Articulatory Features for Conversational Speech Recognition

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe (TH)
genehmigte

Dissertation

von
Florian Metze
aus München

Tag der mündlichen Prüfung:	09. Dezember 2005
Erster Gutachter:	Prof. Dr. Alexander Waibel
Zweiter Gutacher:	Prof. Dr. Jürgen Beyerer
Externer Gutacher:	Dr. Lori Lamel (HdR)

To perform the art, to know the science.

Abstract

While the overall performance of speech recognition systems continues to improve, they still show a dramatic increase in word error rate when tested on different speaking styles, i.e. when speakers for example want to make an important point during a meeting and change from sloppy speech to clear speech. Today's speech recognizers are therefore not robust with respect to speaking style, although "conversational" speech, as present in the "Meeting" task, contains several, distinctly different, speaking styles.

Therefore, methods have to be developed that allow adapting systems to an individual speaker and his or her speaking styles. The approach presented in this thesis models important phonetic distinctions in speech better than phone based systems, and is based on detectors for phonologically distinctive "articulatory features" such as ROUNDED or VOICED. These properties can be identified robustly in speech and can be used to discriminate between words, even when these have become confusable, because the phone based models are generally mis-matched due to differing speaking styles.

This thesis revisits how human speakers contrast these broad, phonological classes when making distinctions in clear speech, shows how these classes can be detected in the acoustic signal and presents an algorithm that allows to combine articulatory features with an existing state-of-the-art recognizer in a multi-stream set-up. The needed feature stream weights are automatically and discriminatively learned on adaptation data, which is more versatile and can be handled more efficiently than previous approaches.

This thesis therefore presents a new acoustic model for automatic speech recognition, in which phone and feature models are combined with a discriminative approach, so that an existing baseline system is improved. This multi-stream model approach captures phonetic knowledge about speech production and perception differently than a purely phone based system.

We evaluated this approach on the multi-lingual "GlobalPhone" task and on conversational speech, i.e. the English Spontaneous Scheduling Task (ESST) and RT-04S "Meeting" data, which is one of the most difficult tasks in Automatic Speech Recognition today. The algorithm is applied to generate context-independent and context-dependent combination weights. Improvements of up to 20% for the case of speaker specific adaptation outperform conventional adaptation methods.

Zusammenfassung

Obwohl die durchschnittliche Erkennungsleistung steigt, leiden selbst moderne Spracherkennungssysteme noch unter sehr schlechten Erkennungsraten, wenn sie auf unterschiedlichen Sprechstilen getestet werden. Diese treten beispielsweise auf, wenn ein Sprecher in einer Besprechung einen wichtigen Punkt präsentiert und von spontaner Sprache in einen deutlichen, besonders klaren Sprechstil wechselt. Die Erkennungsleistung von Spracherkennungssystemen wird unter diesen Bedingungen sinken, da die Erkenner nicht gegen Änderungen des Sprechstiles robust sind.

Es müssen daher Methoden entwickelt werden, die es erlauben, einen Erkenner besser auf einen Sprecher und seine verschiedenen Sprechstile, wie sie z.B. im NIST RT-04S "Meeting-Korpus" vorliegen, zu adaptieren. Der hier vorgestellte Ansatz erlaubt es, wichtige phonetische Unterscheidungen, nämlich "artikulatorische Merkmale" wie GERUNDET oder STIMMHAFT, besser als ein herkömmlicher phonem-basierter Ansatz zu modellieren. Diese "Features" können robust erkannt werden und können verwendet werden, um die Unterscheidung zwischen Worthypothesen, die durch die aufgrund des geänderten Sprechstils schlecht passenden Phonem-Modelle ähnlich geworden sind, zu verbessern.

Diese Arbeit präsentiert Beispiele, wie Sprecher in deutlichem Sprechstil diese phonologischen Merkmale verwenden, um wichtige Unterscheidungen zu betonen, sie zeigt, wie diese im akustischen Signal detektiert werden können und präsentiert einen Algorithmus, um die für die Kombination dieser komplementären distinktiven Merkmale in einem Multi-Stream-Ansatz benötigten Gewichte auf diskriminative Weise automatisch zu bestimmen. Die Optimierung eines Maximum-Mutual-Information Kriteriums erlaubt eine effizientere Modellierung und flexiblere Kombination phonetischer Information als bisherige Ansätze.

Diese Arbeit stellt ein neuartiges akustisches Modell für die Erkennung von Sprache aus Dialogen und Gesprächen vor, welches ein diskriminatives Verfahren einsetzt, um herkömmliche Phonemmodelle bestmöglich mit Featuremodellen zu kombinieren und die Erkennungsleistung eines bestehenden Systems zu verbessern. Dabei wird phonetisches Wissen über die Produktion und Perzeption von Sprache grundlegend anders behandelt, als in einem rein phonem-basierten Ansatz.

Der vorgestellte Ansatz wird sowohl auf dem multi-lingualen “GlobalPhone” Korpus als auch auf Spontansprache, nämlich dem English Spontaneous Scheduling Task (ESST, Verbmobil-II) und dem RT-04S “Meeting” Korpus evaluiert. Dieser gilt als einer der derzeit interessantesten und schwierigsten Korpora für die automatische Spracherkennung. Der Algorithmus wird verwendet, um auf Adaptionsdaten kontext-unabhängige und kontext-abhängige Gewichte zu generieren. Für den Fall der sprecher-abhängigen Adaption wird die Fehlerrate um bis zu 20% relativ reduziert, was die Leistung konventioneller Maximum-Likelihood basierter Verfahren deutlich übertrifft.

Acknowledgments

First, I would like to thank my thesis adviser, Professor Dr. Alexander Waibel, for giving me the opportunity to work at the Interactive Systems Laboratories and making them an exciting, challenging, and fun place to be. I was never bored a minute while working on several international research projects over the course of time, be they related to speech-to-speech translation, Italian fruit, or, finally, FAME!

Although an individual work, this thesis profited greatly from discussions with my thesis committee, Professor Dr. Jürgen Beyerer and particularly Dr. Lori Lamel (HDR). Thanks are also due to Tanja and Sebastian for sharing with me the passion for feature based acoustic modeling and discussing with me and to Professor Mari Ostendorf for kindly proof-reading my thesis.

I would not have come to Karlsruhe, had it not been for my teachers at LMU München, University of St. Andrews and USH Strasbourg, who got me interested in the combination of science and language.

I learned a great deal about good software engineering from reading the JRTk source code, thanks to all who put together this fine speech recognition toolkit. Many thanks also to my first office mate Thomas for answering my first ton of questions, Ivica and ThomaS also frequently shared their insights with me. Formed together with my long-term office mates Hagen and Christian, the "Ibis gang" was known to be sticking together at all times and was feared for its uncompromising use of computing resources during evaluation campaigns. I owe thanks to all members of the ISL, working here would have been much harder without the support of Silke, Annette, Margit, Frank, and Dan. When at CMU, Tanja, Susi, Hua, and Kornel (among many more) offered friendship and hospitality, thank you!

Finally, I want to thank my family and friends, who allowed me to and encouraged me to spend time and effort on this thesis and supported me during rough times. This thesis is dedicated to my loving parents, Hedwig and Wolfgang, my sister Gudula, who already holds her own Ph.D., and Marietta: thank you for reminding me from time to time that there is more to life then talking with machines and working toward usable automatic speech recognition!

The peril of acknowledgments is, of course, to leave someone out unintentionally; to all of you: I owe you one!

Contents

List of Figures

List of Tables

List of Abbreviations

AF	Articulatory Feature
AM	Acoustic Model
ASR	Automatic Speech Recognition
BN	Broadcast News
CA	Contrastive Attribute
CD	Context-Dependent (e.g. states of an HMM)
CFG	Context-Free Grammar
CI	Context-Independent (e.g. states of an HMM)
CLP	Consensus Lattice Processing
C-MLLR	Constrained MLLR, i.e. FSA
CMS	Cepstral Mean Subtraction
CMU	Carnegie Mellon University
CNC	Confusion Network Combination
CSR	Continuous Speech Recognition (CSR corpus, i.e. WSJ-0)
CTS	Conversational Telephony Speech (i.e. SWB)
CVN	Cepstral Variance Normalization
DBN	Dynamic Bayesian Network
DMC	Discriminative Model Combination
EM	Expectation Maximization
ESST	English Spontaneous Scheduling Task (i.e. Verbmobil)
FSA	Feature Space Adaptation, i.e. C-MLLR
FSA-SAT	Feature-Space Adaptation Speaker-Adaptive Training
G	Global (i.e. state independent stream weights)
GMM	Gaussian Mixture Model
GP	GlobalPhone
HMM	Hidden Markov Model
HSC	Hyper-articulated Speech Corpus
ICSI	International Computer Science Institute
IHM	Individual Headset Microphone
IPM	Individual Personal Microphone
ISL	Interactive Systems Laboratories

LDC	Linguistic Data Consortium
LID	Language Identification
LM	Language Model
LVCSR	Large Vocabulary Conversational Speech Recognition
MAP	Maximum A Posteriori
MCE	Minimum Classification Error
MDM	Multiple Distant Microphones
ME	Maximum Entropy
MFCC	Mel-Filtered Cepstral Coefficients
ML(E)	Maximum Likelihood (Estimation)
MLLR	Maximum Likelihood Linear Regression
MMI(E)	Maximum Mutual Information (Estimation)
MPE	Minimum Phone Error
MWE	Minimum Word Error
NIST	National Institute of Standards and Technology
OOV	Out Of Vocabulary
PDF	Probability Density Function
RT	Rich Transcription, e.g. RT-04S (Spring 2004) evaluation
RTF	Real Time Factor
SD	State dependent (stream weights)
SDM	Single Distant Microphone
STC	Semi-Tied Covariances
STT	Speech-To-Text (i.e. ASR)
SWB	Switchboard (i.e. CTS)
TTS	Text-To-Speech (i.e. speech synthesis)
VTLN	Vocal Tract Length Normalization
WA	Word Accuracy
WER	Word Error Rate
WSJ	Wall Street Journal

List of Symbols

O	Time sequence of observation vectors
o, o_t	Observation vector (at a given time)
α	VTLN parameter
ω	Frequency value
W	Word sequence
w	Word
S	State sequence
s, s_t	Single state (at a given time)
$\mathcal{R}$	States in reference sequence
$\mathcal{S}$	States in entire search space
g	Score function
$\mathcal{L}$	Likelihood value
$p(W\|O)$	Posterior probability (function)
$p(O\|W)$	Likelihood (function)
P	Probability (value)
α_t	Forward probabilities
β_t	Backward probabilites
γ_t	State a-posteriori probabilites
Φ	Accumulated statistics
ϵ	Training parameter ("step size")
η	Smoothing parameter
F	MMI optimization criterion
Λ	Stream weights
λ_i	Weight for stream i
Ψ	Acoustic model parameters
Ψ_i	Acoustic model parameters in stream i
Ξ	Language model parameters
μ	Mean of Gaussian mixture component
Σ	Covariance matrix of Gaussian mixture component
w	Distribution weight of Gaussian mixture component

Chapter 1

Introduction

This thesis deals with an Automatic Speech Recognition ("ASR") system using acoustic models based on both sub-phonetic units and broad, phonological classes motivated by articulatory properties as output densities in a Hidden-Markov-Model framework. The aim of this work is to improve speech recognition performance by using non-phonetic units as a basis for discrimination between words whenever possible. For example, the difference between the words *bit* and *pit* would not be determined by looking for the sounds /b/ and /p/, but by calculating the probability that the first sound of the word being VOICED in the case of *bit* and UNVOICED in the case of *pit*, which is a more generic decision problem and should allow for more robust recognition.

In contrast to conventional ASR systems, the acoustic model used in this work is not exclusively trained on a phonetic partition of the training data, i.e. on time alignments on the (sub-)phonetic level, but instead "conventional" probability density distributions are combined with more generic distributions based on broad classes such as VOICED, FRICATIVE, or ROUNDED. During recognition, the combination is achieved in a stream architecture on the log-likelihood (or acoustic score) level.

To automatically learn from data the features which can be used to discriminate between given contexts, the stream weights needed for the model combination are trained using discriminative approaches on training or adaptation data. This thesis compares two training approaches and presents a new scheme to train weights using the Maximum Mutual Information (MMI) criterion, which allows faster training on conversational speech task using lattices when compared to a previously used Minimum Word Error (MWE) criterion. The use of broad phonetic classes to distinguish between sounds allows for more parameter sharing when compared to sub-phone based models, leading to greater robustness of the resulting recognizer with respect to pronunciation variability.

A major advantage of this approach over other speech recognizers based on Articulatory Features (AFs) is that the presented architecture can be integrated easily into an existing state-of-the-art baseline system, as was shown by our experiments on the RT-04S "Meeting" task. Most approaches based on articulatory properties alone are computationally tractable only on small tasks or in a rescoring step, which is impractical in many applications. Our approach adds only little computational complexity and can be integrated in a single decoding pass.

Also, this work demonstrates that the new approach performs better for speaker adaptation than standard approaches based on the Maximum Likelihood principle, which shows that a structured approach based on articulatory properties indeed captures speaker variability better than a "beads-on-a-string" approach [Ost99]. Moreover, the feature based approach improves the recognition of "hyper-articulated" or "clear" speech, which occurs whenever speakers want to particularly emphasize some part of their speech. As recognizers are usually trained on "normal" speech, these parts of speech are usually recognized with below-average accuracy [Sol05], which is counter-intuitive to human speakers, as sentences spoken "clearly" are more intelligible than those spoken "conversationally" under nearly all conditions [PDB85, UCB+96, PUB94].

Other aspects investigated in this work are the trans- and multi-lingual properties of articulatory features, which as detectors for universal phonological features can be shared and re-used across languages.

In contrast to other work on speech recognition based on articulatory features, the approach presented in this work does not explicitly model trajectories of a real or assumed articulator. Also, in our terminology, the term "Articulatory Feature" does not refer to characteristic properties of the speech signal, found only at a specific point in time, as is the case in "landmark-based" automatic speech recognition and similar approaches. Instead, we use the term "Articulatory Features" to describe an acoustic model, in which the likelihood of a specific phone is expressed not as the result of the evaluation of a single phone model's probability density function, but as a combination of likelihoods computed for complementary linguistic features such as VOICED, possibly also in combination with conventional phone models. Apart from the feature inventory and the feature to phone mapping, no further expert knowledge is used for the construction of the speech recognizer and no claim as to the relation of feature values computed with actual articulatory processes is being made. The term "Articulatory" reflects the observation that most of the properties used in linguistic feature theory, which forms the basis for our work, is in fact based on articulation rather than perception.

The remainder of this chapter covers in more detail the (expected) potential of articulatory features in Automatic Speech Recognition, presents the goals and contributions of this work, and gives an overview of the structure of this thesis.

1.1 Motivation

Speech recognition has advanced considerably since the first machines which could convert human speech into symbolic form (i.e. transcribe it) were conceived in the 1950s [JR05]. Today's state-of-the-art text-to-speech systems are based on Hidden Markov Models (HMMs), for which efficient training and evaluation techniques are known.

Still, humans are much better than machines at deciphering speech [Moo03] under changing acoustic conditions, in unknown domains, and at describing somebody's speech characteristics as "sloppy", "nasal" or similar, which allows them to rapidly adjust to a particular speaking style. In many important domains, this results in a human speech transcription performance still unmatched by machines, particularly if the speaker's original goal was human-to-human, not human-to-machine communication.

Leaving aside the domain of the speech, which may also not be available to a machine, phonetics and phonology categorize human speech and describe the process of speech understanding in humans. Today's automatic speech recognition systems use phones, a phonetic description unit, as atoms of the speech model. These, however, are a shorthand notation for a bundle of phonological properties such as voicedness or lip rounding, which characterize a certain region of speech. As many of these categories are related to articulation, they are often referred to as "Articulatory Features", keeping in mind that the physical processes of articulation, i.e. the movements of the vocal cords etc., are *not* observable to a human listener or an automatic speech recognizer operating on audio information only. In this sense, the term "Articulatory Features" describes classes of speech sounds (i.e. voiced sounds and unvoiced sounds), whose names are based on articulatory categories, although the partitioning is based on acoustic observations only. This means the features are a perceptual categorization and should therefore be helpful in the discrimination of speech sounds.

Particularly for spontaneous and conversational speech, it is not generally possible to identify discrete, clearly separated units in human speech, instead it is only possible to mark transient phonetic events, which can be aggregated into canonical phones using knowledge of the language. While there is a long history of studies on speech articulation [Fan60, Fla65, Ste98, RKT03], the focus of this work has usually been on a general understanding of the articulation process in humans, less so on the usefulness of AFs for Automatic Speech Recognition. Combining insights into human articulatory processes and speech understanding in humans with standard HMM based speech recognition system however is interesting for the following reasons:

- Existing, efficient, and well-understood tools can be re-used.

- AFs provide more, and different, degrees of freedom than standard features, but they can still be formulated in a probabilistic framework.

They might therefore complement existing acoustic models, thus directly leading to improved performance.

- AFs allow including linguistic knowledge differently, particularly for non-native speech and hyper-articulated speech; cases which represent a particularly challenging speech recognition tasks.
- AFs should be portable across languages as they are modelled on supposedly universal, i.e. cross-lingual, properties.
- Speech recognition using AFs was shown to be more robust against noise than a standard recognizer trained on the same data [Kir99].
- By using AFs in a stream setup, it could be possible to amend some aspects of the "beads-on-a-string" [Ost99] model while still retaining a computationally efficient system structure.
- AFs might be particularly useful for adaptation, particularly to speaker and speaking style, as they have been used in respective verification [LMSK05] or analysis [Esk93] tasks.

