlich. Über die reine Spracherkennungsfähigkeit hinaus sind aber hochentwickelte Analysewerkzeuge erforderlich, die die notwendigen aussprachebezogenen Diagnoseinformationen aus dem Sprachsignal extrahieren können. Gerade in diesem Bereich zeigen die bisher verfügbaren Systemlösungen jedoch noch die größten Schwächen. Üblicherweise sehen sie die Möglichkeit zur Aufzeichnung und Wiedergabe von Sprachsignalen vor, gekoppelt mit einem einfachen globalen Bewertungsmechanismus und einer visuellen Aufberei-

tung ausgewählter Signalparameter (z.B. Amplituden- und Grundfrequenzverlauf). Im Hinblick auf eine *Verbesserung* der Aussprache sind diese Hilfsmittel allein jedoch nicht ausreichend, da sie die Aufgabe der Fehleridentifizierung ganz dem Schüler überlassen. Gerade dazu ist dieser aber oftmals gar nicht in der Lage. Selbst dann, wenn eine Bewertung auf Wortformenebene das fehlerhaft ausgesprochene Wort direkt ausweist, wird für den Schüler noch nicht unbedingt deutlich, wie die Aussprachequalität erhöht werden kann. Erst durch eine detaillierte Analyse des Signals direkt auf der Ebene der Laute kann das System in die Lage versetzt werden, dem Schüler Erklärungen zur Art des Fehlers, sowie speziell auf das Problem zugeschnittene Sprachbeispiele bzw. gar individualisierte Verbesserungshinweise anzubieten. Darüberhinaus ermöglicht eine solche Diagnose auch die automatische Generierung individuell angepasster Möglichkeiten zum wiederholenden Üben.

Um das Potential der derzeit verfügbaren Spracherkennungs-Technologie für die Entwicklung einer neuen Generation von Werkzeugen zum Aussprachetraining auszuloten, wurde eine Reihe von Komponenten für die automatische Lokalisierung und Diagnose von Aussprachefehlern entwickelt und in eine multimediale Lernumgebung integriert. Das System wurde in erster Linie im Hinblick auf italienische und deutsche Lerner des Englischen optimiert, die zugrundeliegenden technischen Lösungen sind jedoch weitgehend generisch und gestatten daher eine Übertragung auf beliebige Sprachpaare.

Das System bietet dem Schüler eine große Vielfalt kommunikativ orientierter Übungen (wie z.B. das Beantworten von Fragen oder die Teilnahme an einem simulierten Dialog), wobei der Schüler seine Antwort jeweils aus einer kleinen Auswahl von vordefinierten Bausteinen zusammensetzen kann. Die Entscheidung darüber, welche Art von Feedback-Information er im Einzelfall nutzen möchte, bleibt aber stets dem Schüler selbst überlassen.

Systemarchitektur

Das System setzt sich aus verschiedenen Modulen zusammen, die auf drei Ebenen angeordnet sind. Auf der obersten Ebene, realisiert ein Standard-Autorenwerkzeug die Kommunikationsschnittstellen mit dem Schüler einerseits und dem Übungsautor andererseits. Hierunter fallen sowohl die Präsentation des Übungsmaterials, als auch die Verwaltung und Synchronisation der multimodalen Interaktion mit dem Schüler: Ein- und Ausgabe gesprochener Sprache, Visualisierung von Feedback-Information, sowie die Selektion von Auswahloptionen mit der Maus.

Auf der untersten Ebene der Systemarchitektur befinden sich die Komponenten zur Analyse des Sprachsignals. Eine zentrale Montitorkomponente vermittelt den Informationsfluss zwischen diesen beiden Architekturebenen und ruft die erforderlichen Analysemodule in Abhängigkeit von den Eingaben des Schülers auf.

