

# VERBESSERUNG DER AUTOMATISCHEN TRANSKRIPTION VON ENGLISCHEN WÖRTERN IN DEUTSCHEN VORLESUNGEN

*Sebastian Ochs, Matthias Wölfel und Sebastian Stüker*

*Institut für Theoretische Informatik, Universität Karlsruhe (TH)  
Am Fasanengarten 5, 76131 Karlsruhe, Germany*

sebastian.ochs@stud.uni-karlsruhe.de

**Kurzfassung:** Um deutsche, technische Vorlesungen erfolgreich von einem automatischen Spracherkennungssystem transkribieren zu lassen, müssen neben deutschen Wörtern auch fremdsprachliche Wörter, z.B. aus dem Englischen, erkannt werden. Das Erkennen von Wörtern aus anderen Sprachen ist allerdings ein grundlegendes Problem in der automatischen Spracherkennung, denn die Systeme sind in der Regel auf einzelne Sprachen trainiert. Eine Analyse von auf deutsch gehaltenen, technischen Vorlesungen an der Universität Karlsruhe (TH) ergab, dass ca. 2% aller auftretenden Wörter der englischen Sprache zugehörig sind. Es wurde festgestellt, dass ein deutsches Spracherkennungssystem auf diesen englischen Wörtern, obwohl im Wörterbuch vorhanden, signifikant höhere Fehlerraten produziert als auf deutschen Wörtern. Das Ziel dieser Arbeit bestand darin, den Einfluss der englischen Wörter zu untersuchen und verschiedene Maßnahmen zu entwickeln und anzuwenden, um die Erkennungsleistung eines deutschen Spracherkennungssystems in Bezug auf englische Wörter zu verbessern. Es werden zwei Ansätze vorgestellt:

1. Die Abbildung von englischen Phonemen auf deutsche, um die Aussprache von englischen Wörtern zu approximieren.
2. Die parallele Verwendung eines deutschen und eines englischen akustischen Modells.

Durch die hier vorgestellten Verfahren war es möglich die Wortfehlerrate englischer Wörter von 60,0% auf 34,6% zu reduzieren. Hierbei wurde die Gesamtwortfehlerrate von 13,8% auf 12,7% verbessert.

## 1 Einleitung

Die automatische Transkription von Lehrveranstaltungen durch Spracherkennungssysteme ermöglicht die Erschließung neuer Dienstleistungen in der Lehre, denn das Vorhandensein von Transkriptionen ist Voraussetzung für vielfältige Anwendungen wie z.B. die automatische Übersetzung oder Zusammenfassung von Vorlesungen sowie die Volltextsuche in Audio- und Videoaufzeichnungen. Allerdings enthalten gerade technische Vorlesungen vielerlei fremdsprachliche Ausdrücke. Das stellt ein Problem für Spracherkennungssysteme dar, denn ihr akustisches Modell beschränkt sich normalerweise auf eine einzige Sprache. Ebenfalls sind fremdsprachliche Wörter oft nicht im Aussprachewörterbuch enthalten und falls doch,

so ist ihre Aussprache im Allgemeinen nicht gut repräsentiert. Das Aussprachewörterbuch ist eine der Hauptkomponenten eines Spracherkenners, es ordnet Wörtern Aussprachen zu. Beim Training bestimmt es somit, welche Merkmale den Phonemmodellen zugewiesen werden. Bei Erkennungsläufen bestimmt es, zusammen mit dem Sprachmodell, diejenigen Wörter, die mögliche Kandidaten für die Hypothesen des Spracherkenners sind. Schlecht repräsentierte Aussprachen im Wörterbuch führen daher zu zwei Arten von Fehlern, der Verschmutzung der Modelle beim Training und dem Nicht- oder Fehlerkennung von Wörtern bei Erkennungsläufen. Gründe für die Existenz schlechter Repräsentationen sind:

- Beim Hinzufügen neuer Aussprachen werden diese oft automatisch generiert. In unserem System verwenden wir z.B. Festival [1]. Die Regeln zur Ausspracheerzeugung sind sprachspezifisch und liefern für fremdsprachliche Wörter deshalb unbefriedigende Ergebnisse.
- Aufgrund der zum Teil sehr unterschiedlichen Phoneme zweier Sprachen können selbst bei manueller Erzeugung gewisse Aussprachen nicht korrekt dargestellt werden (z.B. kann der englische *TH*-Laut nur mangelhaft durch deutsche Phoneme nachgebildet werden).