Therefore, the approach taken in this work is to improve an existing speech recognizer based on phones and HMMs by adding a description of speech based on articulatory features, while still retaining the HMM based speech model for efficiency. While we call our approach "Articulatory", it does not rely on the actual observation of articulatory parameters and does not assume a particular articulatory process to occur for the production of a particular sound as in a speech production model. Instead, we are using articulatory categories to name and classify acoustic or auditory targets for speech sounds, which is motivated by the findings in [GEWB^{+}99]. "Lexical access" (or speech-to-text) can then be described as the identification, as a perceptive category, of articulation-inspired binary distinctive features, which suffice to discriminate words [Ste02].

1.2 Goals

The goal of this work is to improve an existing state-of-the-art ASR engine, i.e. existing efficient training and decoding algorithms should be re-used in order to avoid extra complexity and to ensure that the resulting recognizer can be used in today's ASR framework. It is therefore not the goal of this work to build a system based solely on AFs or do "recognition through synthesis", i.e. retrieve the actual movements of the articulators from speech data, as these approaches tend to have prohibitively high computational complexity.

Instead, we show how to adapt an existing, general recognizer to new conditions (speaking styles) in the "Meeting" domain using articulatory features. Also, we present results on the possibility of detecting articulatory properties from speech and on multi-lingual data, i.e. in a language different from the training data, and hyper-articulated speech as case studies of further applications of this approach.

Our experiments are conducted on English data from the following tasks:

- Hyper-articulated data ("HSC")
- Multi-lingual data (GlobalPhone, "GP")
- English Spontaneous Scheduling Task data ("ESST")
- "Meeting" data (RT-04S)

1.3 Outline

This work can be divided into three parts. The first part comprises *Chapters 2 to 4* and presents an introduction to phonetics, speech production, fundamentals of current state-of-the-art ASR systems, and gives an overview on speech recognition based on articulatory features. The second part, *Chapters 5 to 7*, presents our results on detection of articulatory features from speech and develop the approach to discriminative combination of knowledge sources used in the ASR experiments on several tasks, presented in the third part, *Chapters 8 to 11*.

More specifically, *Chapter 2* introduces basic concepts of phonology and phonetics and the underlying articulatory properties. It discusses multilingual properties of phones and articulatory features and present the differences occurring between different speaking styles, motivating the use of "articulatory features" in ASR research. In *Chapter 3*, we review fundamental properties of mainstream automatic speech recognition, as far as they are important in the context of this work.

Chapter 4 discusses other relevant work in the fields of large vocabulary conversational speech recognition (LVCSR), using articulatory properties for speech recognition, and approaches to the combination of several information sources for speech recognition.

Chapter 5 describes how we detect articulatory properties from the speech signal. In this work, we do not measure actual movements of the human articulatory apparatus, instead we build models ("detectors") on canonical articulatory properties of the input speech. We show that the articulatory properties used in this work can indeed serve to improve discrimination by building a combined phone and feature based speech recognizer, as the changes predicted when altering the speaking style can be modelled by detectors for articulatory properties.

Chapter 6 presents our stream architecture, which integrates these "feature detectors" with standard context-dependent acoustic models in an HMM based recognizer.

Chapter 7 develops the theory behind the discriminative approach to model combination developed in this work and investigates two different criteria ("DMC" and "MMI") for model optimization. It also introduces global and context-dependent stream weights.

The following chapters present our large-vocabulary experiments on estimation of articulatory features from the speech signal, the combination of classifiers in our stream architecture, and the discriminative estimation of the stream weights. Results are presented on multi-lingual data in *Chapter 8*, spontaneous speech data vs. read speech data in *Chapter 9*, and conversational "Meeting" speech in *Chapter 10*.

This thesis rounds up with a look at how articulatory features can further improve the robustness of speech recognition systems: *Chapter 11* presents results on hyper-articulated speech. This is particularly important in conversational speech, as people want to "stress" important information by altering the way they speak, i.e. they are speaking very clearly. Experiments however show that this may result in a degradation of speech-to-text performance, which can be alleviated by using articulatory features.

Chapter 12 presents a summary and conclusions, after which *Appendix A* shows the full derivation of the new discriminative stream weight estimation scheme, *Appendix B* lists details of the systems used in our experiments and *Appendix C* lists the weights computed with the training approaches discussed in this work.

Chapter 2

Human Speech Production

This chapter presents descriptions of speech at different levels of abstraction, as needed for research on articulatory features for speech recognition. We set out with a brief introduction to the human speech production process, then describe the role of articulatory features in phonetics and phonology, and finally introduce multi-linguality and language independence, hyper-articulation, and sloppy or conversational speech.

2.1 The Phonatory Apparatus

The production of human speech is mainly based upon the modification of an egressive air stream by the articulators in the human vocal tract. Even though different languages can exhibit vastly different sounds, the overwhelming majority of sounds can be described sufficiently enough by marking very few parameters only, as the phonation is limited by the anatomical properties of the speaker. By "sufficiently enough" we mean that, although in spoken speech no two sounds, even when produced by the same speaker, will ever be strictly identical, the intended meaning in the speaker's language will be evident by looking at very few parameters. In other words, *phonological* knowledge helps to describe *phonetic* events with only a few parameters: while *phonetics* deals with how speech sounds are actually produced, transmitted and received in actual spoken language, *phonology* deals specifically with the ways those sounds are organized into the individual languages, hence dealing with abstractions on a virtual basis. The term "articulatory features" strictly speaking is a phonetic term, but its interpretation requires phonological knowledge, too, to be useful in practice.

The goal of this section is to give a *functional* overview of basic phonatory processes as they occur in English and most other languages. Other languages exhibit different, but mostly similar, properties, which will not be discussed here. In order to understand how humans produce speech sounds, it is necessary to identify the essential components of the speech production process and describe how they work. As this section covers only these topics in articulatory phonetics which are relevant to understand this work, more detailed information about general phonetics is available for example in [Lad82, Cat77, CY95]. More information about acoustic phonetics is available in [Ste98]. A description of articulatory processes can be found in [Lav94, Per97].

The production of speech sounds in humans involves three major processes: the *air stream*, the *phonation*, and the configuration of the vocal tract (*oro-nasal* process). Fant's source filter model [Fan60] interprets these processes as a system of linear, time shift invariant components. Figure 2.1 shows a sagittal view of the human head with the organs used for speech production while Figure 2.2 shows a functional view of the source-filter model.

The Air stream process describes how sounds are produced and manipulated by the source of air. The *pulmonic egressive* mechanism is based on the air being exhaled from the lungs while the *pulmonic ingressive* mechanism produces sounds while *inhaling* air. Ingressive sounds however are rather rare. Besides these pulmonic sounds, a closure of the glottis leads to the so-called *glottal* air stream mechanism. There are *ejective* and *implosive* glottal sounds, depending on whether the air is directly pushed outward or if the glottis is lowered. A special sound is the glottal stop produced by trapping of air by the glottis.

The Phonation process occurs in the vocal chords. *Voiced* sounds are produced by narrowing the vocal chords when air passes through them. The Bernoulli effect leads to a fast cycle of opening and closing of the glottis, which produces a strong modulation of the air stream. Depending on the length of the vocal chords, the frequency of this process can be in the range of 120-230 Hz. An open glottis leads to *unvoiced* sounds. In that case, air passes through the glottis without obstruction so that the air stream is continuous.

The Oro-nasal process: from a technical point of view, the vocal tract can be described as a system of cavities. The major components of the vocal tract are illustrated in Figure 2.1. The vocal tract consists of three cavities: the *oral* cavity, the *nasal* cavity, and the *pharyngeal* cavity. These components provide a mechanism for producing different speech sounds by obstructing the air stream or by changing the frequency spectrum. Several articulators can be moved in order to change the vocal tract characteristic.

The sounds therefore depend on the air stream, the phonation, and on how this signal is being modified, e.g. on the place of the modifiers.

Cine-radiographic (X-ray) films of the speech organs in action show that they are in continuous fluent motion during speaking [Per69, SO72,

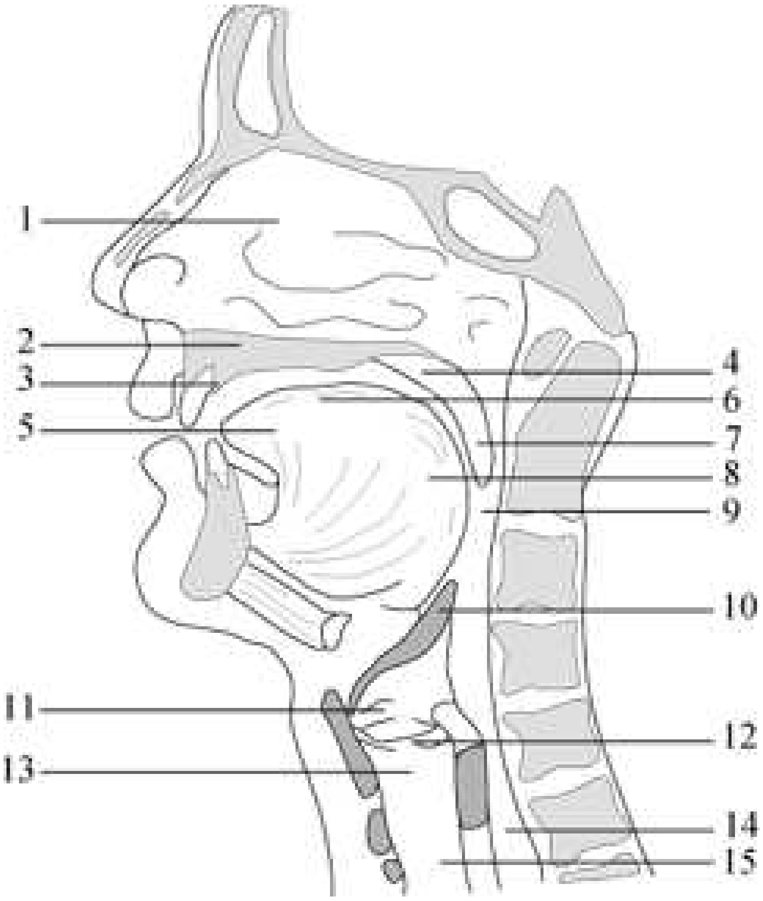


Figure 2.1: Organs of human speech production [Lem99]: (1) nasal cavity, (2) hard palate, (3) alveolar ridge, (4) soft palate (velum), (5) tip of the tongue (apex), (6) dorsum, (7) uvula, (8) radix, (9) pharynx, (10) epiglottis, (11) false vocal cords, (12) vocal cords, (13) larynx, (14) esophagus, and (15) trachea.

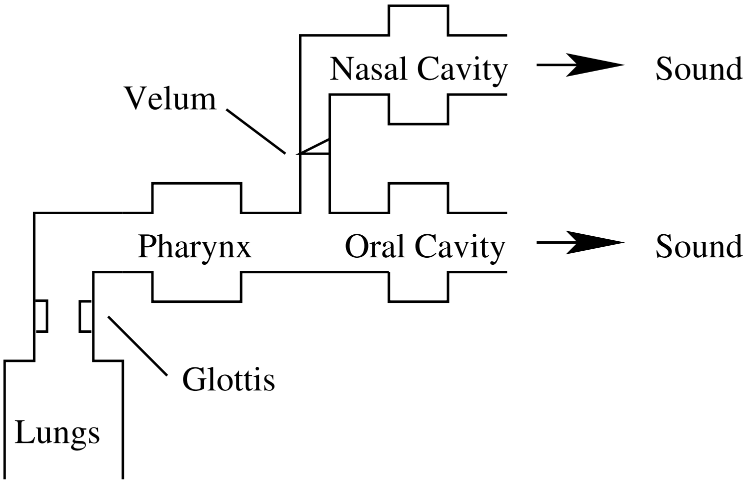


Figure 2.2: Vocal tract as a system of cavities [Sol05]: lungs and glottis are responsible for the air stream process, phonation occurs in the glottis, the resulting sound is modified primarily in the oral and nasal cavities.

MVBT95]. The same can be conjectured when looking at the spectrogram representation of speech. The patterns are changing constantly and clear-cut boundaries between sounds can only be identified for a few cases. Extra knowledge of the underlying language is needed to determine which part of the articulatory process is significant, i.e. carrying a meaning, and which is simply due to the "laziness" of the speaker.

2.2 Distinctive Features

The description of the human phonatory apparatus in the previous section already allows guessing which "features" can be used to describe speech production and speech sounds: the behavior of the vocal cords for example determines if a sound is "voiced" or "unvoiced", while the velum makes it possible to discriminate between "nasal" and "non-nasal" sounds. The configuration of the oral cavity also influences the sound produced.

One of the aims of feature theory is to set up a universal inventory of "distinctive features" (i.e. phonetic or phonological properties) which is sufficient to characterize all sounds in all languages and which permits deriving phoneme systems (i.e. symbolic descriptions) for all the languages in the world. Several feature systems have been proposed over time [JFH52, CH68, Lad82], using a mixture of criteria and approaching feature theory from a range of angles, for example from articulatory, acoustic, or auditory perspectives. This is necessary, as articulation, acoustics and perception all contribute to the transmission of information and therefore it is sensible to integrate them all into one model. As a short-hand notation for a certain combination of features, “phonemes” are used to describe a certain combination of distinctive features, which usually occur together in a specific language. A list of features is presented in the next section together with a categorization of phonemes into these features.

2.3 Phonetic Description of Speech

Linguistic analysis of a language’s vocabulary and its spoken speech representation allows determining which sounds need to be distinguished in a specific language (*phonemes*), because they serve to distinguish between words. Phonetics describes the actual realization of phonemes, and actual speech sounds are called *phones*. If two sounds are phonetically different, i.e. they are produced by different configurations of the vocal tract, but the distinction does not carry lexical information, these sounds are called *allophones*. Germans for example have two different ways of producing the /r/ phoneme, the [r] (alveolar trill) and [ʀ] (uvular trill) sounds, where the preference depends on the dialect of the speaker.

This linguistic knowledge of the underlying language permits segmenting speech by identifying points where linguistically significant changes occur. The existence of such a segmentation is the base of current phonological analysis. It is assumed that every segment has an *articulatory target*, which describes the configuration of the vocal tract and organs that are representative for the described segment and sound respectively. Usually the involved articulators make a continuous movement from and to the target during the speech production. And in some instances the target might be held for a certain amount of time. The transition phase between targets is influenced by *coarticulation*, which can span several sounds. Heavy coarticulation occurs in spontaneous speech and can make the identification of distinct sounds very difficult.

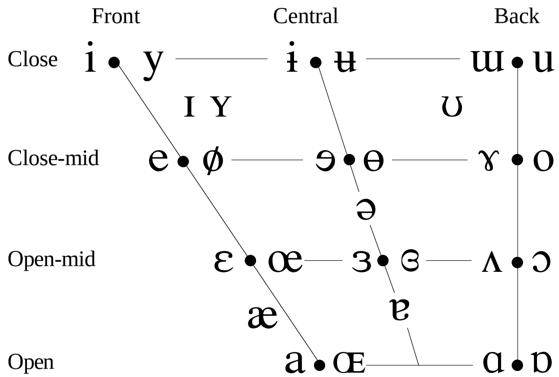
The International Phonetic Alphabet IPA [Int99, Hie93] has been created to describe and categorize the speech segments or sounds occurring in any language. A symbol is created as a short-hand notation for a specific *feature bundle*, i.e. a configuration of the articulators, if the resulting phone has phonemic value in a language. Diacritics serve to mark minor fəˈnɛtɪk variations, which are of interest in specialized cases only.

For the description of the above segments IPA heavily relies on the distinction between vowels and consonants. Speech involves consecutive widening and narrowing of the vocal tract. The openings are used to define syllables and act as the *nucleus* of the syllable. Segments that involve a narrow or closed vocal tract are called *consonants*. Sounds with a wide vocal tract in which the air flows largely uninhibited carry the terminus *vowel*. Because of this general difference between vowels and consonants IPA has decided

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Post alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ	▨	ʔ ▨
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ʀ		▨
Tap or Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative	▨	▨		ɬ ɮ						▨	▨
Approximant		ʋ		ɹ		ɻ	j	ɰ			▨
Lateral approximant	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.



Where symbols appear in pairs, the one to the right represents a rounded vowel.

Figure 2.3: The IPA consonant table (left) and vowel quadrilateral (right) [Int99].

to use different schemes to describe them. This results in an IPA chart for describing phonemes that has separate sections for vowels and consonants. For a detailed description of the IPA chart and the possibilities it offers for describing the sounds of human speech the reader may refer to [Int99].

The generic classification into vowels and consonants as well as the different attributes used to describe the way the sounds from this classes are articulated is what we refer to as "articulatory features" (AFs) in this work.

2.3.1 Consonants

There are commonly three articulatory feature dimensions in which to describe consonants:

- Firstly, there is *place of articulation* which describes the position of the main constriction of the vocal tract on the mid-sagittal plane. The different places are represented by the columns in the IPA consonant chart (see Figure 2.3). Figure 2.4 shows the mid-sagittal plane of the human vocal tract and names possible places of articulation, also compare with Figure 2.1.

- Secondly, *manner of articulation* is used as another dimension. It describes the degree of the constriction of the vocal tract, the position of the velum, and some other attributes referring the behavior of the articulators such as vibration and redirection of the air stream from the middle to the side of the vocal tract.

- The third dimension describes the vocal cord vibration by classifying consonants as either *voiced* (vocal cords vibrate) or *unvoiced* (no vibration). In the IPA table for consonants every cell is split into half. The left half always refers to the unvoiced version of a consonant and the right one to the voiced version.