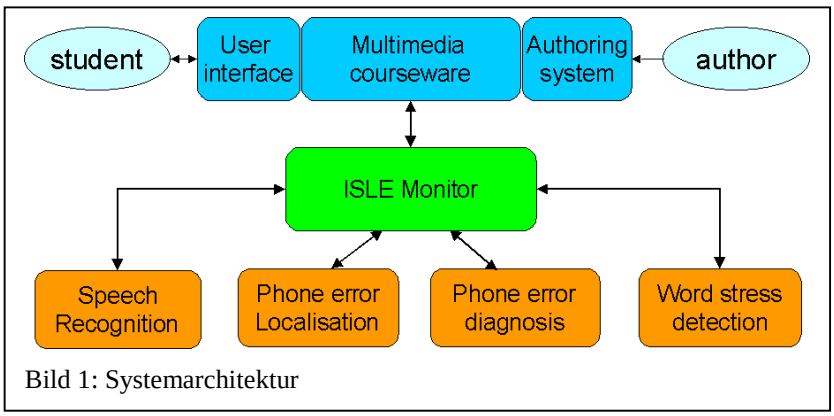

Bild 1: Systemarchitektur

Das vom Schüler eingegebene Sprachsignal wird grundsätzlich zuerst an die Spracherkennungskomponente übergeben, um festzustellen, was der Schüler

gesagt hat. Somit kann bei den anderen Modulen zur Qualitätsbewertung und Fehlerdiagnose stets davon ausgegangen werden, dass die tatsächlich gesprochene Äußerung bereits bekannt ist. Für den Fall einer erfolgreichen Erkennung versucht das System, zuerst diejenigen Stellen im Signal zu identifizieren, die mit hoher Wahrscheinlichkeit einen Erkennungsfehler enthalten. Diese Fehlerlokalisierung basiert auf den Konfidenzwerten des Spracherkenners für Worte und Phone. Auf Nachfrage wird das Signal weitergehend analysiert, um Phonvertauschungen, -einfügungen und –auslassungen, sowie fehlerhafte Wortbetonungen festzustellen.

Die Sprachanalysekomponenten

Die *Erkennung* von hochgradig akzentbehafteter Lernersprache ist eine ausgesprochen schwierige und heutzutage noch weitgehend ungelöste Aufgabe. Sie läßt sich derzeit nur dann mit ausreichender Sicherheit durchführen, wenn die Erkennungsaufgabe selbst sehr einfach gehalten wird. Hierzu kann die Menge der zulässigen Äußerungen mit Hilfe einer Grammatik auf eine relativ kleine Auswahl eingeschränkt werden. Zudem sollte das lexikalische Inventar so ausgewählt werden, dass phonetisch ähnliche Wörter an der gleichen Satzposition vermieden werden. Der im System eingesetzte Erkenner (Morton et al., 1999) erzielt nach einer kurzen Adaption auf den jeweiligen Sprecher 100% Erkennungsrate auf sehr einfachen Erkennungsaufgaben, wie sie etwa bei der (akustischen) Auswahl aus drei oder vier kurzen Phrasen anzutreffen sind (Perplexität kleiner als 1.5). Ein Beispiel für eine komplexere Erkennungsaufgabe ist in Bild 2 wiedergegeben.