Da die meisten fremdsprachlichen Wörter in unseren deutschen Vorlesungen vermutlich aus dem Englischen stammen, beschränken wir uns im Folgenden auf die Sprachen Deutsch und Englisch. Prinzipiell lassen sich die hier vorgestellten Verfahren aber auch auf weitere Sprachen erweitern.

## 1.1 Identifizierung englischer und deutscher Wörter

Um die Repräsentation englischer Wörter in unserem deutschen Aussprachewörterbuch zu verbessern, müssen wir für sie korrekte Aussprachen generieren, die auf englischen Regeln basieren und nicht auf deutschen. Hierfür müssen die englischen Wörter zuerst identifiziert werden. Zu diesem Zweck verwenden wir die freie Software Hunspell [2] mit zwei passenden Wörterbüchern, einem deutschen und einem englischen. Hunspell ist eine Rechtschreibprüfung, die z.B. von OpenOffice benutzt wird. Sie zeichnet sich dadurch aus, dass sie speziell für Sprachen mit komplexen Wortbildungsregeln entworfen wurde und daher auch Wörter in flektierter und komponierter Form erkennt. Zur Bestimmung seiner Sprachzugehörigkeit lassen wir ein Wort von beiden Hunspell-Wörterbüchern prüfen. Wenn das Wort nur dem deutschen Wörterbuch bekannt ist, klassifizieren wir es als *deutsch*. Entsprechend klassifizieren wir es als *englisch*, wenn es nur dem englischen Wörterbuch bekannt ist. Für den Fall, dass ein Wort beiden Wörterbüchern oder keinem bekannt ist, wird es in die Klassen *ambivalent* bzw. *unbekannt* eingeteilt. Beispiele für ambivalente Wörter sind „bald“, „kind“ und „man“, da sie je nach Kontext deutsch oder englisch ausgesprochen werden können. Auch englische Wörter, die sich im Deutschen etabliert haben, wie z.B. „handy“ und „service“, werden hierzu gezählt.

Wir benutzen diese Klassifizierungsmethode, um alle potentiell englischen Wörter (d.h. die Klassen *englisch* und *ambivalent*) in unserem deutschen Aussprachewörterbuch zu identifizieren und lassen für sie korrekte, den englischen Regeln folgende Aussprachen generieren. Diese englischen Aussprachen werden dann unserem deutschen Aussprachewörterbuch als neue Varianten hinzugefügt. Allerdings ist dies nicht ohne Weiteres möglich, denn die neuen Aussprachen bestehen aus englischen Phonemen, welche unserem

deutschen Spracherkennungssystem unbekannt sind. Zur Lösung dieses Problems untersuchen wir zwei mögliche Vorgehensweisen.

## 2 Abbildung von englischen Phonemen auf deutsche Phoneme

Ein naheliegender und im Grunde einfach zu realisierender Ansatz ist das Abbilden der englischen Phoneme auf deutsche. Hier stellt sich jedoch die Frage, wie eine solche Abbildungsfunktion zu wählen ist. Um möglichst gute Abbildungen zu finden, vergleichen wir manuell erstellte als auch datengetriebene, errechnete Funktionen. Für das datengetriebene Verfahren setzen wir zur objektiven Bestimmung der Ähnlichkeit von Phonemen drei verschiedene Distanzmaße ein [3]: die euklidische, die Mahalanobis- und die Kullback-Leibler-Distanz. Hierfür werden zunächst Merkmale aller deutschen und englischen Phoneme aus entsprechenden, sprecherspezifischen Trainingsdaten extrahiert und danach alle Modelle mit jeweils einer Gaußkurve trainiert. In früheren Experimenten haben wir festgestellt, dass für ein Phonemtraining zum Zweck der Distanzberechnung Gaußmixturen keine nennenswerten Vorteile gegenüber dem Training mit nur einer Gaußkurve erbringen [4, 5].