2.3.2 Vowels

Because of the generally open character of vowels they cannot be described by means of "place of articulation" etc. as consonants can. For vowels it is more appropriate to classify them by describing the horizontal and vertical position of the highest point of the tongue called the *dorsum*. The two dimensions of the dorsum position lead to the notion of an abstract vowel space that is usually visualized using the vowel quadrilateral depicted in Figure 2.3. In order to incorporate the use of the lips, un-rounded vowels are placed to the left of the back or front line of the quadrilateral and rounded ones to the right. Also, all vowels are classified as voiced sounds.

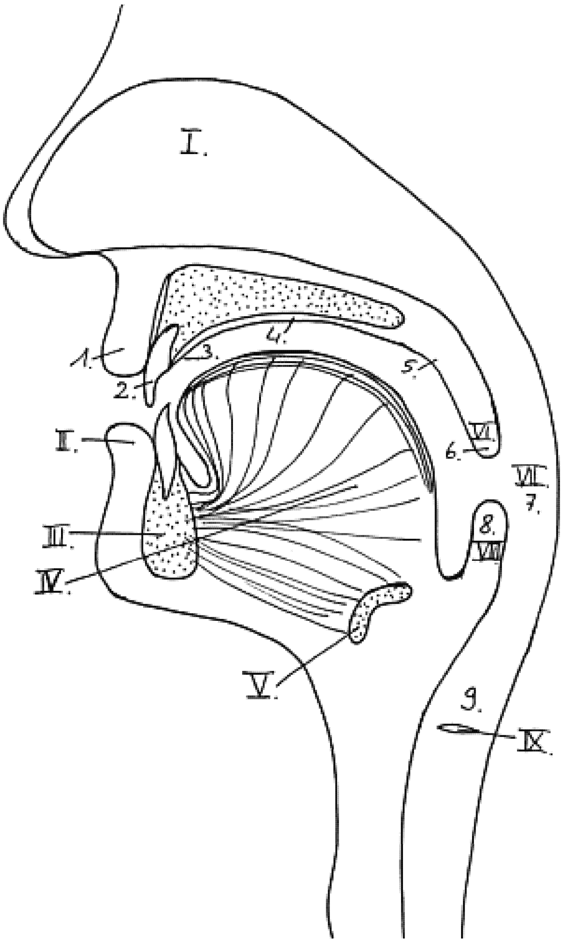


Figure 2.4: Mid-sagittal plane of the human head [Ell97]: Articulators are marked by Roman numbers: I. nasal cavity, II. lower lip, III. mandible, IV. tongue, V. hyoid, VI. uvula, VII. pharynx, VIII. epiglottis, IX. glottis. Places of articulation are marked by Arabic numbers: 1. lips, 2. incisors, 3. teeth-ridge, 4. hard palate, 5. soft palate, 6. uvula, 7. pharynx, 8. epiglottis, 9. glottis.

2.4 Conversational Speech

In spontaneous or conversational speech, pronunciations differ significantly from their canonic representations usually found in dictionaries and assumed in the IPA chart. For example, the tongue will not reach its "target" position in sloppy speech, voiceless sounds become voiced when they assimilate to voiced neighbors or voiced sounds can become devoiced. For this reason, a significant amount of ASR literature on conversational speech is devoted to pronunciation modeling, i.e. finding appropriate phonetic descriptions for conversational speech. Other factors, such as prosody and number as well as type of disfluency, also change with speaking style, but we neglect them as they are beyond the scope of this work. This does not generally invalidate the concept of phones and the "beads-on-a-string" model of speech [Ost99], however it makes it more difficult to use in practice. Still, the question for the "atomic" units of speech remains unsolved.

Several studies have compared the degree of attention to the articulation between conversational speech and clear speech. A good review can for example be found in [Esk93]. Attention to articulation is defined to be the degree of attainment of articulatory targets, such as a given formant frequency or stop releases. In general, articulatory targets are reached much more often in clear/ read speech than in conversational/ sloppy speech, for both consonants and vowels. Especially for vowels, there is much evidence suggesting increased articulatory efforts in clear speech, or equivalently, decreased articulatory efforts in sloppy speech:

- Formant values tend to achieve the extremes of the "vowel triangle" in clear speech, compared to more "central" values in sloppy speech. Variability of formant values is also found to be smaller in clear speech, indicated by a smaller cluster in a plot of F1/F2 values.

- Transition rates measure the movement of the formants at the onset and the offset of a vowel. They reflect the coarticulation of the vowel with its neighbors and indicate whether articulatory targets are achieved for the vowel or not. Some authors relate this to the casualness of speech. [PDB86] finds longer transition rates in clear speech, and more CV (consonant-vowel) coarticulation in spontaneous speech for most speakers [Esk93].

- Sloppy speech exhibits increased phonological variability. In the Switchboard Transcription Project [GHE96], linguists manually transcribed a portion of the Switchboard corpus at the phonetic level. It is clear that many words are not pronounced in the canonical way. Phonemes could be either deleted, or have their phonetic properties drastically changed, to such a degree that only the barest hint of a phoneme

> segment can be found. Greenberg consequently questioned the appropriateness of the phonetic representation in this project. Portions of the data are found to be quite hard to transcribe phonetically. It was reported that 20% of the time even experienced transcribers cannot agree upon the exact surface form being spoken. The transcribing process was unexpectedly time consuming, taking on average nearly 400 times real time to complete. For this reason, it was decided to transcribe only at the syllable level later on. Greenberg argues that syllables are a more stable, and therefore, a better unit for representing conversational speech as they are much less likely to be deleted.

The changes occurring in conversational speech at the articulatory and phonological level have also been studied with respect to ASR performance in [SK00, WTHSS96], differences between elicited and spontaneous speech are described in [SPSF00]. [Yu04] reports a re-speaking experiment, in which three participants of a meeting, which was recorded using close-talking microphones, were asked to re-read the transcript using (1) a clear voice and (2) a simulated ("acted") "spontaneous" speaking style. Recognizing these three data sets (which contain identical text) with a Broadcast News recognizers results in an error rate of 36.7% for the "read" part, 48.2% for the "acted" part, and 54.8% for the original "spontaneous" part. It is therefore clear that speaking style is a major factor influencing the performance of speech recognition systems.

Pronunciation change in conversational speech as opposed to read or "acted" speech is only partial most of the time; a phone is not completely deleted or substituted by another phone but it is modified only partially. Analysis of manual phonetic transcription of conversational speech reveals a large number ($> 20\%$) of cases of genuine ambiguity [SK00], where even human labelers disagree on the identity of the surface form. This observation leads us to our approach of modeling phonetic units as a combination of distinctive phonological features, which can then be varied according to speaker, speaking style and context. This follows an approach described in [Ste02], which argues for a model of "lexical access" (or speech-to-text), in which words are discriminated not by entire phones, but by a bundle of binary distinctive features, or "landmarks". While there is ongoing discussion about the process of spoken word recognition [FT87] and the units underlying perceptual processes in humans [GA03], there is evidence that sub-segmental cues play an important role in auditory lexical decision [MWW94, SB81] by providing acoustic invariants in speech [BS85].

[EB92] and [HHB89] have observed increased robustness against speaker changes in ASR systems (particularly speaker-dependent ones) based on phonological features as opposed to systems based on phonemes. This observation also supports the notion that phonological features should represent a useful invariant property to be used in the recognition of speech.

Chapter 3

Fundamentals of Statistical Speech Recognition

This chapter presents the key concepts of today's statistical speech recognition systems, as far as they are necessary for the understanding of this thesis. After formulating ASR as a statistical classification problem which maps speech to text, we describe typical feature extraction schemes and parameter estimation techniques for both acoustic and language models. A more comprehensive introduction can be found in most text books, for example [Rog05, WL90, Jel98].

Current state-of-the-art speech recognition systems are based on the concept of Hidden Markov models (HMM, see Section 3.4) to represent acoustic units. HMMs make it possible to model temporal variations in speech. The structure or syntax of a language is usually captured by statistical n-gram language models (LMs). Together with the acoustic model (AM), they form the "backbone" of a modern speech recognizer. From an algorithmic point of view, there are two basic problems:

Training: Techniques for robustly estimating the model parameters are required. Typically, today's training databases contain several hundreds of hours of speech and several millions of words.

Testing: The complexity of the acoustic and language models generated during training requires efficient search techniques in order to find the state sequence with the highest probability for a given test utterance in a reasonable amount of time.

3.1 Speech Recognition as a Classification Problem

The recognition process of a word sequence W can be formulated as the search for the maximum a-posteriori probability over all elements $p(W|O)$

in the search space, given the acoustic observations as a time series of feature or observation vectors and linguistic knowledge about the language. Given an utterance represented by a sequence of T feature vectors $O = (o_1, .., o_T)$, the classification problem, i.e. the search for the most likely word sequence W^*, can be expressed as:

$$\begin{aligned} W^* &= \operatorname*{argmax}_W p(W|O) \\ &= \operatorname*{argmax}_W \frac{p(O|W) \cdot P(W)}{P(O)} \\ &= \operatorname*{argmax}_W p(O|W) \cdot P(W) \end{aligned} \tag{3.1}$$

The maximization process of the a-posteriori probabilities allows for a separation of the a-priori probabilities $P(W)$ and the class conditioned probabilities $p(O|W)$. The best word sequence W^* is independent of the observation probability $P(O)$, which can therefore be ignored. The a-priori probabilities $P(W)$ are computed using the "language model" (LM) $P(W, \Xi)$. The class probabilities $p(O|W, \Psi)$ with parameters Ψ are called "acoustic model" (AM). Given this framework, research in speech recognition focuses on the estimation of the parameter of the language model Ξ and of the acoustic model parameters Ψ based on large training corpora.

3.2 Optimality Criterion

The previous section established a framework for statistical speech recognition and defined the "best hypothesis" for a given test utterance as the most likely hypothesis given a set of knowledge sources, namely acoustic and language model.

In a Bayesian framework, training of acoustic and language models then means creating good model estimates Ψ and Ξ for $p(O|W, \Psi)$ and $P(W, \Xi)$. Given two different sets of knowledge sources (Ψ_1, Ξ_1) and (Ψ_2, Ξ_2), the one producing the best expected probability (or likelihood) $\langle p(W|O) \rangle$ over a test corpus O is indeed the one producing better hypotheses in the sense that these better correspond to what was actually said. However, in order for this approach to be valid, complete knowledge about the process that generates the probability distributions is assumed, which is not achievable in reality. Nonetheless, the likelihood is usually used as an optimization criterion for (acoustic) model training.

The quality of a trained ASR system is better measured in terms of the "word error rate" (WER), which is defined as the minimum edit distance between a reference transcription and a given hypothesis divided by the length of the reference transcription, which means that the performance of an ASR system is evaluated using a criterion which is fundamentally different from the system's internal optimization model. This approach however works well in practice. Other approaches, which do not optimize the likelihood, but the a posteriori confidence of the word hypothesis also exist [MBS00]. Nevertheless, this pragmatic approach has lead to significant improvements (also in WER) over the last few years.

As there are three different kinds of errors (deletions, insertions, and substitutions), the WER can be computed as

$$\text{WER} = \frac{\texttt{\#DEL} + \texttt{\#INS} + \texttt{\#SUB}}{\texttt{\#REF\_WORDS}}$$

In the following example the word error rate is $\frac{1+1+1}{5} = 60\%$:

Reference	I	HAVE	HEARD	YOUR	VOICE	
Hypothesis	I		HEARD	YOU'RE	VOICE	IT
Error		DEL		SUB		INS

The "word accuracy" (WA) is defined as WA $= 1 - \text{WER}$. Other approaches to speech recognition evaluation focus more on the end-to-end performance of an application, these include:

- Weighted WER (i.e. evaluated on content words only)
- Summarization score
- Information retrieval score

Also, systems participating in "rich transcription" evaluations [MFP+04, MJF+04, MSW+04, NIS04a] are increasingly demanded to annotate their output (word hypotheses) with meta information such as confidence measures [SK97], which can then be used in further processing, such as summarization, IR, or language identification [MKS+00].

As mentioned before, ASR systems are evaluated using the WER metric, the "best" recognizer is not necessarily the one producing the best $p^*(W^*|O)$. Therefore, models do not have to be trained using maximum likelihood, other criteria can be employed as well, or criteria can be mixed. In this work, therefore use discriminative training criteria for the acoustic model, which do not improve the likelihood of the training data, but instead reduce the WER ("minimum word error", MWE) or related criteria. Nonetheless, the search process still computes the hypothesis with the best expected probability, although the models have not been trained using maximum likelihood in the strict sense. This approach of using maximum likelihood and discriminative training criteria in one system is prevalent in modern state-of-the-art ASR systems. Currently popular discriminative criteria include: Maximum Mutual Information Estimation (MMIE) [WP02], and Minimum Phone Error (MPE) [Pov05].

Figure 3.1 gives an overview of the progress of speech recognition over the years on different corpora with different speaking styles. It shows clearly that

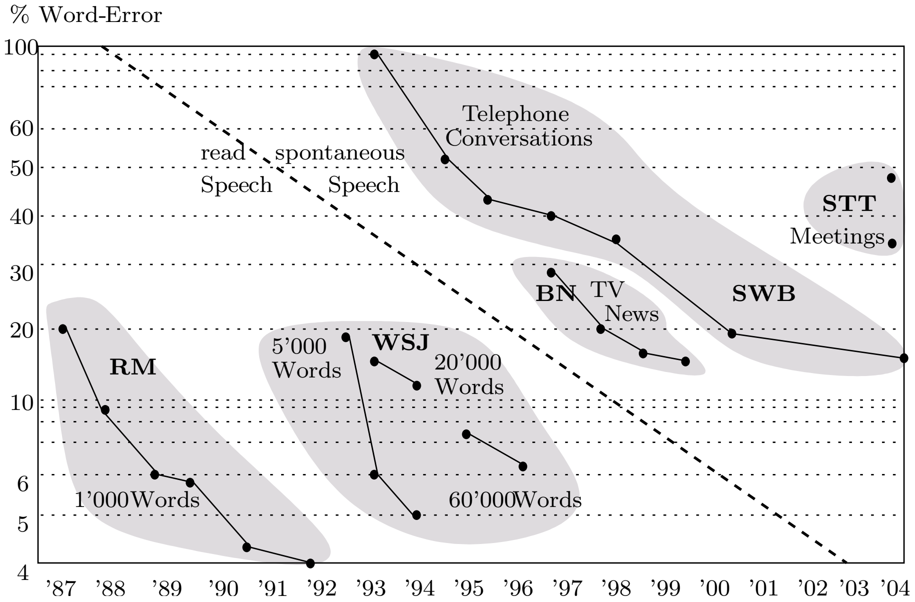


Figure 3.1: Progress (word error rates) in speech recognition over the years [Rog05]. In 2005, best CTS numbers were around 15%, while "Meeting"-type speech continues to pose a challenge at around 30% WER.

read or planned speech is much easier to recognize than spontaneous speech. As of 2005, word error rates for the Conversational Telephony Speech (CTS) task have dropped to around 15%, while the conversational "Meeting" task has replaced CTS as the "most difficult" task in ASR research. State-of-the-art systems have word error rates of around 30% on close-talking "Meeting" data [NIS04a]. While absolute numbers vary from language to language, the development of non-English speech recognition systems generally follows the same pattern. As the largest and most active research community is currently centered around English speech recognition, it defines the state-of-the-art and most ideas are only ported to non-English systems once their effectiveness has been confirmed on one of the tasks presented above.

3.3 Recognizer Design

A modern speech recognition system consists of three main information sources, which have to be generated and trained:

Acoustic Model (AM): The AM contains the HMM's observation probabilities $p(O|S)$ for a given observation O and states S, using the dictionary for the mapping between word sequence W and state sequence S, one can also write $p(O|W)$.

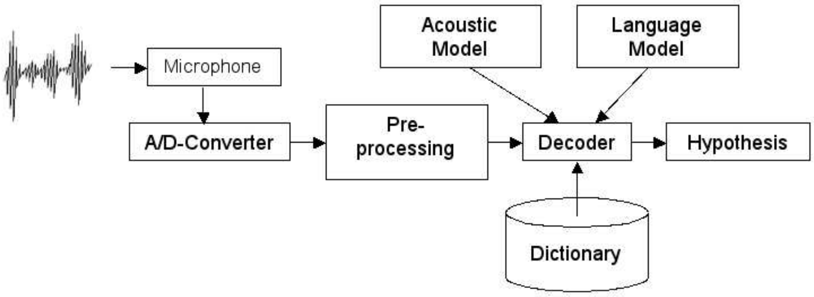


Figure 3.2: Components of a statistical speech processing system: The acoustic model contains the HMM emission probabilities while the structure of the HMM, i.e. the possible transitions and their probabilities, are determined by the dictionary and the language model. In this work, we are only concerned with the acoustic model.

Language Model (LM): The LM contains the a-priori probability $P(W)$ of a given word sequence W.

Dictionary: The dictionary (together with the language model) determines, which state sequence S should be used to model a given word sequence W, i.e. it contains a mapping of words to speech sounds, for example:

`speech` → /spiːtʃ/

When writing $p(O|W)$ for an acoustic model probability, the dictionary (i.e. the mapping of words to speech sounds) is implicitly included into the acoustic model.

Figure 3.2 presents a functional diagram of a modern statistical speech recognition system: in this work, we will improve the acoustic model of an existing speech recognizer, but leave the other components unchanged.