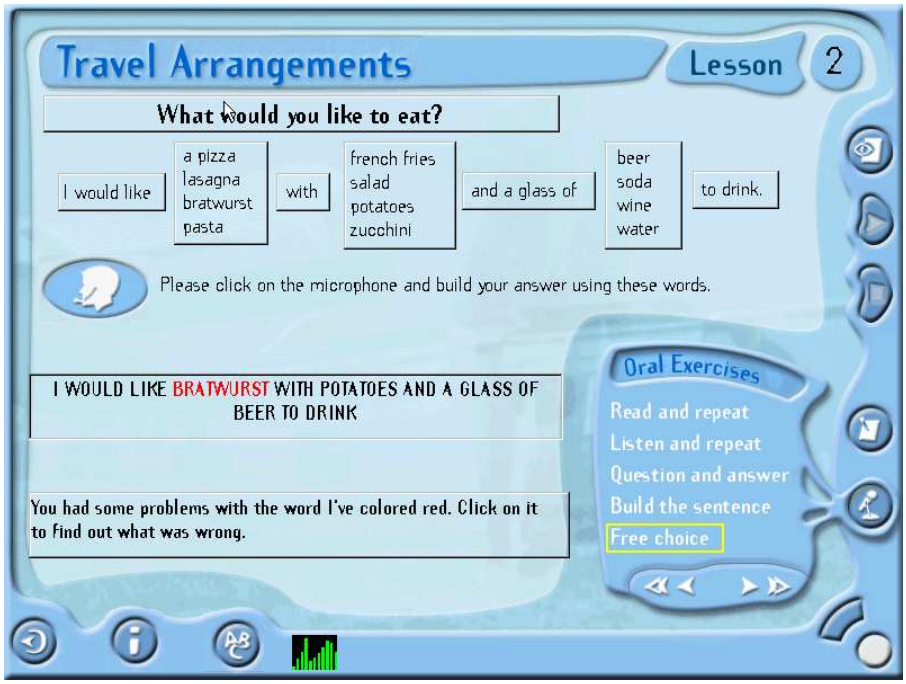

Bild 2: Beispiel einer komplexen Menüselektion

Bei Aufgaben dieses Typs werden noch Erkennungsraten oberhalb von 96% erreicht. Als besonders schwierig gelten Aufgaben mit hoher phonetischer Verwechselbarkeit, z.B. Minimalpaarübungen mit Sätzen der Art: *I said ship not sheep.* Hier sinkt die Erkennungsrate auf ca. 94%.

Fehlerlokalisationsverfahren verwenden üblicherweise direkt die akustischen Bewertungen des Spracherkenners. Es hat sich aber herausgestellt, dass mit einem indirekten Verfahren, das Konfidenzwerte auf der Basis von Normalverteilungsklassifikatoren berechnen, eine höhere Zuverlässigkeit erreicht werden kann. Die Prädiktion berücksichtigt drei Wahrscheinlichkeitswerte, die zu jedem Analysezeitpunkt aufgrund der Beobachtung des Sprachsignals geschätzt werden: (1) die Bewertung des erkannten Pfads durch das akustische Modell, (2) die Ausgabewahrscheinlichkeit des

wahrscheinlichsten Modellzustandes und (3) die akustische Wahrscheinlichkeit eines Hintergrundmodells. Dabei wird davon ausgegangen, dass der beste Zustand eine obere und das Hintergrundmodell eine untere Schranke für die Bewertung liefern und dass die akustische Bewertung der erkannten Lautfolge demnach zwischen diesen beiden Extremen liegen sollte. Separate Klassifikatoren wurden für verschiedene Phonklassen (z.B. Vokale, Frikative, Liquide) trainiert.