Um eine Abbildungsfunktion zu errechnen, wird für jedes englische Phonem dasjenige deutsche Phonem mit der größten Ähnlichkeit, d.h. dem kleinsten Abstand, ermittelt. Auf diese Weise erhält man drei verschiedene Abbildungsfunktionen, eine pro Distanzmaß. In unserem Fall waren diese drei berechneten Abbildungsfunktionen zu ca. 80% identisch. Sie ersetzen jedes englische Phonem durch genau ein deutsches Phonem. Unsere manuelle Funktion ersetzt jedes englische Phonem ebenfalls durch genau ein Abbild, wobei dieses Abbild jedoch statt einem einzelnen deutschen Phonem auch ein deutsches Phonempaar sein kann (so wird z.B. das englische *CH* durch das Paar *[T SCH]* ersetzt). Zusätzlich erstellen wir eine weitere, *gemischte* Abbildungsfunktion, indem wir die drei errechneten Funktionen und die manuelle Funktion kombinieren. Diese gemischte Funktion ist in Tabelle 1 dargestellt. Sie ersetzt jedes englische Phonem durch bis zu vier Alternativen, nämlich durch das ähnlichste deutsche Phonem nach jedem der verwendeten Distanzmaße und optional durch das Phonem bzw. Phonempaar aus unserer manuellen Funktion. Jede dieser Alternativen führt zu einer eigenen Aussprachevariante, die dem Wörterbuch hinzugefügt wird. Die Testergebnisse unserer Experimente mit den Abbildungsfunktionen werden in Kapitel 4.2 präsentiert.

## 3 Parallele Verwendung von englischem und deutschem akustischen Modell

Durch den Abbildungsansatz lassen sich die englischen Aussprachen nur approximieren, denn die englischen Phoneme können durch die deutschen nicht exakt nachgebildet werden, da sie sich teilweise sehr voneinander unterscheiden. So ist z.B. das deutsche Phonem *SCH* nur ein unzureichender Ersatz für die englischen Phoneme *JH* und *ZH*. Möchte man die korrekten, englischen Aussprachen ohne Verfremdung dem deutschen Aussprachewörterbuch hinzufügen, so ist dies durch die Aufnahme aller englischen Phoneme in das deutsche Spracherkennungssystem möglich. Hierzu genügt es nicht, nur die englischen Phoneme in den deutschen Phonemsatz zu integrieren, sondern es muss auch das komplette akustische Modell für Englisch aufgenommen werden, d.h. die Modelle der *Hidden-Markov-Model-Zustände* inklusive des phonetischen Entscheidungsbaumes. Somit wird das englische akustische Modell mit dem deutschen verschmolzen, hierbei werden die englischen Modelle markiert und getrennt von den deutschen gehalten.

Englisch	Deutsch	Englisch	Deutsch	Englisch	Deutsch
AA	AH, A	F	F	R	R
AE	AI, E	G	G	S	S
AH	A, ER	HH	H	SH	SCH
AO	O, AU	IH	I	T	T
AW	AU, AH	IX	EH, I	TH	TS, F, Z, S
AX	E2	IY	IE	UH	UH, UE
AXR	R, ER	JH	SCH, [T SCH], D	UW	UH, UE
AY	AI	K	K	V	V
B	B	L	L	W	U
CH	SCH, [T SCH], CH	M	M	XL	U, OH, L
D	D	N	N	XM	N, UH, M
DH	D, V	NG	NG	XN	N
EH	E	OW	OH, [O U]	Y	J, IE
ER	R, OE, OEH	OY	E, EU	Z	S, Z
EY	AEH, [EH I]	P	P	ZH	SCH

**Tabelle 1** - Gemischte Abbildungsfunktion von englischen auf deutsche Phoneme. Die deutschen Alternativen sind durch Komma getrennt, ihre Reihenfolge hat keine Bedeutung.

Zur Gewinnung des englischen akustischen Modells wird ein neues, englisches System trainiert, welches dieselbe Vorverarbeitung und dieselben Transformationsmatrizen wie unser deutsches System benutzt. Die *Linear Discriminant Analysis* (LDA) und *Semi-Tied Covariance* (STC) [6] Matrizen werden ausschließlich auf deutschen Daten trainiert, da unser Zielsystem ein deutsches ist und unser Testset überwiegend deutsche Wörter enthält. Wir haben in früheren, sprachübergreifenden Experimenten festgestellt, dass zumindest die Berechnung der LDA Matrix sowieso nicht besonders abhängig von der Zielsprache ist. Die Testergebnisse unserer Experimente mit diesem parallelen System werden in Kapitel 4.3 präsentiert.