3.4 Hidden Markov Models

Today's statistical speech recognition systems usually employ HMMs for building acoustic models. Speech production is seen as a stochastic process: we describe words, phones, etc. as "states" in a linear sequence describing the speech production process. Each state "emits" (observed) sounds with a certain probability distribution. ASR then becomes the problem of finding the "most likely state sequence" for a given observation. This "decoding" problem is discussed in Section 3.9.

HMMs are defined as a tuple of:

- A set of states $S = \{s_1, s_2, \ldots, s_n\}$
- The initial probability distribution: $P(s_i)$ is the probability of s_i being the first state in a sequence
- The state transition probability matrix $A = (a_{ij})$ for transitions from s_i to s_j
- The set of emission probability distributions or densities: $\{p^1, p^2, \ldots, p^n\}$ where $p^i(o)$ is the probability $p(o|s_i)$ of observing o when the system is in state s_i
- The feature space O can be discrete or continuous. Accordingly, the HMM is called a discrete HMM or a continuous (density) HMM (CDHMM)

This model is called a "Hidden" Markov Model as we observe the emitted symbols, but not the associated state sequence. This formulation is very compact and can be trained and evaluated efficiently, as there are no dependencies between states apart from the transition probabilities. Systems usually use several thousand context-dependent acoustic models which are generated using various tree-based clustering schemes [Rog97], starting from about fifty base phones. Hence, each context-dependent model is trained on a very small subset of the training data only, which can make generalization to unseen contexts difficult.

HMMs have a number of properties:

- For the initial probabilities we have $\sum_i P(s_i) = 1$
- Frequently, we choose $P(s_{\texttt{<s>}}) = 1$ (and call `<s>` "begin of sentence")
- $\sum_j a_{ij} = 1$ for all i and $a_{ij} = 0$ for most j in ASR

Examples for typical HMM topologies are shown in Figure 3.3. HMMs pose three main problems, which are solved by different algorithms [Rog05]:

The evaluation problem: Given an HMM state sequence $S = (s_1, s_2, \ldots, s_n)$ and an observation sequence $O = (o_1, o_2, \ldots, o_T)$, compute the probability $p(O|S)$ that the observation was produced by H (typically: $n \neq T$)

→ *Forward Algorithm*

The decoding problem: Given an HMM state sequence $S = (s_1, s_2, \ldots, s_n)$ and an observation sequence $O = (o_1, o_2, \ldots, o_T)$, compute the most likely state sequence $(q_1, q_2, \ldots, q_T)$, i.e.

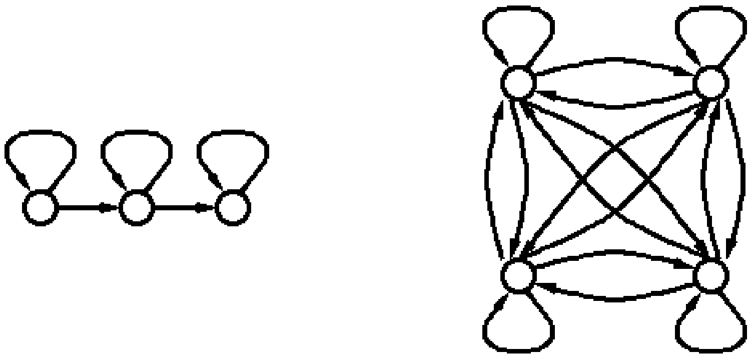


Figure 3.3: Typical HMM topologies: (left-to-right) linear (left) and ergodic (right). All examples employ self-loops, i.e. $a_{ii} > 0$.

$$\underset{(q_1,q_2,\ldots,q_T)}{\operatorname{argmax}} \, (q_1, q_2, \ldots, q_T | O, S)$$

→ *Viterbi Algorithm*

The learning or optimization problem: Given an HMM state sequence $S = (s_1, s_2, \ldots, s_n)$ and an observation $O = (o_1, o_2, \ldots, o_T)$, find a new model S' so that $p(o_1, o_2, \ldots, o_T|S') > p(o_1, o_2, \ldots, o_T|S)$

→ *Expectation Maximization (EM) Algorithm*, which makes use of the *Forward-Backward-Algorithm* (see Section 3.6)

The left HMM in Figure 3.3 shows the structure used in Janus [FGH+97] to model phones: A phone is modelled as a linear sequence of begin-, middle-, and end-state ("tri-state architecture"). Transitions are allowed into the same state ("self loop") or the next state only (i.e. all other $a_{ij} = 0$. Given this type of phone model, words can be modelled by simply appending the HMM phone models in the order the respective phones appear in the dictionary, using the correct context-dependent model. Word transitions can be modelled in the same way. Language model probabilities appear as transition probabilities at word boundaries.

3.5 Extraction of Relevant Features

The goal of the pre-processing step is to remove problem-invariant features from the digitized acoustic signal and to construct an "optimal" feature space for the acoustic models Ψ. "Optimal" of course means resulting in a lower word error rate and containing as few parameters as possible.

"Features" in the context of this section refer to time-series of n-dimensional parameters describing the *acoustic* signal only (e.g. the energy in the 200Hz frequency bin), as opposed to "*Articulatory* Features", which try to describe the articulatory process that generated the acoustic signal. Of course, these features could also spawn an "optimal" feature space usable for ASR, particularly as a low-bitrate coding scheme for transmission or recognition of speech [ZLR+95].

In the first step, a short-time spectral analysis is performed to extract features in the spectral domain. This step is valid, since it can be assumed that the speech signal is stationary over a short period of time. The next assumption is that the phase spectrum does not contain meaningful information for speech recognition. Consequently, only the power spectrum is passed to the next step. The properties of human perception of audio signals are emulated by a logarithmic scaling of the signal energy and a frequency scaling by applying a filter bank, e.g. mel or bark coefficients. Based on Fant's source-filter model [Fan60], a so-called liftering process is used to separate the vocal tract's transfer function from the periodic excitation signal. To that end, an inverse cosine function is applied to transform the signal from the spectral to the cepstral domain. These features are called mel-filtered cepstral coefficients (MFCC). Channel normalization is performed by cepstral mean subtraction (CMS). Additionally, the feature values can be divided by their variances (cepstral variance normalization, CVN) on a per-utterance or global basis. The next step induces temporal context information: cepstral features from adjacent windows are concatenated into a single feature vector. A linear discriminant analysis (LDA) is used as a final step to transform the feature space. The LDA transform attempts to maximize the inter-class variances while minimizing the intra-class variances. At the end of so-called "pre-processing", the original audio file has been transformed into a sequence of T N-dimensional feature vectors $O = (o_1, o_2, \ldots, o_T)$. T is the length of the utterance expressed in frames of typically 10ms and N is usually in the order of 16 to 42.

Vocal Tract Length Normalization (VTLN) is a feature transform which attempts to normalize the frequency changes due to different vocal tract lengths [AKC94]. Fant's source-filter model suggests that the formant frequencies are scaled with the length of the vocal tract. Systematic speaker variations can be compensated for by warping the frequency axis. To that end, a piece-wise linear function $f(\omega)$ can be employed:

$$f(\omega) = \begin{cases} \alpha\omega & : \quad \omega < \omega_0 \\ \beta\omega + \gamma & : \quad \omega \geq \omega_0 \end{cases}$$

where β and γ can be obtained via constraints at $f(\omega_0)$ and $f(\omega_N)$. The "warping factor" α can be estimated using maximum likelihood [ZW97]:

$$\mathcal{L}(\alpha) = \sum_t \log(J(\alpha)P(f(o_t, \alpha)|S))$$

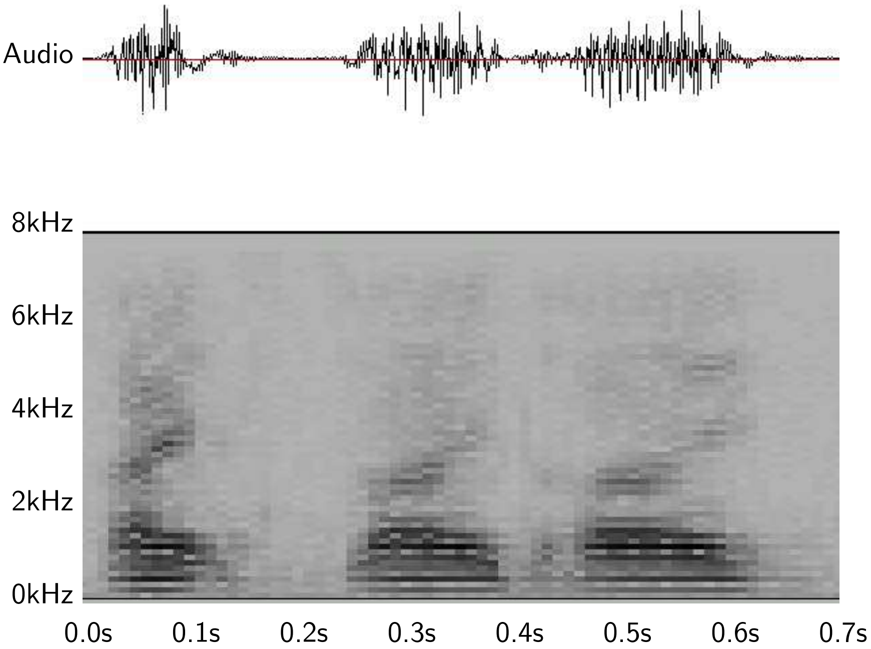


Figure 3.4: Pre-processing for ASR: the digitized audio signal (top) is converted to a spectral representation for a succession of short segments ("visible speech", bottom).

A Brent search is often used since no closed-form solution is available. Furthermore, the derivative $J(\alpha)$ is ignored and the resulting function formally no longer satisfies the requirements of a probability density function (PDF).

Example

Two different representations of speech are shown in Figure 3.4. The visual effect of the MFCC mel transformation, which is applied to the short-term Fourier spectrum ("visible speech") shown in Figure 3.4 in addition to the digitized signal, is a dimensionality reduction and smoothing of the frequency axis. Again, details or further explanations can be found in most text books, for example [Rog05, WL90, Jel98].

The sequence of feature pre-processing steps as presented here is fairly standard in the ASR community, although countless variations and flavors exist. Details of the pre-processing employed for every experiment in this work are given in the respective sections.

3.6 Acoustic Models

Acoustic modeling deals with the probabilities $p(O|S)$, where S denotes a state sequence and O is a sequence of feature vectors. Since speech signals exhibit differences in temporal and spectral domain, an appropriate model must deal with both dimensions in a statistically consistent way. The temporal changes can be modelled as a finite state automaton with associated transition probabilities between the states. Attaching observation probabilities to each state extends the automaton to an HMM. This model is also called "first order Markov process" since the state probability depends only on the predecessor. Defining $S = \{s_1, s_2, \ldots, s_n\}$ as a set of n HMM states and $\mathcal{S} = S^T$ as the set of all state sequences of length T, the probability $p(O|S)$, given the model Ψ, can be computed as:

$$p(O|S,\Psi) = \sum_{q\in\mathcal{S}} \prod_t a_{q_t q_{t+1}} p(o_t|q_t) \tag{3.2}$$

The element $q \in \mathcal{S}$ represents one path through the state automaton and q_t denotes the state index at time t. The variable a_{ij} represents the probability for the transition from state s_i to s_j. The *Forward-Backward Algorithm* computes these probabilities via dynamic programming with a complexity of $O(Tn^2)$. The forward (α_t) and backward (β_t) probabilities are defined as:

$$\begin{aligned} \alpha_t(j) &= p(o_1..o_t, q_t = s_j|\Psi) \\ \beta_t(j) &= p(o_{t+1}..o_T|q_t = s_i, \Psi) \end{aligned}$$

The conditional probability $p(O|S,\Psi)$ can be expressed as a sum over the α and β:

$$p(O|S,\Psi) = \sum_i \alpha_T(i)\beta_T(j)$$

The α and β can now be computed using a recursion:

$$\alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{ij} p(o_t|q_t = s_j) \tag{3.3}$$

$$\beta_t(j) = \sum_i \beta_{t+1}(i) a_{ji} p(o_{t+1}|q_{t+1} = s_i) \tag{3.4}$$

The *Viterbi* algorithm is similar to the *Forward-Backward* algorithm but requires only one pass: if the $\sum$ operator in Equation 3.3 is replaced by the max operator, the best state sequence can be obtained as follows:

$$q^* = \operatorname*{argmax}_{q\in\mathcal{P}} \prod_t a_{q_t q_{t+1}} p(o_t|q_t)$$

This expression is evaluated by Viterbi-Decoding algorithms as discussed in Section 3.9.

Despite the availability of efficient algorithms to work with HMMs, there are several drawbacks. One important point is that the emission probabilities depend only on the current state. Thus, certain dependency relations between states cannot be expressed. For example, the observed feature vectors may depend on several factors such as speaking rate, dialect, gender, error recovery mode, microphone, or environmental noise. In an HMM framework, these factors must be treated as one state, although conditional independence between these factors may be an issue. A factorization of these random variables would allow for a better parameter sharing scheme. In the HMM framework, a state must represent all of these combinations to express the emission probabilities. As a result, the number of HMM states would grow exponentially. Factorial HMMs [Gal02] or dynamic Bayesian Networks (DBNs) [ZR98] make it possible to factorize such dependencies. However, parameter estimation and decoding in a BN framework is complex and computationally demanding, so that this approach is impractical even with today's resources for the systems described in this thesis.

Kullback-Leibler Statistics

Parameter Estimation for ASR usually focuses on the emission probabilities, which are usually modelled by Gaussian Mixture Models (GMMs). Practical considerations restrict the covariance matrix Σ to diagonal form. The PDFs for emission probabilities now look as follows:

$$\begin{aligned} p(o|s,\Psi) &= \sum_i w_i N(o|\mu_i,\Sigma_i) \\ N(o|\mu,\Sigma) &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(o-\mu)^t\Sigma^{-1}(o-\mu)} \end{aligned}$$

The HMM model is now fully specified. The parameters consist of the transition probabilities, mixture weights, diagonal covariances, and mean vectors.

Baseline parameter estimation is based on the ML principle. A direct application of the ML-principle on HMMs is, however, not possible. Instead, the Kullback-Leibler statistics are used to establish an iterative algorithm, known as the Baum-Welch re-estimation procedure. Introducing a variable q for the (hidden) state sequence and initial parameter Ψ^0, the Log-Likelihood $\mathcal{L}$ of parameter Ψ for an HMM can be expanded as:

$$\begin{aligned}
\mathcal{L}(\Psi) &= \log p(O|\Psi) \sum_{q\in\mathcal{P}} p(q|O,\Psi^0) \\
&= (\log p(O,q|\Psi) - \log p(q|O,\Psi)) \sum_{q\in\mathcal{P}} p(q|O,\Psi^0) \\
&= \sum_{q\in\mathcal{P}} \log p(O,q|\Psi) p(q|O,\Psi^0) - \\
&\quad \sum_{q\in\mathcal{P}} \log p(q|O,\Psi) p(q|O,\Psi^0)
\end{aligned}$$

The likelihood can be expressed as the Kullback-Leibler statistics

$$Q(\Psi,\Psi^0) = \sum_{q\in\mathcal{P}} \log p(O,q|\Psi) p(q|O,\Psi^0)$$

and a remainder term.

Maximizing the parameters Ψ with respect to the Kullback-Leibler statistics, $Q(\Psi,\Psi^0) \geq Q(\Psi^0,\Psi^0)$ increases the likelihood $\mathcal{L}(\Psi) \geq \mathcal{L}(\Psi^0)$. In the HMM framework, the term $p(q|O,\Psi^0)$ in $Q(\Psi,\Psi^0)$ denotes the state occupancies obtained using initial model parameters. The Baum-Welch algorithm increases the likelihood in each training iteration until saturation is reached. However, the model parameters depend on the initial settings Ψ^0, as convergence is to a local maximum only. Various schemes such as deterministic annealing strategies, which are also biologically inspired [ABK+00], or merging and splitting of Gaussians [UNGH98] during training exist to improve the quality of the resulting PDFs on realistic data.

Semi-tied Full Covariances

Semi-tied full Covariances (STC) [Gal99] or Maximum Likelihood Linear Transform (MLLT) [Gop98] introduce linear transforms for covariance modeling. The motivation for this approach is that diagonal covariances are used for practical reasons (i.e. speed), but the observation space does not really support this restriction since the features are correlated, which results in significant off-axis probability mass. A better parameter sharing scheme may be achieved by sharing the full transform matrices. The PDF is structured as follows:

$$P(o|s,\Psi) = \sum_i w_i N(o;\mu_i, A^T \Sigma_i A)$$

where Σ_i is a diagonal matrix per component and A is supposed to be a full matrix which may be shared across components and states. Since the term $A^T\Sigma_i A$ represents a full matrix, the PDF evaluation becomes computationally expensive. If the inverse matrix $B = A^{-1}$ is used, a more efficient feature and mean transform can be obtained:

$$P(o|s,\Psi) = |B| \sum_i w_i N(Bo; B\mu_i, \Sigma_i)$$

The resulting Kullback-Leibler statistics are of the same form as for the feature adaptation with the exception that the same matrix B is applied as an extra transform to μ:

$$Q(B,B^0) = c + \sum_{i,t} \gamma_i(t)(\log|B| - c_i - \frac{1}{2}(Bo_t - B\mu_i)^T \Sigma_i^{-1}(Bo_t - B\mu_i))$$

3.7 Adaptation

Statistical models are optimized on a large amount of training data, which should resemble the test data as closely as possible. As test conditions are usually unknown in advance and subject to change, robust systems are usually trained on a number of input conditions, to cover as many test conditions as possible. The resulting system can then be improved by adapting it to a specific test condition. Adaptation can be done using the reference transcription of the adaptation data ("supervised adaptation") or using a recognizer hypothesis for the transcription ("unsupervised adaptation").