Die eigentliche *Diagnose* von Phonfehlern folgt als dritter Schritt. Sie ermittelt, in welcher Hinsicht das vom Schüler gesprochene Wort von der im Aussprachewörterbuch angegebenen kanonischen Aussprache abweicht. Dazu wird eine große Anzahl von Fehlerhypothesen gegen das Sprachsignal evaluiert. Die Erzeugung der Fehlerhypothesen erfolgt mit Hilfe von muttersprachenspezifischen Phon-Phon- bzw. Graphem-Phon-Transformationsregeln, sowie zielsprachenspezifischen Graphem-Phon-Regeln. Die generierten Fehlerhypothesen werden temporär dem Wörterbuch hinzugefügt und durch einen erneuten Erkennerlauf mit nichtadaptierten Phonmodellen auf ihre Plausibilität bewertet. Auf diese Weise lassen sich artikulatorische Schwierigkeiten der Zielsprache (z.B. der notorisch schwierige /th/-Laut im Englischen), muttersprachlich bedingte rezeptive Schwierigkeiten (z.B. der Unterschied zwischen /ih/ and /iy/ für italienische Sprecher), orthographische Interferenz der Muttersprache (z.B. die Aussprache des Buchstabens z als homorganer Affrikat im Deutschen) und orthographische Schwierigkeiten der Zielsprache (z.B. der hohe Grad an phonologischer Mehrdeutigkeit im Englischen) modellieren.
Die Verwendung solcher Transformationsregeln erlaubt es, die Größe des zu untersuchenden Hypothesenraums relativ genau zu beeinflussen. Letztendlich ist selbst die Hypothesenvielfalt, die ein einigermaßen vollständiger Regelsatz erzeugt, noch klein im Vergleich zu einer gänzlich unbeschränkten Phonerkennung. Dennoch werden für den praktischen Einsatz zusätzliche Einschränkungen benötigt, um den Berechnungsaufwand in akzeptablen Grenzen zu halten. So berücksichtigt das Demonstrationssystem maximal einen Fehler pro Wort, während bei der Ermittlung der unten angegebenen Performanzdaten maximal drei Fehler pro Wort zulässig waren.

Die *Ermittlung der Wortbetonung* ist der letzte der Verarbeitungsschritte. Er basiert ebenfalls auf einem binären Normalverteilungsklassifikator, der als Merkmale Lautdauer, Energie, sowie einige spektrale Kenngrößen der Vokalregion berücksichtigt. Für alle mehrsilbigen Wörter wird die so ermittelte Position der Hauptbetonung anschließend mit dem im Wörterbuch vermerkten kanonischen Betonungsmuster verglichen und jede Abweichung als mögliches Diagnoseresultat der Fehlererklärung zur Verfügung gestellt.

Zur Evaluation der drei Diagnosekomponenten wurden die Resultate der automatischen Verarbeitung analogen Entscheidungen von phonetisch geschulten Linguisten gegenübergestellt. Zu diesem Zweck war es erforderlich, ein größeres Korpus von englischer Lernersprache manuell auf der Phonebene zu annotieren (Menzel, Atwell, et al., 2000; Bonaventura, et al., 2000). Dieses Korpus enthält ca. 20 Minuten Sprache von jeweils 23 deutschen und italienischen Muttersprachlern mit mittlerem Sprachniveau im Englischen. Die Annotation wurde so durchgeführt, dass sie auch Aussagen über die Konsistenz der Gutachterurteile ermöglicht. Bild 3 zeigt eine Gegenüberstellung der Zuverlässigkeit für automatisch und manuell ermittelte Diagnoseergebnisse.
Es ist unmittelbar erkennbar, dass die Trefferquoten für alle drei Komponenten verhältnismäßig gering ausfallen, jedoch auch nicht dramatisch schlechter sind, als die der menschlichen Experten. Eine Ausnahme bildet offenbar die