## 4 Experimente

Die unten aufgeführten Spracherkennungsexperimente wurden mit dem *Janus Recognition Toolkit* (JRTk) durchgeführt, das gemeinsam von den Schwesterlaboratorien, den *Interactive Systems Laboratories* an der Universität Karlsruhe (TH), Deutschland und an der Carnegie Mellon University in Pittsburgh, Pennsylvania, USA entwickelt und gepflegt wird.

Alle Erkennungsläufe wurden auf deutschen, technischen Vorlesungen durchgeführt, die an der Universität Karlsruhe (TH) aufgezeichnet wurden. Das Training fand auf einem anderen Set statt, das ebenfalls aus Vorlesungen bestand.

Unser akustisches Modell verwendet 2000 Codebücher mit bis zu 64 Gaußglocken. Von dem mit 16 kHz abgetasteten Signal werden alle 10 ms neue Merkmale berechnet. Hierfür zerlegen wir das kontinuierliche Sprachsignal durch ein 16 ms Hamming-Fenster und berechnen die skalierte Mel-Frequenz *Minimum Variance Distortionless Response* (MVDR) Einhüllende [3]. Danach werden 20 Cepstralkoeffizienten, die durch cepstrale Mittelwertsubtraktion normalisiert wurden, zusammen mit ihren sieben linken und rechten Nachbarn durch eine diskrete Kosinustransformation berechnet. Schließlich wird die

Merkmalsdimension durch eine LDA auf 42 reduziert.

Das verwendete 4-Gramm Sprachmodell berücksichtigt deutsche Fernsehnachrichten, Zeitungstexte sowie manuelle Transkripte um Effekte der gesprochenen Sprache zu modellieren. Es wurde auf Vorlesungen und technische Präsentationen adaptiert.

#### 4.1 Ausgangssystem

Durch unsere in Kapitel 1.1 beschriebene Klassifizierung der Sprachzugehörigkeit kann ein Überblick über den Einfluss englischer Wörter auf die Spracherkennungsleistung gewonnen werden. Wir analysierten eine 6589 Wörter umfassende Vorlesung an der Universität Karlsruhe (TH), es ergab sich folgende Verteilung:

- 64% oder 4195 Wörter sind eindeutig *deutsch*.
- 2% oder 110 Wörter sind eindeutig *englisch*.
- 13% oder 887 Wörter konnten weder Deutsch noch Englisch zugeordnet werden und wurden als *unbekannt* eingestuft. Zu diesem Anteil gehören auch Häsitationen .
- 21% oder 1397 Wörter sind *ambivalent* und können sowohl deutsch als auch englisch sein.

Der Anteil der eindeutig englischen Wörter erscheint recht gering, ein Blick auf die individuellen Wortfehlerraten unseres Ausgangssystems in Tabelle 2 zeigt jedoch, dass diese englischen Wörter sehr viel schlechter erkannt werden als eindeutig deutsche Wörter. Der Einfluss dieser hohen englischen Fehlerrate auf die Gesamtfehlerrate von 13,8% ist nicht zu unterschätzen. Gelänge es, die englische Fehlerrate zu halbieren, so würde die Gesamtfehlerrate bereits um 0,5% absolut sinken. Die Tatsache, dass die Wortfehlerrate bei ambivalenten Wörtern dichter am deutschen als am englischen Wert liegt, lässt sich dadurch erklären, dass die meisten, aber nicht alle, dieser ambivalenten Wörter deutsch und nicht englisch ausgesprochen wurden.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	52	1	44	0
Insertions	58	9	37	2
Substitutions				
Deutsch	258	37	91	113
Englisch	7	6	8	7
Ambivalent	68	10	33	56
Unbekannt	5	3	2	4
Gesamtfehler	448	66	215	182
Wortfehlerrate	10,7%	60,0%	15,4%	20,5%