Acoustic Model Adaptation

The ML-criterion can also be used for estimating a linear transform of the model parameters [Leg95] in the Maximum Likelihood Linear Regression "MLLR" framework. In the context of mixtures of Gaussians, an adaptation of the means of Gaussians can be represented by such PDFs:

$$p(o|s,\Psi) = \sum_i w_i N(o; A\mu_i, \Sigma_i)$$

Keeping the Gaussian parameters w_i, μ_i, Σ_i fixed, the Kullback-Leibler statistics can be used to estimate the linear transform A. The Kullback-Leibler statistics can be written as:

$$Q(A,A^0) = c - \sum_{i,t} \gamma_i(t)(c_i + (o_t - A\mu_i)^T \Sigma_i^{-1}(o_t - A\mu_i))$$

The state probabilities $\gamma_i(t)$ are computed using the initial parameter A^0. Terms not relevant for the optimization are denoted by c and c_i. The maximization of Q requires solving:

$$\frac{d}{dA} Q(A,A^0) = 0$$

Differentiating Q with respect to A leads to a set of linear equation systems, which can be solved row by row.

$$\sum_{i,t} \gamma_i(t)\Sigma_i^{-1} o_t \mu_i = \sum_{i,t} \gamma_i(t)\Sigma_i^{-1} A \mu_i \mu_i$$

In analogy, a similar transformation can also be applied to the covariance matrices Σ_i of the Gaussians.

Feature Adaptation

Linear transforms can also be applied in the feature space. This technique has computational advantages over model adaptation since combinations with adaptive training schemes and Gaussian selection algorithms are easy to realize. When transforming the features, it is not possible to transform means and covariances differently as is the case when transforming models, so this approach is also called "constrained MLLR" or FSA ("feature space adaptation").

Given a PDF $p(x)$ and a feature transform $f(x)$, an appropriate PDF with respect to f would be $\hat{p}(x) = p(f(x))\frac{d}{dx}f(x)$. This ensures that the probability mass is conserved:

$$\int p(x)dx = \int p(y)dy = \int p(y)\frac{dy}{dx}dx = \int p(f(x))\frac{df(x)}{dx}dx = \int \hat{p}(x)dx$$

When $f : \vec{x} \to \vec{y}$ is a vector function, the corresponding substitution rule is extended to the functional determinant or Jacobian. The corresponding Kullback-Leibler statistics for a linear transform $f(x) = Ax$ therefore are:

$$Q(A, A^0) = c + \sum_{i,t} \gamma_i(t)(\log|A| - c_i - \frac{1}{2}(Ao_t - \mu_i)^T \Sigma_i^{-1}(Ao_t - \mu_i))$$

The Jacobian $|A|$ term complicates the optimization process. However, the Laplace development for a row j results in the following representation of the Jacobian:

$$\begin{aligned} |A| &= \sum_{jk} a_{jk}\tilde{a}_{jk} \\ \tilde{a}_{jk} &= (-1)^{j+k}|A_{jk}| \end{aligned}$$

where $\tilde{a}_{jk}$ denotes the adjunct of A, given j and k. This allows for the implementation of an iterative row-by-row optimization scheme. The adjuncts $\tilde{a}_{jk}$ are being kept fixed while optimizing row j.

3.8 Language Models

The language model (LM) describes the a-priori probabilities $P(W)$, where $W = w_1, w_2, \ldots, w_n$ denotes a sequence of words.

For small, limited domains, context free grammars (CFG) are used to introduce constraints for the search space. The disadvantage of CFGs is that current algorithms to learn the structure from data do not work very well. Human labor is, therefore, required to a great extent during the preparation of grammars.

On tasks covering large domains, statistical n-gram models are popular. The word "memory" is constrained to n words, so that an n-gram model predicts the probability of the next word given a "history" of $n-1$ words. Typical systems use 3-grams, 4-grams, or sometimes 5-grams. Higher order are impractical because of lack of training data and disc space. The mathematical formulation of a trigram is as follows:

$$P(W) = \prod_i P(w_i|w_{i-1}, w_{i-2})$$

Backing-off schemes are used to capture unseen n-grams. The models may be "refined" by adding word classes, phrases, and interpolations of them. The models can be trained by several criteria, such as maximum likelihood or maximum entropy.

3.9 Decoding

The task of the decoder is to find the best solution W^* to the problem

$$W^* = \operatorname*{argmax}_W p(O|W) \cdot P(W)$$

as quickly as possible. Decoding can be done in two ways

Depth first: expand every hypothesis and decode it until it no longer is the best hypothesis, at which point it is discarded (time-asynchronous).

Breadth first: expand all hypotheses of the current frame into the next frame (time-synchronous) and prune the worst hypotheses.

Sometimes, it is necessary to have more than one decoding pass to incorporate for example the full language model, new words, or complex acoustic models, which cannot be handled in a time-synchronous way because of context dependency. Also, acoustic models are frequently adapted on the hypothesis of a previous decoding pass in multi-pass search strategies.

An example for depth first strategy would be "stack decoding" [Jel69], which keeps a sorted list of partial hypotheses and sequentially expands the best hypothesis with words, re-inserting them into the list where appropriate. This approach works similar to the well-known A^* search.

Stack decoders can easily deal with complex language models. However, for efficiency most ASR systems today use a time-synchronous beam search. In these breadth-first designs, all hypotheses are pursued in parallel as the decoder evaluates all time frames sequentially. Given a set of "active" hypotheses (or HMM states) $\mathcal{S}_{t-1}$ at time $t-1$, the decoder expands them to time t and compute $\mathcal{S}_t$ by evaluating the HMM topology, the AM, and LM accordingly. In order to avoid factorial explosion, equivalent states are usually recombined at t using the Viterbi approximation, after which the best state s_t^* at t can be determined, so that all states in $\mathcal{S}_t$ worse then s_t^* by a certain score difference (or "beam") can be "pruned away", i.e. discarded before the states from $\mathcal{S}_t$ are expanded into $\mathcal{S}_{t+1}$.

The main challenge during decoding is the organization of the search space and the evaluation of the output probabilities for several thousand distinct acoustic models. The trade-off in decoding is usually speed vs. accuracy. The decoder used in this work is a Viterbi decoder capable of using arbitrary language models and cross-word tri-phone acoustic models, it is described in [SMFW02]. An overview of current decoding techniques is given in [Aub02].

The output of a beam decoder is usually retrieved from the state sequence using a back-trace from the best node in the final frame to the start frame. In many applications, it is however necessary to retrieve not only the single best hypothesis, but to get more information about the search space, i.e. alternative (less likely) word sequences.

N-best lists

An N-best list is an ordered (by score) list containing not only the most likely hypothesis, but alternative phrases, too. An example would be:

```
show me the interface   please
show me     in her face please
show me     in her face see
show    the in her face please
show    the in her face see
show    the inner  face see
```

In a stack decoder, this corresponds simply to the N best complete hypotheses at the end of the evaluation phase. Using a Viterbi decoder, it is also possible to retrieve such a list by performing multiple trace-backs.

Very often, in particular when working with spontaneous speech, it is found that $N > 1000$ is often required in order to capture variability not only in "minor" words such as `a`, `an`, `the` or noises or to contain the correct transcription `show me the interface please` as in the example above.

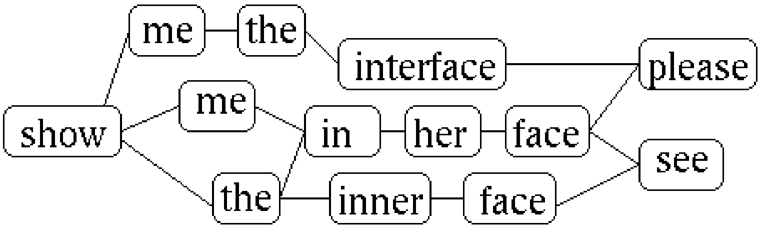


Figure 3.5: A directed a-cyclic graph of words or “lattice” typically used in speech recognition.

Lattices

An alternative approach to N-best lists, which can also conveniently be extracted from the back-trace of a Viterbi decoder, is a directed, a-cyclic graph of words, i.e. a “lattice”. Such a structure is shown in Figure 3.5. Typically, every node (or word) in a lattice is annotated with start and end times and acoustic score. Links are often annotated with acoustic cross-word model scores and, if given, language model scores. Lattices are usually significantly more compact then N-best lists, the above example of 6 sentences using 34 words can be represented in a lattice of just 13 nodes.

Lattices or N-best lists map the search space employed by the decoder, showing the most likely competing hypotheses. Therefore, discriminative training, also in this work, very often relies on them. Lattices are also used to compute a-posteriori probabilities or confidence scores for words [SK97]. N-best lists and Confusion Networks [MBS00] can also readily be derived from lattices.

Chapter 4

Related Work

The goal of this work is to improve an existing state-of-the-art speech recognizer and show how the inclusion of articulatory information can improve performance. This chapter will therefore present relevant work in the field of conversational speech recognition, use of AFs for ASR, combination of several information sources, and give an overview on work in related areas such as speaker recognition and language identification using AFs.

The experiments presented in this work were conducted using the Janus [FGH+97] speech recognition toolkit and the Ibis [SMFW02] decoder, which provides a state-of-the-art environment for ASR research and includes all the techniques mentioned in Chapter 3.

4.1 Large Vocabulary Conversational Speech Recognition

For HSC, GlobalPhone, and ESST the experiments presented in this thesis have been performed on the best available systems in-house, however there are to our knowledge no comparable external systems to compare against. The robustness experiments on conversational speech were performed using ISL's RT-04S "Meeting" evaluation system [MFPW05, MJF+04, MFP+04], which is a state-of-the art system on one of the most difficult tasks in Large Vocabulary Conversational Speech Recognition (LVCSR) today.

Systems for the close-talking condition of the speech-to-text task on meeting are typically trained on around 100h of dedicated meeting training data pooled with BN data [MJF+04] or are trained on large corpora of telephony speech and adapted on the meeting data [MSW+04] to compensate for the mismatch in bandwidth and channel. The best error rates in 2004 on the RT-04S meeting development data using manual segmentation are 29.8% for the ICSI system [SWM+04] and 28.0% for the ISL system [MFP+04], which will be used for AF experiments in this work. The best number published on this task as of 2005 is 27.9%, which uses a larger amount of CTS acoustic training data and web-data for language modeling [SAB+05] than was available in 2004. “Meeting” error rates are therefore more than twice as high as CTS error rates, for which a comparable number on the NIST 2004 development set is 13.0% [SKM+05].

4.2 Pronunciation Modeling for Conversational Speech

Most of the work on dedicated models for conversational speech has focused on finding dictionaries suitable for conversational speech. Pronunciation variants of frequent words are added to the dictionary, allowing for typical and predictable changes in conversational speech. However, changes are only possible at the phonetic level, i.e. one phone can be added, deleted, or substituted. Changes at the sub-phonetic level, i.e. partial changes, are not possible. This approach assumes that the deviation from the canonical pronunciation can be represented in terms of complete changes from the base-form phonemes to the surface-form phones. One problem with this approach is that pronunciation variants are related to a variety of factors such as speaking style, speaking rate, individual speaker habits and dialectal region. If however the variants added to the dictionary are badly chosen with respect to the particular task, recognizer, or speaker, the overall performance may decrease. Great care therefore has to be taken when creating pronunciation variants [ADL99], which are often generated using expert linguistic knowledge or trained on data [Wes03]. The recognizers used in this work employ this approach by containing on average approximately 1.2 pronunciations per word. Multi-words, which model reductions from “going to” to “gonna” etc. are also used by our conversational speech systems [FW97b].

Another approach to handling pronunciation variability is to create several dictionaries using different phone sets and combine the hypotheses from the separate recognizers [LG05] using ROVER [Fis97] or Confusion Network Combination [MBS00]. Phone sets then have to be created and maintained separately, also the decoding effort increases linearly with the number of phone-sets used. Other recent approaches to overcome the limitation of having to model conversational speech with a fixed and limited sets of acoustic units replace phones with articulatory instances of their phonetic attributes [FFKW99], while allowing other attributes (e.g. nasalization), too, which results in a richer pronunciation model that can be learned on data. Another approach generalizes the context clustering tree [Yu04] by sharing the root nodes between different phones. In this approach, models can be shared for phones, where the surface pronunciation is not well represented by base form pronunciations, resulting in a more robust estimation of models.

4.3 Articulatory Features for ASR

4.3.1 Overview

Inspired by the process a human expert uses to “read” a spectrogram, i.e. which cues he or she uses to identify and classify segments in a graphical representation of speech, there have been several attempts at incorporating articulatory and phonetic expert knowledge into systems for automatic speech recognition, e.g. [ZL86]. Roughly, they can be grouped into the following classes ranked according to complexity:

- Include AFs as additional features into the front-end of an otherwise standard recognizer. These approaches basically assume that AFs are a better projection of the speech signal for recognition than standard auditory-based pre-processing and can therefore be used to augment such a system.

- Segment-based recognizers using AFs. These systems can either solely rely on AFs or combine AFs with existing acoustic models. Depending on the kind of segmentation and integration (Hidden-Markov-Models, Dynamic Bayesian Networks , ...), some degree of asynchrony between features is permitted. However, AFs are regarded as abstract phonological or perceptual classes, which do not necessarily exactly correspond to physical movements.

- Explicit modeling of articulatory trajectories: these generative approaches (“analysis by synthesis”) try to recognize speech by evaluating physical models and comparing them with the speech signal. Therefore, dynamic constraints have to be used in order to overcome a many-to-one mapping problem, in that many configurations of the vocal tract can result in the same acoustic signal.

The approach pursued in this thesis fits in the second class, because it promises a good compromise between theoretical motivation and performance improvements (the main drawbacks of the first class) and computational complexity, the main disadvantage of most systems in the third class.

4.3.2 AFs as Additional Features

Eide [Eid01] used articulatory attributes to enhance the front-end of a speech recognizer. She trained a classifier based on GMMs for articulatory attributes. The output of these GMMs is then combined with the original cepstral observation vector to form an extended front-end, which is then used to train the “real” acoustic models. She observed an error reduction of up to 25% on car audio data. She argues that the direct modeling of phonemes from the waveform as it is usually done in the beads-on-a-string model [Ost99] disregards some of the phenomena of conversational speech such as the relaxation of the requirements on the production of certain distinctive features. She claims that variations in the pronunciation may cause big phonemic differences while in terms of articulatory features the difference may be considerably smaller because only few articulatory features actually change their value. Therefore she argues that the task of recovering a word sequence from a feature representation is more feasible than from a phonemic representation. In earlier work, binary linguistic features have been used for phoneme classification and word-spotting [ERGM93].

Approaches for feature fusion can also be regarded as articulatory approaches, if merging for example the "Voicing" feature with standard features using Linear Discrimination Analysis (LDA), as is done by *Zolnay* in [ZSN05], where a gain of up to 7% relative was observed on the German Verbmobil II data.

4.3.3 Segment-based Articulatory Features

Different explanations for the poor performance of HMM based recognizers on spontaneous speech as well as reasons why articulatory features used in pseudo-articulatory classes might help in overcoming the encountered problems have been proposed by different researchers.

Ostendorf [Ost99], for example, argues that pronunciation variability in spontaneous speech is the main reason for the poor performance. She claims that though it is possible to model pronunciation variants using a phonetic representation of words the success of this approach has been limited. Ostendorf therefore assumes that pronunciation variants are only poorly described by means of phoneme substitution, deletion, and insertion. She proposes that the use of linguistically motivated distinctive features could provide the necessary granularity to better deal with pronunciation variants by using context dependent rules that describe the value changes of features.

Coarticulation and assimilation had been identified as a major source of variability in the speech signal long before that time and a phone recognizer was built based on the detection of place and manner of articulation in an intermediate "Articulatory Feature Vector" level [Sch89]. Overlapping articulatory features are used in [EF96] in an HMM-based recognizer. [RBD03] extends this approach to using diphones in the so-called "Hidden Articulator Markov model" (HAMM). In this approach, articulatory states are factorized into different parallel HMMs, which are synchronized at the diphone boundaries. While the HAMM performs worse than the baseline phone HMM, combining the two at the log-likelihood level improves word error rate.

Kirchhoff [Kir00] also acknowledges that it is easier to model pronunciation variants with the help of articulatory features. She points out that articulatory features exhibit a dual nature because they have a relation to the speech signal as well as to higher-level linguistic units. Furthermore, since a feature often is common to multiple phonemes, training data is better shared for features than for phonemes. Also for AF detection fewer classes have to be distinguished (e.g. binary features). Therefore statistical models can be trained more robustly for articulatory features than for phonemes. Consequently feature recognition rates frequently outperform phoneme recognition rates.

Another reason for the poor performance of automatic speech recognition systems on spontaneous speech is the increased occurrence of coarticulation effects as compared to planned or read speech. In [Kir98], Kirchhoff makes the assumption that coarticulation can be modelled more robustly in the production based domain than in the acoustic one. She also assumes articulatory features are more robust toward cross speaker variation and signal distortions such as additive noise. Kirchhoff developed in her thesis [Kir99] an approach using articulatory information for robust speech recognition. She used neural networks to classify attributes and a second classifier to combine the attribute scores to a phone score. Furthermore, these scores can be combined on the HMM state level with a traditional system [KFS00].

Wester, Chang, and Greenberg [CWG05] suggest that corpora are optimally annotated at the articulatory-acoustic feature level. They argue that the transformation from AF to phonetic segments does not transport sufficient detail and richness common to the speech signal at the phonetic level. This work extends to a more general approach integrating information about syllables, articulatory features, as well as stress accent in a “syllable-centric multi-tier model of speech recognition” [Cha02]. Methods for deriving the needed information from the audio signal are developed and improvements are shown on a limited-vocabulary task.