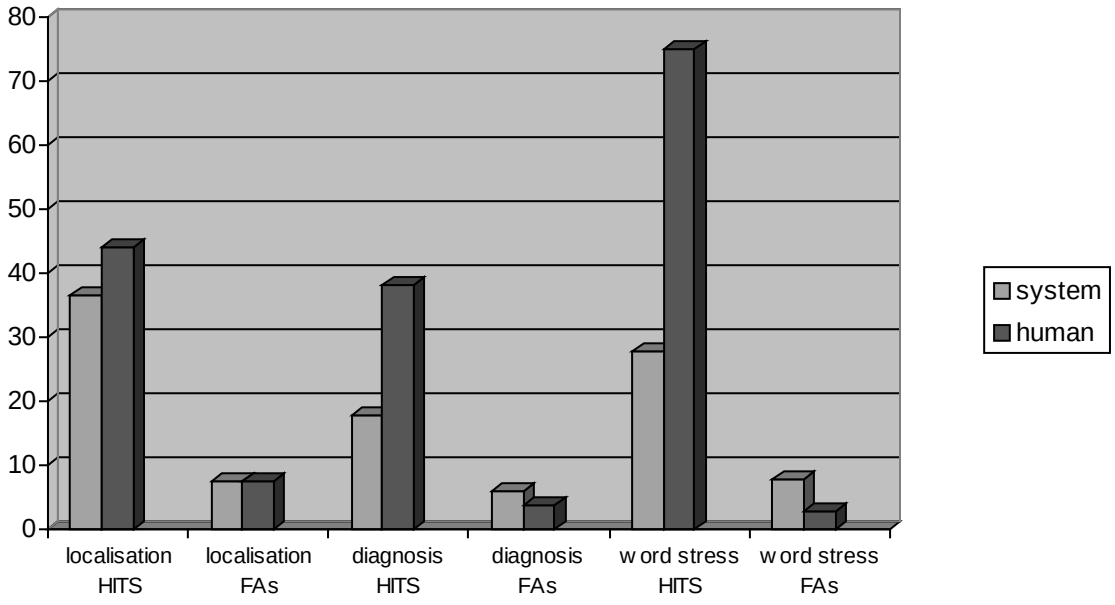

Bild 3: Leistungsfähigkeit der Diagnosekomponenten im Vergleich zu menschlichen Experten. Angegeben ist die Trefferquote (HITS) und die relative Anzahl von Fehlalarmen (FAs) beide in Prozent.

Ermittlung der Wortbetonung, die von Menschen mit hoher Sicherheit, von der implementierten Komponente hingegen nur sehr mangelhaft durchgeführt werden kann. Niedrige Trefferquoten dürften für den praktischen Einsatz jedoch nicht als problematisch angesehen werden, da die verbleibende absolute Zahl der gefundenen Fehler immer noch recht hoch ist. Unangenehm machen sich hingegen die relativ hohen Fehlalarmraten bemerkbar, da ungerechtfertigte Fehlerdiagnosen erhebliche Verwirrung beim Schüler auslösen können. Gerade für fortgeschrittene Lerner mit einer verhältnismäßig niedrigen *a priori*-Fehlerwahrscheinlichkeit kann dies kaum toleriert werden. Allerdings bieten sich auf der Ebene der Übungsgestaltung einige „Tricks“ an, mit denen sich die vom Schüler wahrnehmbaren Auswirkungen einer hohen Fehlalarmrate deutlich reduzieren lassen: (1) die Übungen sollten schwierig genug sein, damit sie den Schüler zu möglichst vielen Aussprachfehlern provozieren und (2) die Fehlererklärungen sollten sich auf die Fehler mit hoher *a priori*-Wahrscheinlichkeit bzw. hoher Diagnosezuverlässigkeit konzentrieren.

Schließlich kann auch noch durch eine stärkere Integration der Diagnosekomponenten eine Erhöhung der Zuverlässigkeit selbst erreicht werden. So ist z.B. eine Verlagerung der Wortbetonung praktisch immer von einer Änderung der Vokalqualitäten bei den betroffenen Silbenkernen begleitet. Wortbetonungsfehler sollten daher nur dann zur Erklärung herangezogen werden, wenn sie mit den Phonfehlerdiagnosen verträglich sind.

Feedback

Basierend auf den vielfältigen Diagnosergebnissen, erhält der Student die Möglichkeit, zunehmend detailliertere Informationen über die Position und die Art der möglichen Aussprachefehler anzufordern. Dabei werden zuerst nur die

möglicherweise fehlerbehafteten Wörter markiert und es besteht die Gelegenheit, zum wiederholten Anhören und Vergleichen mit der Modellrealisierung eines Muttersprachlers. Alternativ dazu kann auch eine detailliertere Diagnose angefordert werden, wie sie etwa in Bild 4 wiedergegeben ist.