**Tabelle 2** - Fehler sortiert nach Sprachen für das Ausgangssystem.

## 4.2 Abbildung von englischen Phonemen auf deutsche Phoneme

Wie Tabelle 3 zeigt, lieferte unsere manuelle Abbildungsfunktion in Tests mit 13,2% eine leicht bessere Gesamtfehlerrate als die drei errechneten Abbildungsfunktionen. Dies liegt wahrscheinlich daran, dass die manuelle Funktion, wie in Kapitel 2 beschrieben, manche englischen Phoneme durch ein deutsches Phonempaar ersetzt statt nur durch ein einzelnes deutsches Phonem. Von den errechneten Abbildungsfunktionen lagen die Mahalanobis- und Kullback-Leibler-basierten mit 13,3% Gesamtfehlerrate gleichauf, die euklidische etwas dahinter mit 13,5%. Betrachtet man nur die englische Wortfehlerrate, so führte Kullback-Leibler mit 46,4% auch vor der manuellen Funktion. Dieser immer noch hohe Wert veranlasste uns zum Erstellen der *gemischten* Abbildungsfunktion, die auch tatsächlich die englische Wortfehlerrate auf 34,6% senken konnte und mit der im Testfeld niedrigsten Gesamtfehlerrate von 12,7% aufwarten konnte. Die Ergebnisse dieses Systems sind in Tabelle 4 zusammengefasst.

All diese Testläufe fanden auf dem ursprünglichen akustischen Modell statt. Nach einem Neutraining des akustischen Modells mit einem Wörterbuch, das aus unserer gemischten Abbildungsfunktion resultierte, sank die englische Fehlerrate nochmals auf 27,3%. Allerdings stellten wir fest, dass die Gesamtfehlerrate wieder leicht anstieg, da nun deutsche Wörter schlechter erkannt wurden. Als Grund dafür vermuten wir eine Verringerung der akustischen Eindeutigkeit der deutschen Phoneme, da diese beim Neutraining durch englische Aussprachen verschmutzt werden.

Sprache	Wortfehlerrate				
	Gesamt	Deutsch	Englisch	Ambivalent	Unbekannt
Manuell	13,2%	11,2%	47,3%	14,7%	16,4%
Euklidisch	13,5%	11,3%	51,8%	15,1%	16,7%
Mahalanobis	13,3%	11,2%	48,2%	15,0%	16,6%
Kullback-Leibler	13,3%	11,1%	46,4%	15,0%	17,4%
Gemischt	12,7%	11,1%	34,6%	13,8%	16,1%
Gemischt (neu trainiert)	13,0%	11,7%	27,3%	14,1%	15,3%

**Tabelle 3** - Wortfehlerraten der verschiedenen Abbildungssysteme im Vergleich.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	50	0	38	0
Insertions	54	7	35	0
Substitutions				
Deutsch	260	18	76	84
Englisch	14	5	6	8
Ambivalent	81	8	37	47
Unbekannt	5	0	1	4
Gesamtfehler	464	38	193	143
Wortfehlerrate	11,1%	34,6%	13,8%	16,1%

**Tabelle 4** - Fehler sortiert nach Sprachen für das gemischte Abbildungssystem.

### 4.3 Parallele Verwendung von englischem und deutschem akustischen Modell

Das Parallelsystem konnte mit 26,4% Wortfehlerrate das insgesamt beste Ergebnis für englische Wörter erzielen. Allerdings produzierte es etwas mehr Fehler auf deutschen Wörtern als die anderen Systeme, wodurch die Gesamtfehlerrate bei 13,4% liegt. Die Fehlerraten der anderen Wortklassen wurden im Vergleich zum Ausgangssystem zwar reduziert, jedoch erreichte das gemischte Abbildungssystem hier niedrigere Werte. Die Ergebnisse des Parallelsystems werden in Tabelle 5 zusammengefasst.