Glass proposes another model for segment-based speech recognition *Glass* [Gla03]. Here, decoding is done on a-posteriori probabilities derived from a segment (feature) sequence, which is a subset of all possible feature vectors in the total observation space, which consists of a graph of features instead of a sequence of frames.

Lee [Lee04] suggests a “knowledge-rich” paradigm to ASR, which makes it possible to include different speech event detectors [LL05] into ASR.

Landmark-based ASR [SMSHL92, Liu96], in which phones are replaced by times where the acoustic manifestations of linguistically motivated distinctive features are most salient, which can be binary and sparse, is both linguistically motivated [Ste02] and has recently received increased attention in the form of the 2004 Summer Workshop at Johns-Hopkins [HJa05].

Another approach is followed by *Reetz* in [Ree00], where features detected from the signal are directly converted into a lexical representation by using a ternary classifier.

4.3.4 Articulation-model-based Acoustic Modeling

Work on estimation of Vocal Tract Shapes/ Articulatory Trajectories from Acoustic Data or actual "inversion":

Dusan is working toward incorporating phonetic and phonological knowledge in Speech Inversion [Dus01]. As the acoustic-to-articulatory mapping is essentially a one-to-many relationship, phonological constraints are used to restrict the trajectories to realistic values using an extended Kalman filter.

The same problem is tackled in [RA97] by using a constrained form of a HMM to attain a smooth and slow trajectory.

Deng [DS94, Den97] sees "residual" variability in speech that is difficult to explain in terms of general properties as the main obstacle in achieving a high word recognition accuracy. He argues that today's speech recognition systems make use of statistical methods and automatic learning procedures in order to model speech at a detailed level because of a lack of reliable speech knowledge. He proposes to use constellations of overlapping articulatory features as speech units that should be able to model these variations in speech incorporating all necessary contextual information. At the same time the number of units is small enough as not to demand too high an amount of training data.

In [Den98] Deng developed a framework based on neural networks and extended Kalman Filter. The Kalman filter was used to model the temporal structure of speech units while the neural network induced a nonlinearity in the system. In the same work, he proposed the concept of trended HMM, whereby polynomials serve as trend functions describing the temporal structure of vocal tract resonances.

Recent work by *Livescu* [LGB03, LG04] develops a feature-based pronunciation model, which realizes an explicit representation of the evolution of multiple linguistic feature streams using Dynamic Bayesian Networks. Pronunciation variation is viewed as the result of asynchrony between features and changes in feature values, which can be learned from data. However, the benefits of this modeling approach could only be shown on feature values derived from annotations and not from real data.

Blackburn [Bla96] describes the design and implementation of a self-organizing articulatory speech production model which incorporates production-based knowledge into the recognition framework. By using an explicit time-domain articulatory model of the mechanisms of co-articulation, it obtains a more accurate model of contextual effects in the acoustic signal, while using fewer parameters than traditional acoustically-driven approaches, although the system employs separate articulatory and acoustic models.

Tang, Seneff, and *Zue* [TSZ03] model manner and place of articulation separately for sub-word units. The multi-stage configuration permits comparing early, intermediate, and late integration of different information sources. However, they do not find significant differences for these kinds of integration.

Juneja and colleagues also developed a speech recognizer based on phonetic features and acoustic landmarks [Jun04]. Their approach uses a probabilistic phonetic feature hierarchy and support vector machines (SVMs) to classify input speech into five classes and outperforms a context-independent HMM based on an MFCC front end [JEW03]. The input to the SVMs consists of acoustic parameters like zero crossing rate, formant average frequency, energy in frequency band, etc. Other detectors were proposed and evaluated, semi-vowels for example are discussed in [EW94]. The "Lexical Access From Features" project's detection module for nasal sounds is described in [Che00]. SVMs are also employed for stop consonant detection using energy and spectral flatness features in [NBR99].

Most of the approaches presented in this class however could only be tested on small corpora or in N-best list rescoring experiments due to their model complexity.

4.4 Physical Measurements

There have been a number of studies which have investigated the potential of directly measured speech production parameters to improve the accuracy of ASR systems. The MOCHA ("Multi-CHannel Articulatory") database [Wre00] contains actual articulatory measurements which could be used for verification of articulatory properties derived from speech or for speech recognition experiments on articulatory data. This database contains (1) Acoustic Speech Waveforms, (2) Laryngograph Waveforms, (3) Electromagnetic Articulograph, (4) Electropalatograph Frames, and (5) Labeled data for 460 sentences from 2 speakers, although more are planned. In [WR00] *Wrench* shows that the measured articulatory information can be used to improve speech recognition by integrating it with acoustic features using LDA. However, the authors report that "preliminary attempts to estimate the articulatory data from the acoustic signal and use this to supplement the acoustic input have not yielded any significant improvement in phone accuracy." There also exist several non-public data sets with similar specifications.

The authors of [KTFR00] worked on both detection of articulatory properties from acoustic evidence only and on real physical measurements, the paper also presents a brief overview of other related work.

Small command and control applications can be mastered using surface electro-myographic readings of muscular activity when speaking silently alone [MHMSW05]. This approach does not try to identify articulatory features (i.e. rounded lips or opening of lips) as such, instead it measures the muscular activity needed to move the articulators into that position. Recent work on whispered speech uses data collected through a throat microphone and also reports improved recognition through the use of Articulatory Features [JSW05].

Other examples that combine acoustic measurements with actual physical measurements to improve automatic speech recognition can be found in [PHT+92, MND04].

4.5 Combination of several Information Sources for ASR

Having several independent information sources available for a particular decision allows us to reduce the error in that decision-making process. Several approaches exist to combine information sources in the speech-to-text process, although the information sources are usually hardly genuinely independent:

Feature fusion combines different feature streams and constructs a common classifier. This approach allows for a simple and efficient construction of classifiers. Most modern speech recognizers use this technique when incorporating a feature context window for the final feature $o_t = [\vec{\omega}_{t-n}, \vec{\omega}_{t-n+1}, \ldots, \vec{\omega}_t, \ldots, \vec{\omega}_{t+n}]$ although the MFCC spectra $\vec{\omega}_t$ can hardly be regarded as independent; other recent examples include [ZSN05] and [Li05], which also deals with other combination approaches. Also, combining the signal from several microphones into one audio signal through beam-forming to reduce background noise could be regarded as "feature fusion".

Classifier combination attempts to build dedicated classifiers for each feature stream and combines the probabilities or likelihoods during search. The stream approach presented here uses this approach. Most audio-visual work prefers this approach over the feature fusion approach [PNIH01] because of its flexibility and quality. Some approaches also decouple streams by allowing for a slight amount of asynchrony between streams.

Decision Fusion builds dedicated recognizers for each information source and combines the resulting hypotheses by some suitable algorithm. This leads to good results, but is only a viable solution if the different streams or classifiers produce hypotheses of comparable quality, which usually requires models of comparable complexity, which is not the case in the "asymmetric" stream architecture presented here. Examples for this approach include the ROVER algorithm [Fis97] and Confusion Network Combination (CNC) [SYM+04].

Model Changes aim to use a model in the first place, which allows to treat several input streams properly and handles dependencies between them accordingly. An example for such a framework, which is also used in speech recognition, are Dynamic Bayesian Networks [ZR98]. However, these approaches typically have high computational demands.

A currently debated problem is how to generate different features, classifiers, or hypotheses by running several independent recognizers for example by varying the context decision tree or phone set [LG05, SKM$^+$05]. The stream approach presented in this work can be classified as "classifier combination".

4.5.1 Audio-visual Speech Recognition

Although humans can understand speech without seeing their partner, they make use of additional information such as visual cues, when they have the opportunity to see their partner speaking. In fact, humans tend to rely more on visual cues, when acoustic communication is difficult, for example in noisy environments. There is a wealth of literature on audio-visual speech recognition [DMW94, PBBB88], including a comparison of human and automatic large-vocabulary audio-visual speech recognition [PNIH01], a comparison of different weight estimation schemes [GAPN02], discriminative weighting of information sources [PG98], asynchronous streams [GPN02]. At this time, model fusion seems to perform better on audio-visual data then feature fusion (early integration) or hypothesis fusion (late integration) [GSBB04].

4.5.2 Multi-Stream Models

Many researchers have explored the potential of multi-stream speech recognition [JEM99]. In most cases, the motivation comes from either combining information from different spectral ranges [BDR96] to improve noise robustness or from being able to combine different acoustic models, different time scales or a combination of both.

Estimation of stream weights has mostly been based on the ML [RW94, Her97], MMI [EB00], Maximum a Posteriori (MAP) [MHB01], Maximum Entropy (ME) [SISHB04] or directly on Minimum Classification Error (MCE) [GAPN02] criteria.

Asynchrony between streams was explored by *Mirghafori* [MM99] with limited success; other work [NY02] reports improvements on a isolated word task using an approach trying to approximate "loosely couple articulators".

4.5.3 Classifier Combination

The structure of Equation 3.1 implies that instead of two knowledge sources, language model and acoustic model, there could be more, presumably independent, knowledge sources to be taken into account. Several approaches have therefore undertaken to create a unified framework to integrating several knowledge sources [Bey00], be they language models or acoustic models, or even further side information [Ver00], into the recognition process.

This is frequently achieved using log-linear interpolation, as one can write

$$p(o|\Psi) := C \prod_i p_i(o|\Psi_i)^{\lambda_i} \tag{4.1}$$

using a normalization constant C. In log-space, the above multiplication of exponentially weighted terms simplifies to a linearly weighted sum, which is easy to compute for every state. The $p_i(o|\Psi_i)$ can be N different independent knowledge sources and the combination of the classifiers is achieved by choosing the λ_i appropriately. In some cases, knowledge source integration is also achieved by rescoring lattices from a standard recognizer with other, possibly non-local, information [LTL05]. As in Equation 4.1 $p(o|\Psi)$ is written as a probability density function (PDF), it is important to notice that the exponential weighting destroys the normalization property of the individual p_i, even if $\sum_i \lambda_i = 1$, which makes re-normalization through C necessary.

4.5.4 Dynamic Bayesian Networks

A Bayesian network is a general way of representing joint probability distributions with the chain rule and conditional independence assumptions. The advantage of the Bayesian network framework over HMMs is that it permits for an arbitrary set of hidden variables s, with arbitrary conditional independence assumptions. If the conditional independence assumptions result in a sparse network, this may result in an exponential decrease in the number of parameters required to represent a probability distribution. Often there is a concomitant decrease in the computational load [ZR98].

Recent advances in inference and learning of DBNs allow using in real-world applications and it is therefore not surprising that many researchers are using the extra modeling power stemming from the factored state representation enabled by DBNs to model articulatory, or pseudo-articulatory processes, in ASR: *Wester, Frankel* and *King* [FK05, WFK04] describe a training scheme, which allows to learn a DBN recognizer for articulatory feature on asynchronous labels, where supported by the data. This results in a more structured DBN, which results in less feature combinations in the recognition output. In addition, this recognizer performs better than a Neural Network.

Some of this work has already been discussed in Section 4.3. The main pitfall of these approaches is that they are usually still too complex to be tractable on large tasks.

4.6 Speaker Verification and Language Identification

While speech recognition aims to build statistic models for speech, which focus on speaker (and language) invariant properties as much as possible, techniques developed for these purposes can also be applied to investigate inter-class variability in order to identify speakers or languages.

In the simplest case, a Language Identification (LID) system evaluates the output of speech recognizers specialized for specific languages, assigning the language to the recognizer which produces the highest confidence output [MKS+00] or the best acoustic likelihood [LDGP94, MC92].

Information on articulatory properties derived solely from the acoustic signal is also valuable to identify speakers or languages. For language identification, an approach based on n-gram modeling of parallel streams of articulatory features has shown better performance on shorter test signals compared to baseline systems based on statistical modeling of phone sequences extracted from the speech signal [PK03].

Speaker Verification (SV) using articulatory features has been demonstrated in [LMSK05], the same authors have also investigated phone-level confidence measures using articulatory features [LS03].

Chapter 5

Detecting Articulatory Features from Speech

The aim of this research is to incorporate the concept of articulatory features into a speech recognition system. A first step in that direction is to build dedicated "detectors" for these features in order to examine whether it is possible to reliably extract the feature information from the acoustic signal. As we are expecting articulatory properties to be portable across languages, we did also perform multi- and cross-lingual experiments at this stage.

By "detector", we mean acoustic models which can be used to classify a given speech frame as either "feature present" or "feature absent" by comparing the class-conditioned probabilities $p(o|a)$ for the feature attributes a, or the associated likelihoods.

As the goal of this work however is not to use the detectors for feature classification on a per-frame basis, results in this chapter only serve to verify our assumptions that

- (Pseudo-)articulatory features can be detected robustly from speech.
- Detectors for articulatory features can be transferred across languages.
- Articulatory features can be detected for different speaking styles, including hyper-articulated speech.

As a case study on how speaking style influences articulatory features as we are using them in this work, we present an analysis of AFs on hyper-articulated speech, where we find that the changes occurring when switching to a hyper-articulated speaking mode do not affect phones, but the feature needed to distinguish between the words the speaker wants to discriminate (see Section 5.3).

5.1 Model Training for Articulatory Features

Detectors for articulatory features can be built in exactly the same way as acoustic models for existing speech recognizers. Using time alignments from an existing speech recognition system, we separated the training data into "feature present" and "feature absent" regions for every articulatory property we are interested in and trained acoustic models using MLE estimation as described in Section 3.6. We trained our models on the *middle* states of every phone only, assuming that features such as VOICED would be more pronounced in the middle of a phone than at the beginning or the end, where the transition into neighboring, maybe unvoiced, sounds has already begun.

Acoustic models for articulatory feature detectors were trained on the ESST (English Spontaneous Scheduling Task) database collected during the Verbmobil project [WSS+00], phases VM-I and VM-II. It consists of American speakers, who were simulating dialogs to schedule meetings and arrange travel plans to Germany with a business partner. The participants were in separate rooms, talking over a telephone, but could usually see each other. Many also knew their conversation partner.

The ESST dialogs contain a large amount of spontaneous effects (partial words etc.) and also contain a high proportion of foreign (mostly German) proper names (restaurants, businesses, places, ...) pronounced by native American speakers without knowledge of German. Because the ESST data contains foreign words to an unusually high degree, it has been transcribed narrowly, frequently employing "phonetic English paraphrases" in order to facilitate acoustic model training. The hotel "Prinzenhof" for example receives the following "phonetic English" transcriptions in the training data: `Preezenhof, Presenhoff, Prinzenhof, Prinzenhof, Prinzenhoff, Prisenhoff, Prisonhof, Prizenhof, Prizhof, Prosinhof`.

Training data consists of approximately 32h of audio data recorded with 16kHz/ 16bit using high quality close-talking microphones. On the ESST training labels, *begin-, middle-*, and *end-* states represent 32.9%, 34.1%, and 33.0% of phone-labeled data respectively. Even the rarest feature (ALV-FR, 0.4%) could still be trained on 101s (middle states only) of data. Every feature model used 256 Gaussians with diagonal covariance matrices.

The general system setup and the pre-processing of the audio signal is identical to the system used for the experiments on spontaneous speech, which is described in Section 9.1 and Appendix B.2, although the feature detectors used no STC matrix, a different LDA matrix and were evaluated in a 32-dimensional feature space.

The feature detectors were evaluated on two different tasks: "ESST" and "ReadBN". The ESST test set was recorded under the same conditions as the ESST training data and consists of 58 recordings from 16 speakers with a total duration of 2h25. Details are presented in Appendix B.2. ReadBN data consists of 198 sentences from the Broadcast News database, which

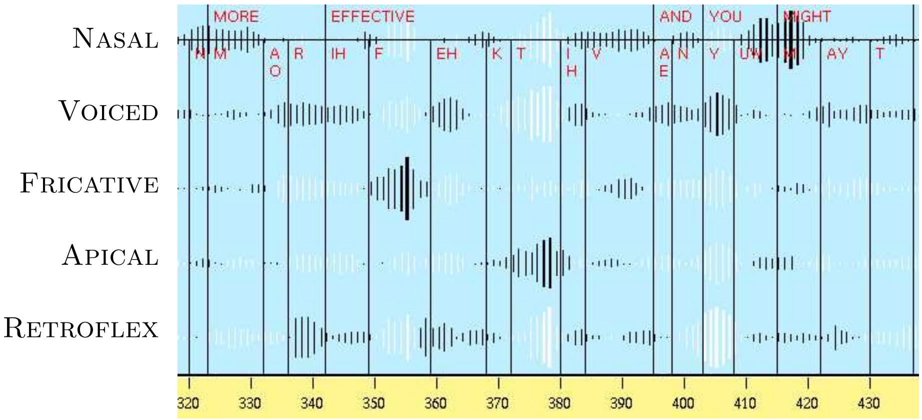


Figure 5.1: Output of the feature detectors for part of the utterance "... be more effective and you might even ..."; black bars mean *feature present* and white bars mean *feature absent*. The height of the bars is proportional to the score difference, i.e. the higher a black (white) bar, the more likely it is that the corresponding feature is present (absent) at this point in time. The numbers at the bottom represent the frame numbers for this excerpt: 1sec = 100 frames.

were re-read in a quiet environment by two speakers, so they are comparable in channel and recording quality to the ESST data, although they are not spontaneous (see Appendix B.3). There is no separate ReadBN training corpus.