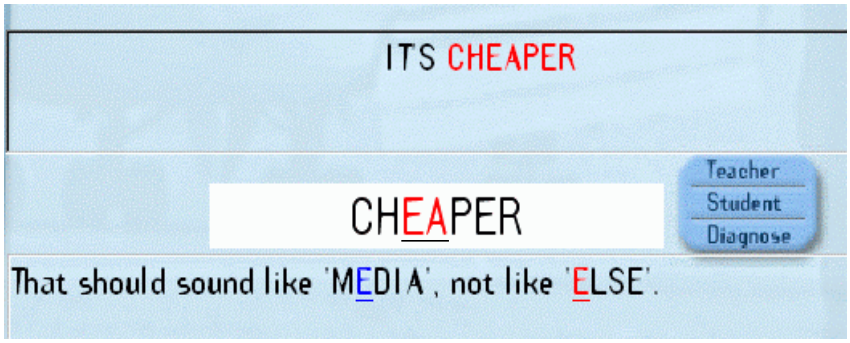

Bild 4: Fehlererklärung einer Phonvertauschung mit Hilfe von Beispielwörtern

Durch Anklicken der betreffenden Wörter und Phone kann der Schüler erneut die eigene Aussprache mit der eines Muttersprachlers vergleichen, wobei nunmehr aber die betroffenen Laute in einem anderen als dem ursprünglich verwendeten Kontext präsentiert werden! Da hierbei die gesamte grafische Interaktion auf der Ebene der Schriftsprache erfolgt, wird jeder Bezug auf ein möglicherweise unbekanntes phonetisches Notationssystem vermieden.

Über die Fehlererklärungen hinaus wird dem Schüler auch angeboten, von speziell auf den jeweils beobachteten Fehler zugeschnittenen Folgeübungen Gebrauch zu machen, um den Problemlaut in identischen bzw. variierenden Kontexten, oder aber auch in einem Minimalpaarkontrast zu üben. Zusätzlich verwaltet das System eine Liste der häufigsten Fehler und hält auch hierzu spezielle Übungsangebote bereit.

Das Demonstrationssystem wurde in einem ersten Feldtest von 15 Lernenden (darunter 6 Italiener und 9 Deutsche), sowie 13 Lehrern (von denen 8 Deutsch und 5 Englisch als Muttersprache angaben) bewertet. Alle Versuchspersonen nahmen an einer betreuten Sitzung von etwa einer Stunde teil und wurden anschließend gebeten, einen Fragebogen auszufüllen, in dem sie die Systemleistung auf einer Skala von eins (sehr gut) bis fünf (sehr schlecht) einstufen sollten. Dabei ergab sich für die drei wichtigsten Fragestellungen das folgende Bild:

War das Feedback	Lehrer		Schüler		Durchschnitt
	Englisch	Deutsch	Italienisch	Deutsch	
... verständlich?	2.2	2.6	1.8	2.3	2.2
... präzise?	3.8	2.4	1.5	2.3	2.5
... hilfreich?	2.6	2.0	1.7	2.0	2.1

Mit Ausnahme der Genauigkeitsurteile der englischsprachigen Lehrer zeigen diese Ergebnisse eine erstaunlich große Akzeptanz in den verschiedenen Nutzergruppen. Da unter den Versuchspersonen die italienischen Schüler einen deutlich geringeren Grad der Fremdsprachenbeherrschung aufwiesen, könnten diese Ergebnisse auch darauf hindeuten, dass die Akeptanz unter Schülern mit geringerem Lernfortschritt besonders groß ist.

Zusammenfassung

Die Erfahrungen mit der Entwicklung eines computer-gestützten Aussprachetrainers haben gezeigt, dass die derzeit verfügbare Technologie zur Verarbeitung gesprochener Sprache durchaus in der Lage ist, die Qualität und die Aussagekraft des Feedbacks bei Aussprachefehlern deutlich zu verbessern. Dank seiner Fähigkeiten zur Phonfehler-lokalisierung und -diagnose ist das System nicht nur in der Lage, dem Schüler hilfreiche Hinweise zur Verbesserung seiner Aussprache zu geben, sondern auch Übungen bereitzustellen, die genau auf seine individuellen Bedürfnisse zugeschnitten sind. Ein solches Leistungsangebot stellt einen deutlichen Fortschritt gegenüber anderen Systemlösungen dar, die in dieser Hinsicht viel stärker eingeschränkt sind.