Sprache	Deutsch	Englisch	Ambivalent	Unbekannt
Deletions	54	0	39	0
Insertions	61	10	48	0
Substitutions				
Deutsch	266	11	74	81
Englisch	19	3	5	14
Ambivalent	74	5	37	67
Unbekannt	6	0	2	6
Gesamtfehler	480	29	205	168
Wortfehlerrate	11,4%	26,4%	14,7%	18,9%

**Tabelle 5** - Fehler sortiert nach Sprachen für das Parallelsystem.

## 5 Zusammenfassung und Ausblick

Beim Vergleichen der Ergebnisse in Tabelle 6 sehen wir, dass die beiden vorgestellten Methoden die insgesamt Erkennungsleistung des Spracherkennungssystems verbessern. Während der parallele Ansatz die niedrigste Fehlerrate für rein englische Wörter hat, ist die gesamte Wortfehlerrate beim Abbildungsansatz am besten. Interessanterweise verschlechtern beide Ansätze die Erkennungsleistung auf deutschen Wörtern. Der Vergleich der Tabellen 2, 4 und 5 zeigt, dass besonders die Ersetzung von englischen Wörtern durch deutsche Wörter abnimmt, während gleichzeitig die Ersetzung von deutschen Wörtern durch englische zunimmt.

Tabelle 7 zeigt, dass beide Ansätze auch eine längere Laufzeit als das Grundsystem benötigen. Hauptursache dafür ist das gewachsene Suchvokabular. Beim Abbildungssystem wurden durch die gemischte Abbildungsfunktion sehr viele neue Aussprachevarianten in das Wörterbuch eingeführt. Neben der Auswirkung auf die Laufzeit haben diese Varianten einen weiteren, negativen Effekt. Aus der Literatur [7] ist bekannt, dass die Präsenz von vielen Aussprachevarianten der akustischen Eindeutigkeit eines Wortes abträglich ist und aus Sicht des Spracherkenners die Gefahr der Verwechslung mit anderen Wörtern erhöht.

Wenn man bedenkt, dass ein vergleichbares englisches Spracherkennungssystem auf rein englischen Daten eine Wortfehlerrate von ca. 10% hat, wird offensichtlich, dass hier noch weiteres Optimierungspotenzial besteht. Um die vorgestellten Methoden weiter zu entwickeln, werden wir daran arbeiten, präzisere und bessere Verfahren für die Identifizierung deutscher und englischer Wörter zu finden. Durch die Betrachtung des Wortkontexts könnten beispielsweise die bisher als ambivalent eingestuft Wörter eindeutig als deutsch oder englisch klassifiziert werden. Außerdem haben wir uns bisher nur auf die Akustik konzentriert und wollen deshalb in Zukunft auch das Sprachmodell untersuchen.

Sprache	Wortfehlerrate				
	Gesamt	Deutsch	Englisch	Ambivalent	Unbekannt
Ausgangssystem	13,8%	10,7%	60,0%	15,4%	20,5%
Abbildung (gemischt)	12,7%	11,1%	34,6%	13,8%	16,1%
Parallel	13,4%	11,4%	26,4%	14,7%	18,9%

**Tabelle 6** - Wortfehlerrate nach Sprache für die unterschiedlichen Ansätze.

System	Ausgangssystem	Abbildung (gemischt)	Parallel
Echtzeitfaktor	1,50	1,79	1,62
Wörterbuchgröße	63510	97247	70550

**Tabelle 7** - Laufzeit und Anzahl der Aussprachen im Aussprachewörterbuch.

## 6 Danksagung

Diese Arbeit wurde innerhalb des Quaero Projektes realisiert. Quaero wird durch OSEO finanziert, der Französischen Agentur für Innovation.

## Literatur

- [1] A. W. Black and P. A. Taylor *The Festival Speech Synthesis System: System documentation*. Human Communication Research Centre, University of Edinburgh, 1997. <http://cstr.ed.ac.uk/projects/festival/>
- [2] Hunspell: open source spell checking, stemming, morphological analysis and generation. <http://hunspell.sourceforge.net>
- [3] M. Wölfel and J.W. McDonough *Distant Speech Recognition*. John Wiley & Sons, 2008.
- [4] D. Heger *Speech Feature Enhancement using Particle Filters with Class-Based Phoneme Models*. Studienarbeit, Universität Karlsruhe (TH), 2007
- [5] S. Stüker *Integrating Thai Grapheme based Acoustic Models into the ML-Mix framework - for language independent and cross-language ASR*. Proceedings of the first International Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU), 2008
- [6] M.J.F. Gales *Semi-Tied Covariance Matrices for Hidden Markov Models*. Cambridge University, Engineering Department, 1998
- [7] M. Saraclar and S. Khudanpur *Pronunciation Ambiguity versus Pronunciation Variability in Speech Recognition*. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000