The output of some of the feature detectors as used in the classification experiment on ReadBN data is shown in Figure 5.1. It seems that the output of the detectors indeed approximates the canonical feature values quite well, as is also indicated by the classification rates in Table C.1, although various co-articulation effects (e.g. nasalization of /uː/ before /m/) are detected.

The same feature detectors were used to classify the test data into *feature present* and *feature absent* categories on a per-frame basis, by comparing the likelihood scores produced for the test-data, also taking into account a prior value computed on the frequency of features in the training data. The reference for testing was given by the canonical feature values associated with the phonetic label obtained through flexible transcription alignment [FW97a] (Viterbi) using the non-feature baseline system. The results shown in the left two columns of Table C.1 were obtained on ReadBN test data, while the right column was obtained on ESST (spontaneous speech). Overall binary feature classification rates for ReadBN data reach 90.8% on middle states and 87.8% on all states. As begin- and end-states account for about two thirds of all speech data, this means that there is a 50% increase in feature classification error at the beginning and end of phones. As the phonetic alignment however was produced automatically, these numbers can not be

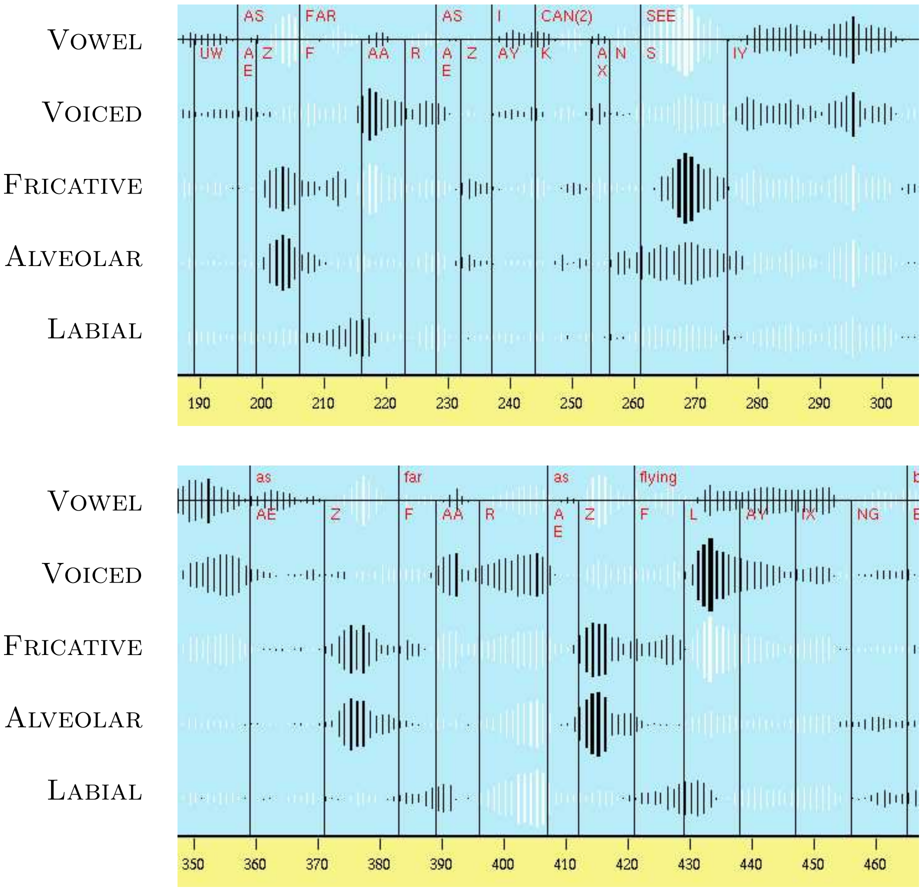


Figure 5.2: Output of the feature detectors for part of the phrase "... as far as ..." in both read speech (top) and spontaneous speech (bottom) from the same speaker. The numbers at the bottom represent the frame numbers for this excerpt: 1sec = 100 frames.

used to compare the accuracy rates presented here with feature detection rates computed on corpora for which detailed annotations at the feature level are available. On ESST data, feature classification accuracy is 87.3% when measured on all states, so that there is no significant degradation between controlled and spontaneous speech, which confirms our impression from visual inspection that feature detection works nearly equally robust for all kinds of speaking styles.

Although not directly comparable, the numbers reported here are in the same range as the results reported in [KT00] for the detection of phonological features using different feature systems on the TIMIT database using neural networks.

Figure 5.2 shows a comparison for an utterance spoken by the same speaker in both controlled mode ("Rob Malkin" in the ReadBN database) and sloppily ("RGM" in the ESST database, testing part). Phone durations are markedly different in spontaneous speech and transitions are less marked, although the output of the feature detectors again seems to be remarkably similar.

5.2 Multi-Lingual Articulatory Features

Next, we built articulatory feature detectors for the five languages Chinese, English, German, Japanese, and Spanish [Stü03, SSMW03, SMSW03] on the GlobalPhone database [SWW97]. These feature detectors were then evaluated on their individual languages as well as on the other four languages in order to investigate the potential of detecting articulatory features across languages.

Using the ML-mix technique [SW01] for language independent acoustic modeling we trained and evaluated a set of multi-lingual detectors, using all possible combinations of the five selected languages.

5.2.1 Mono-Lingual Detectors in Five Languages

In the experiments on the GlobalPhone database we built models for the articulatory features as defined by IPA in the phoneme charts to describe the sounds of human speech (see 2.3), in also adding linguistically motivated questions, that are commonly used during the construction of the decision tree for context-dependent acoustic modeling [FR97].

Every "feature present" and "feature absent" detector was modelled by a mixture of 256 Gaussians. The input vectors for the mixtures were obtained from 13 dimensional mel frequency scaled cepstral coefficients (MFCC) combined with their deltas and delta-deltas, the zero crossing rate of the signal, its power, and the first and second derivative of the power. The resulting 43 dimensional feature vector was then reduced to 32 dimensions using an LDA transformation.

Recognizers based on context dependent sub-phonetic units already existed for the five languages used here. In those recognizers every phoneme is modelled by three states (begin, middle, end). Using these recognizers we produced state alignments of the training and test data on a sub-phonetic level starting from word transcripts.

The first step in training the feature detectors was the calculation of the LDA transformation with the context independent sub-phonetic units as classes. Then the models for the feature detectors were initialized using the k-means algorithm and trained with four iterations of label training. The mapping of the sub-phonetic transcription to the features was done using the IPA table that describes phonemes in terms of articulatory features (see 2.3). For example the phoneme /ɘ/ is attributed with the features Central, Close-Mid, and Un-round. So feature vectors that according

	Test Set				
	CH	EN	GE	JA	SP
CA	93.5%	93.8%	92.9%	95.2%	93.5%

Table 5.1: Average Classification Accuracy (CA) of the AF detectors.

to the transcription belong to /ɘ/ were used to train the present models for Central, Close-Mid, and Un-round, as well as the absent models of all the other features. The feature detectors were only trained with acoustic material that belonged to sub-phonetic middle states. This was done because articulatory features are not static but rather change dynamically. Since we only model abstract classes of articulatory features, we assume that the acoustic data that belongs to middle states is the most representative data for the respective classes.

In addition to the acoustic models for the detectors, we also estimated prior probabilities for the occurrence of the individual features by counting the number of training vectors each model got. Using the acoustic models for the features and the calculated prior probabilities we evaluated the feature detectors by determining their classification accuracy on the development set of their language.

Just as during training, evaluation was performed on the acoustic vectors that, according to the transcription, belong to sub-phonetic middle states. Again, this alignment was automatically generated from the word transcription using phone models. For each test vector every feature was classified into either present or absent. To do so the likelihood score of the absent model was subtracted from the score of the present model and an offset was added that was the difference between the score of the feature present prior probability and the score of the absent prior probability. If the resulting value was below or equal zero the frame was classified as feature present, otherwise as feature absent.

The resulting classification accuracies [SSMW03] averaged over all features are shown in Table 5.1. Detailed results for every single feature can be found in the appendix to [Stü03].

Average classification accuracy is consistently high across all languages. This is consistent with the expectation mentioned in 4.3.3, that statistical models for binary features can be estimated very robustly. The individual results are listed in Appendix C.1, one can see that within a language the classification of the individual features lies roughly in the range from 80% to 99%. On the English GlobalPhone data the classification accuracy for AFs is even higher than for the "ReadBN" data used in Section 5.1 (93.8% vs. 90.8%), we attribute this to the matched conditions for training and test. No experiments were performed with unsupervised training.

5.2.2 Cross-Lingual AF Detection

The next experiment establishes whether articulatory feature detection is robust to inter-language variability. For this purpose we tested each monolingual feature detector on the other four languages that it was not trained on. For this cross-lingual classification we used the prior probabilities that were estimated on the language that the classifiers were trained on. As the GlobalPhone database was recorded under nearly identical conditions for every language, differences can be attributed to language, not channel. Speakers were unique to their language and their training, development or evaluation set.

Table 5.2 shows the results of this evaluation. Every row gives the results of the detectors trained on one of the five language when tested on each of the five languages. The results are averaged over the classification accuracy of the detectors for the individual features. Since not all features of the test set language might be covered by the detectors from the language that is being tested, the classification accuracies could only be averaged over the detectors for features that exist in both, the test and training language. So for example, when testing the Japanese feature detectors on Spanish, we could not determine the classification accuracy for the features Trill, Dental, and Flap. These features are attributed to some Spanish phonemes, however no Japanese phonemes with these features exist, and thus no Japanese feature detectors for them. At the same time there are Japanese feature detectors for Glottal and Uvular. Testing them on the Spanish test set however would only produce false alarms, as these features do not occur with a distinctive function in the Spanish phonemes. Similarly, German consonants are generally aspirated, but as this feature is not used to distinguish two words with different meaning [Hes03, Wie00], i.e. this feature does not form a minimal pair, it is not retained in the German feature set and serves to distinguish regional variants [Wik05]. The amount of false alarms also differs between language pairs and was not measured. The diagonal of the result matrix naturally gives the mono-lingual results mentioned earlier. The detailed results for the individual feature detectors from all languages tested on all languages can be found in Appendix C.1.

As one can see the highest relative drop in average classification accuracy is 11.5%, and occurs when decoding Spanish with Chinese features. The least loss occurs when using English feature detectors to classify the German data. For this constellation the average classification accuracy drops only 4% relative.

However, for every test set there are detectors from languages other than the test language that show a relative increase in performance. For example, the classification error for Central is reduced by 25% relative to 8.4% when using German feature detectors on the English data as opposed to the English detectors' 11.2% (see appendix C.1). Therefore, gains can

	Test Set				
Training	CH	EN	GE	JA	SP
CH	93.5%	87.4%	88.2%	86.5%	83.2%
EN	87.7%	93.8%	89.2%	88.4%	87.9%
GE	88.6%	87.9%	92.9%	86.5%	82.7%
JA	87.1%	87.7%	86.8%	95.2%	87.4%
SP	84.8%	86.4%	83.3%	87.8%	93.5%

Table 5.2: Average classification accuracy of the AF detectors.

	Test Set				
Detector	CH	EN	GE	JA	SP
CH	29	24	23	21	21
EN	24	30	26	24	25
GE	22	26	27	24	24
JA	21	24	24	25	23
SP	21	25	24	23	26

Table 5.3: Number of features shared by different language pairs.

be expected from combining data from different languages to build multilingual speech detectors.

The number of distinctive features shared by different language pairs is shown in Table 5.3. Chinese does not share any of its five tonal features while the European languages German, English, and Spanish share more features among themselves than with Chinese and Japanese.

5.2.3 Multi-Lingual Classification

We trained multi-lingual AF detectors by sharing the training data from n languages to train detectors that are no longer language specific but can be used to detect features in many languages. Since we used the training method "Multi-Lingual Mixed" (see Section 5.2.3) we call a set of feature detectors trained on n languages MMn. If we refer to a set of specific languages that the detectors were trained on, we do so by simply combining the training language identifiers with underscores. E.g. MM3 feature detectors trained on the languages English, German, and Japanese are be called EN_GE_JA detectors.

When training acoustic models with the method "Multi-Lingual Mixed", combining n languages by simply using the training material from all n languages would mean that the available training material would roughly increase n fold. Therefore, in order to ensure that the observed effects do not just occur because of an increase in training material, we only took a fraction of the training material of each involved language depending on how many languages were involved. E.g. for MM AF detectors trained with German and English data we used half of the German training utterances and half of the English.

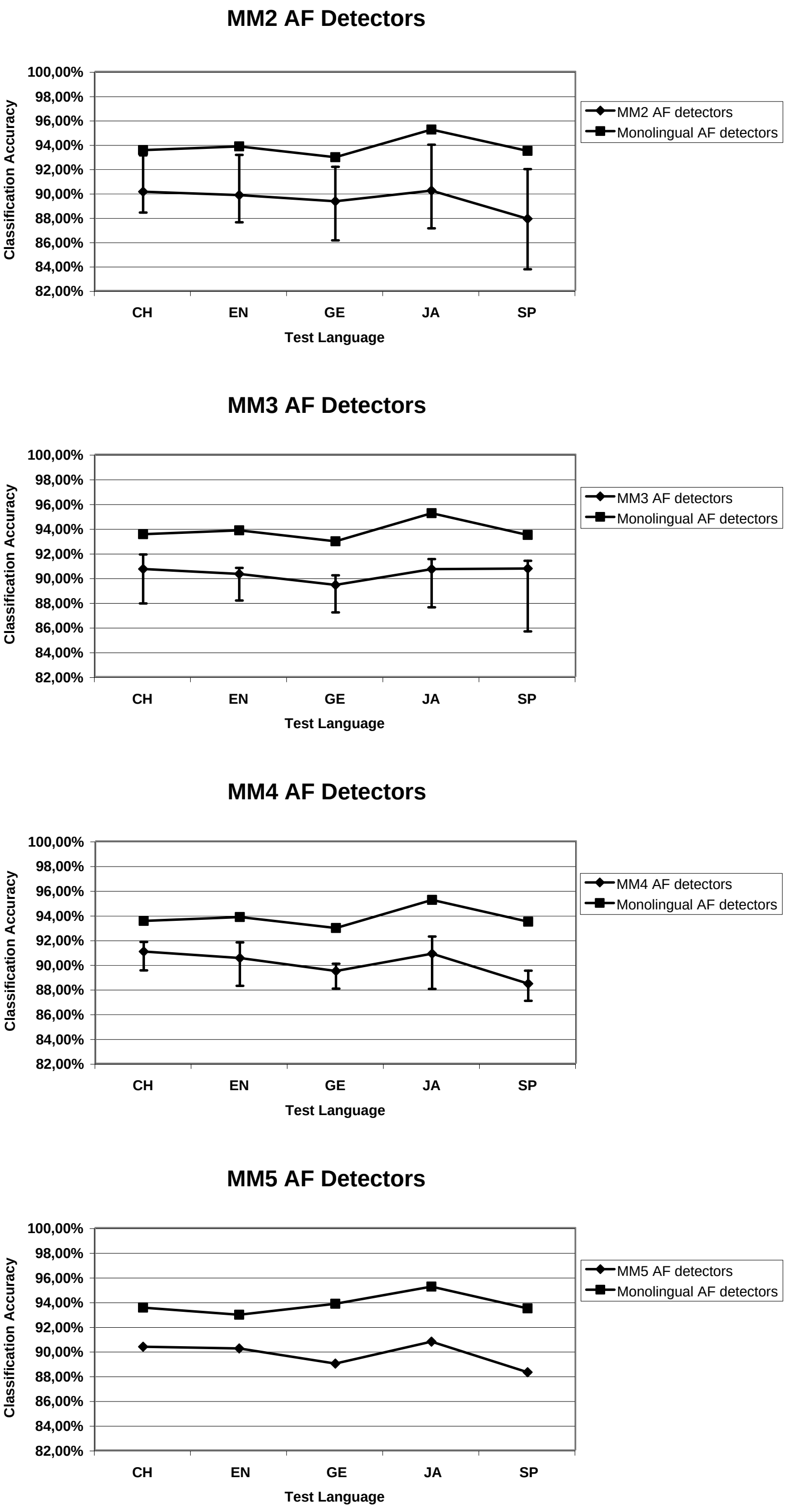
Since we are working on five languages, we can build MM2, MM3, MM4, and MM5 feature detectors. When training on n out of five languages there are $\binom{n}{5}$ possible combinations of languages. In order to explore the multilingual possibilities, we trained all possibilities for combining two to five languages.

Figure 5.3 gives an overview over the performance of the MMn detectors. For every MMn detector the corresponding chart shows the range of the performance of all possible MMn detector sets on all possible test languages compared to the performance of the mono-lingual AF detectors that were trained on the test language. The performance averaged over the individual AF detectors for all possible combinations training data can be found in the appendix of [Stü03]. We can see that if we choose the right combination of languages for a given test set the performance of the MMn detectors is only slightly worse than that of the corresponding mono-lingual ones.

In order to see whether using all available training data instead of just a fraction for training the multi-lingual detectors would improve their performance, we trained the MM5 detector on the complete training data of the five languages. However the evaluation only showed very little absolute improvements of 0.8% on the Chinese test set, 0.2% on English, and 0.2% on Japanese. On the German and Spanish set the performance suffered slightly by just 0.1%. So given the number of parameters of the feature detectors the fraction of training material from the individual languages seems to be sufficient to learn the language dependent properties of the features. This might be an indication that the acoustic manifestation of articulatory features is indeed very similar for different languages, so that there are only few language dependent characteristics in the acoustic signal.