Verbesserungsbedarf besteht insbesondere noch bei der Zuverlässigkeit der Diagnoseresultate. Dennoch deuten die Ergebnisse einer ersten Nutzerbefragung darauf hin, dass akzeptable Lösungen mit möglicherweise beschränkten Diagnosefähigkeiten schon mit den derzeit verfügbaren technologischen Mitteln erreicht werden können.

Danksagung

Diese Forschung wurde von der Europäischen Gemeinschaft als Projekt ISLE im Language Engineering Programme unter dem Kennzeichen LE4–8353 gefördert. Beteiligt waren Entropic Cambridge Research Labs (Cambridge, UK), Ernst Klett Verlag (Stuttgart), Dida*El (Milano, Italy), Leeds University (UK), Università degli Studi di Milano-Bicocca (Italy) und Universität Hamburg (Germany). Mein besonderer Dank gilt Daniel Herron, Rachel Morton, Peter Howarth, Roberto Bisiani, Alberto Petitti, Dario Pezzotta, Patrizia Bonaventura, Eric Atwell, Clive Souter, Jürgen Schmidt und Hartmut Witt, die mit ihrem Einsatz und ihren unterschiedlichen Beiträgen alle zum Erfolg des Projekts beigetragen haben.

Literatur

Bonaventura P., Howarth P., and Menzel W. (2000) *Phonetic annotation of a non-native speech corpus*. Proc. Int. Workshop on Integrating Speech technology in the (Language) Learning and Assistive Interface, InStil 2000, Dundee, p. 10-17.

Herron D., Menzel W., Atwell E., Bisiani R., Daneluzzi F., Morton R., Schmidt J.A. (1999). *Automatic localization and diagnosis of pronunciation errors for second-language learners of English*, Proceedings 6th European Conference on Speech Communication and Technology, Eurospeech 99, Budapest, v. 2, p. 855-858.

Dalby J., Kewley–Port D., and Sillings R. (1998). *Language–specific pronunciation training using the HearSay system*. Proc. Int. Workshop on Speech Technology in Language Learning, STiLL 98, Marholmen, Sweden, p. 25-28.

Eskenazi M. and Hansma S. (1998). *The Fluency pronunciation trainer*. Proc. Int. Workshop on Speech Technology in Language Learning, StiLL 98, Marholmen, Sweden, p. 77-80.

Menzel W., Atwell E., Bonaventura P., Herron D., Howarth P., Morton R., and Souter C. (2000). *The ISLE corpus of non-native spoken English*. Proc. 2nd International Conference on Language Resources and Evaluation, LREC 2000, Athens, Greece, p. 957-963.

Menzel, W., Herron, D., Bonaventura, P., Morton, R. (2000) *Automatic detection and correction of non-native English Pronunciation.* . Proc. Int. Workshop on Integrating Speech technology in the (Language) Learning and Assistive Interface, InStil 2000, Dundee, p. 49-56.

Morton R., Whitehouse D., and Ollason D. (1999). *The HAPI book*. Entropic Cambridge Research Laboratory, Ltd.

Witt S.M. and Young S.J. (1997). *Language learning based on non–native speech recognition*. Proc. 5th European Conference on Speech Communication and Technology, Eurospeech 97, Rhodes, Greece, p. 633-636.

Witt S.M. and Young S.J. (1998). *Performance measures for phone–level pronunciation teaching in CALL*. Proc. Int. Workshop on Speech Technology in Language Learning, StiLL 98, Marholmen, Sweden, p. 99-102.