Given the five languages it is also of interest which influence the presence of the test language among the training languages has. Table 5.4 compares the performance of the MM4 detectors that were trained on all four languages except the test language with the performance of the the detectors trained on all five languages (MM5 detectors), thus including the test language. Again there is the problem that not all features of the test language might be covered by the MM4 feature detectors. Therefore the classification accuracy of the MM5 detectors is only averaged over the features of the test language that are also covered by the corresponding MM4 detectors.

As is to be expected the MM5 detectors always outperform the MM4 detectors, since the test language has been seen during training, the difference is however smaller than 5% relative.



MM2 AF Detectors

MM3 AF Detectors

MM4 AF Detectors

MM5 AF Detectors

Figure 5.3: Performance overview of the MMn recognizers (from [Stü03]).

	Test language				
	CH	EN	GE	JA	SP
MM5	90.6%	90.4%	88.9%	90.9%	88.7%
MM4	89.5%	88.3%	88.0%	88.0%	87.1%
Rel. loss	1.6%	2.4%	1.0%	3.2%	1.9%

Table 5.4: Comparison between MM4 detectors that were not trained on the test language and MM5 detectors.

5.3 Articulatory Features as Contrastive Attributes

Hyper-articulation is a particular speaking style which occurs when people want to stress particular aspects of a linguistic message. As such, it is not a global effect and the changes occurring depend on several factors. In the context of a human-machine dialog system, which pretends to confuse two words, a first-order description would predict changes in the articulatory features used to distinguish the phones which are in turn used to distinguish the confused words.

Contrastive Attributes (CAs) [Sol05] can be used for describing changes occurring while disambiguating recognition errors. A CA is an attribute in context of a word error which can be used to discriminate between the true and the recognized token. In a hyper-articulated speaking mode, such a contrastive attribute could be inverted to stress the mis-recognized part of the word. The following example illustrates this process:

Assuming we have the word BITTER (canonically pronounced /bɪtər/) [Cam06]. Let us now suppose that the word BETTER was recognized, e.g. the recognized phone sequence is /betər/.

The difference is the quality of the vowels in the first syllable [Int99], namely

ɪ	the near-close near-front un-rounded vowel
e	the close-mid front un-rounded vowel

In the Janus recognizer lexicon and linguistic question set these words are represented by

```
BITTER  {{B WB} IH T {AXR WB}}  and
BETTER  {{B WB} EH T {AXR WB}}.
```

WB marks a “word boundary” and can be ignored for our purposes. We can use the Articulatory Features (defined as a set of phones, see Chapter B.2)

High-Vow	(IY IH UH UW IX)	and
Mid-Vow	(EH AH AX)	

to distinguish between the two words.

Using contrastive attributes we can now *predict* what kind of changes will occur during hyper-articulation, i.e. when the speaker tries to produce `BITTER` a second time, but this time tries to produce it phonetically distinct from `BETTER`. In order to avoid the mis-recognized word `BETTER`, a hyper-articulated variant of `BITTER` will exhibit *activated* attributes for HIGH-VOW. To demonstrate that the predicted effects actually occur in real utterances, we can look at the output of our feature detectors:

Let an utterance (word sequence) W be represented as a sequence of observable feature vectors $(o_1, o_2, \ldots, o_T)$, where T denotes the length of the utterance in terms of number of frames. The probability density functions for $p(o_t|a)$ are modelled by mixtures of Gaussian densities. The PDFs are used for defining the conditionals for the articulatory attributes a. In the same way, anti-models are available, e.g. $p(o_t|\bar{a})$. The models are trained in a speaker and speaking mode independent fashion as described before. The conditionals are used to define a distance function as shown in Figure 5.1:

$$\Delta(o_t, a) = \log p(o_t|a) - \log p(o_t|\bar{a})$$

The two acoustic signals are shown in Figure 5.4. Figure 5.5 shows two curves: the solid line represents the output of the feature detector for HIGH-VOW for the word `BITTER` in a normal speaking mode. In a hyper-articulated speaking mode, the same word `BITTER` results in the $\Delta(o_t, a)$-curve shown by the dashed line. Both words were uttered by the same speaker. The hyper-articulated variant arose when the speaker tried to resolve the recognition error `BETTER` vs. `BITTER` in the framework of a dialog system [Sol05].

The output of the feature detectors in Figure 5.5 also clearly shows the different lengths of the closure period and the overall longer duration of the hyper-articulated variant of `BITTER`. However, the initial syllable clearly has a higher mid-vowel character for the hyper-articulated variant then for the normal variant. The second syllable seems relatively unaffected. Figure 5.6 on the other hand shows that the peak and area of the feature detector for MID-VOW is hardly influenced by the speaking style. The visible change is mainly due to the longer closure of the plosive.

On the other hand, we can look at what happens when the same speaker tries to disambiguate `BETTER` from `BITTER`: Figure 5.7 shows that in this case also, the MID-VOW feature is stressed for the duration of the first syllable during production of the "stressed" variant.

This example illustrates that our articulatory feature detectors can indeed capture information that humans use to disambiguate words from each other. Computing a Viterbi alignment of `SIL BITTER SIL` and `SIL BETTER SIL` on our four examples (both words produced normally and hyper-articulated) results in the following acoustic scores:

Table 5.5 shows that for the case where the speaker said `BITTER` (columns

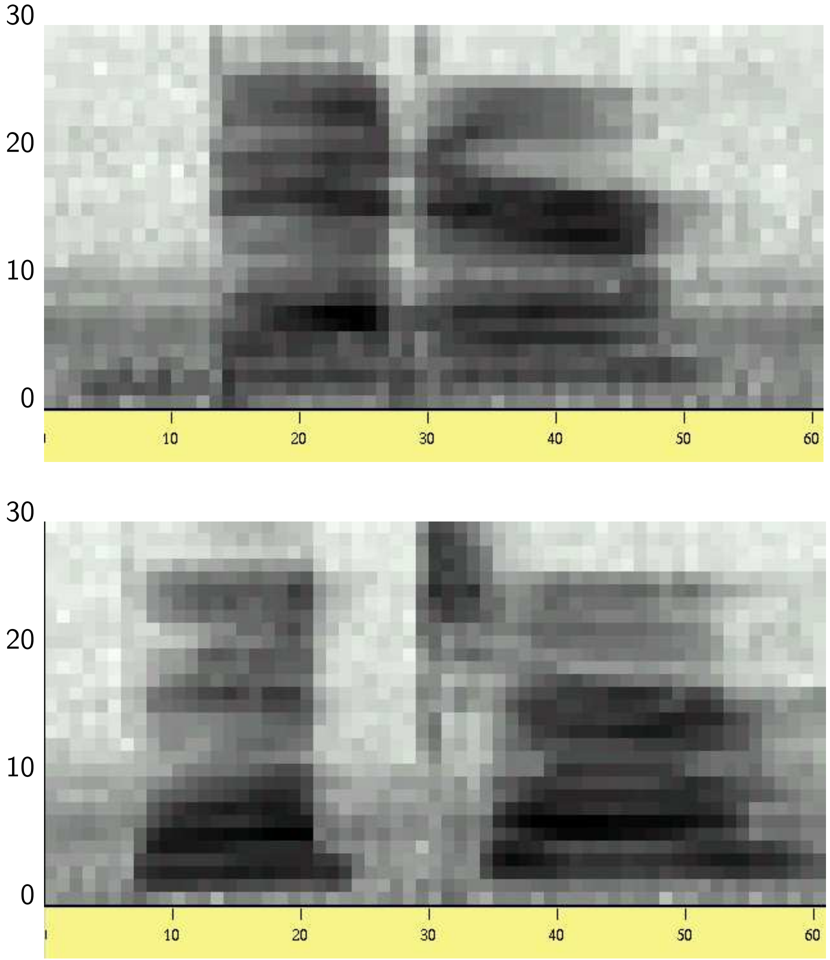


Figure 5.4: Log-MEL features for the word `BITTER` pronounced both "normally" (top) and "stressed" (bottom). It is obvious that the "stressed" version includes a couple of differences, notably the length of the closure and release of the /t/, leading to a larger overall length. Time (horizontal axis) is in frames of .01s, vertical axis shows 30 Log-MEL feature bins spanning a frequency range from 0 to 8kHz.

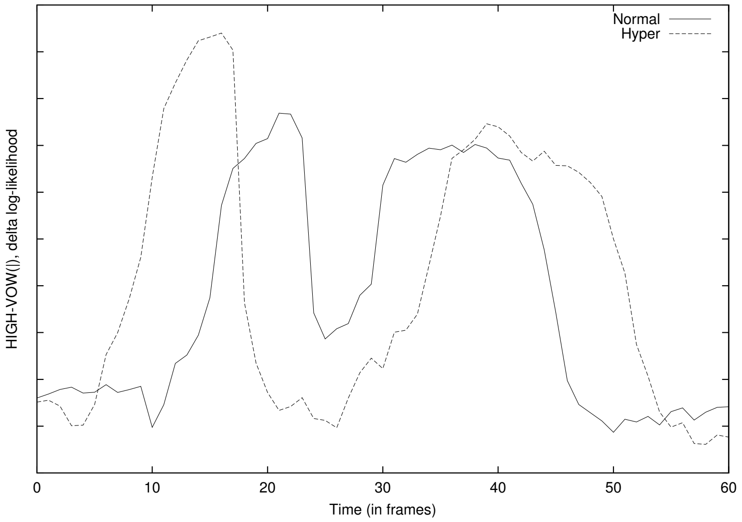


Figure 5.5: $\Delta(o_t, a)$ for attribute High-Vow while pronouncing BITTER, both normally and hyper-articulated.

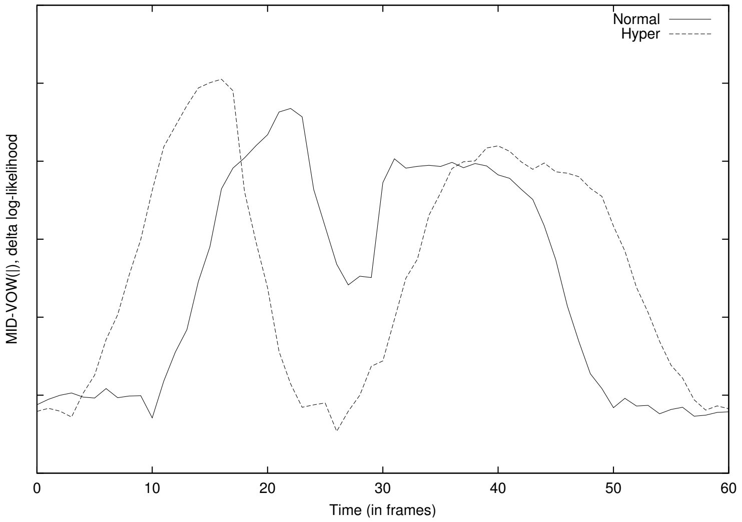


Figure 5.6: $\Delta(o_t, a)$ for attribute MID-Vow while pronouncing BITTER, both normally and hyper-articulated.

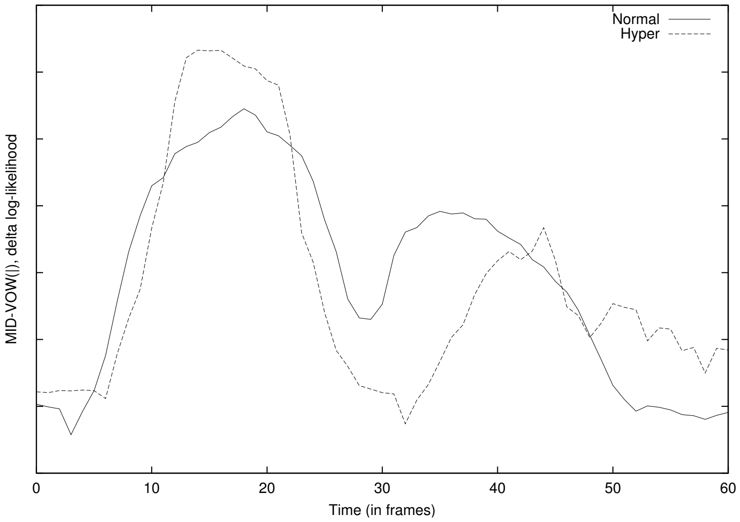


Figure 5.7: $\Delta(o_t, a)$ for attribute MID-VOW while pronouncing BETTER, both normally and hyper-articulated.

	Normal		Hyper-articulated	
Hypothesis	BITTER	BETTER	BITTER	BETTER
BITTER	3.6789	3.830262	3.6386	3.888735
BETTER	3.70339	3.751807	3.648059	3.800349
Margin	0.02449	0.07455	0.009459	0.088386

Table 5.5: Acoustic scores (negative log-likelihoods $\cdot 10^3$) for alignment of two hypotheses on normal and hyper-articulated versions of these two words. "Margin" is the score difference between the correct and the wrong hypothesis.

marked `BITTER`), the score difference deteriorates from 0.024 for the "normal" version to 0.009 for the "hyper-articulated" variant, i.e. contrary to the speaker's intention, the hyper-articulated version of `BITTER` is even more similar to `BETTER`, although the overall score has improved. For the case of `BETTER`, the overall score deteriorates, but the two versions also become more different acoustically (from 0.074 to 0.088).

Therefore, articulatory feature detectors can model the changes occuring when speakers change from normal to hyper-articulated speaking style better than standard acoustic models.

Chapter 6

Including Articulatory Features in HMM based ASR

The previous chapter introduced a method to build dedicated detectors for articulatory features using Gaussian mixture models. The detectors are based on two models with complementary distribution, one for "feature present" and one for "feature absent". This forms a simple binary decision tree, which can also be used in the acoustic model of a speech recognizer. This chapter therefore takes the step from simply detecting features to actively using them in the task of recognizing speech.

The goal of the research in this work is not to build a recognition system solely based on articulatory features. Instead, we concentrate on supporting an existing HMM based recognizer with models for M articulatory features as an additional source of information. Therefore, our approach integrates dedicated detectors for articulatory features with conventional context-dependent sub-phone models, using a stream architecture [MW02]. Although the individual extra classifiers are very simple, they can contribute to an improved overall classification [Kit98], as they are "different" in the sense that they have been trained on different partitions on the training data.

In the taxonomy of approaches to combination of information sources presented in Section 4.5, this "classifier combination" approach avoids the overhead of creating separate hypotheses for the different information sources and having to fuse them "late", while it also avoids the relatively inflexible "feature fusion" approach, which is simple to realize, but very inflexible, as the training and classification occurs on a probability distribution trained jointly over all information sources. It is therefore not possible to change the relative weights of the different information weights at a later stage. The stream approach taken in this work on the other hand, makes it possible

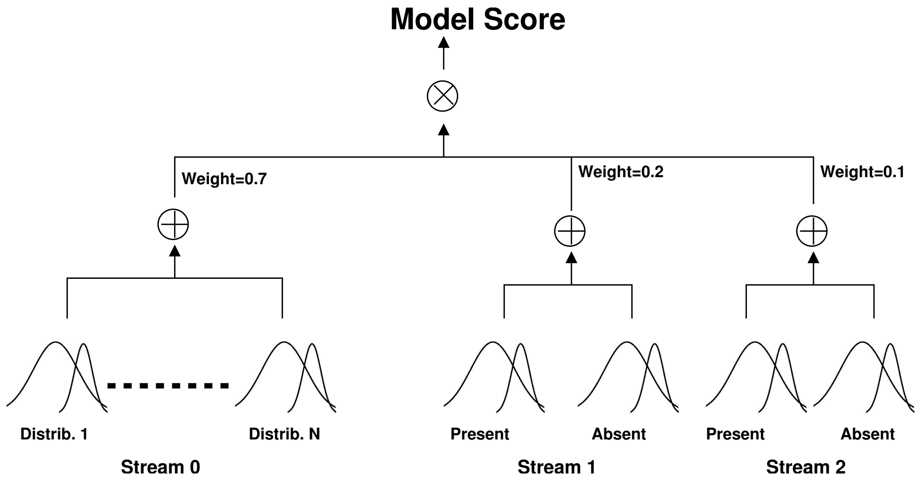


Figure 6.1: Stream setup that combines a “main” stream (“Stream 0”, left) using N context dependent models with $M = 2$ “feature” streams, each containing only two feature “absent” and “present” detectors (neglecting silence and noise models for clarity of presentation). Every stream has a different stream weight λ_i (examples here: 0.7, 0.2, 0.1) for additive combination in log-likelihood space ($\otimes$ symbol). The $\oplus$ symbol represents the selection of exactly one model per state in the decision tree.

to use the relative weighting of different feature streams for adaptation to speaker and speaking style.

6.1 Stream Architecture

Kirchhoff [Kir99] investigated several approaches to combine information about different articulatory features and found the most promising approach to be the combination of scores at the log-likelihood level. After initial experiments with front-end approaches, we therefore used this approach to combine information sources, be they “feature” or “main” stream, too. The conventional models that we use in this research are context dependent subphonetic units that are modeled as a mixture of Gaussians. Because of that, and because of the design of our feature detectors as described in Chapter 5, the acoustic *score* (negative log probability) for a model is now computed as the weighted sum of several Gaussian mixtures models, which represent the standard models and “feature” probability distribution functions. The result is a stream-based architecture which is illustrated in Figure 6.1. The 0-th stream consists of the context dependent standard models. For every articulatory feature that we wish to use, we add an additional stream that contains the “present” and “absent” models for this feature as described in the last chapter. When the decoder now computes the score of a state

s given a feature vector o it adds the score of the corresponding context dependent model from the 0-th stream to the scores from either the "absent" or "present models" from the other streams, depending on whether s is attributed with the respective feature or not. The mapping to determine whether a particular phone is attributed with a feature or not is done according to the linguistic question set used during the construction of the context decision tree, shown in the system descriptions in Appendix B. This question set holds the same information as the IPA chart (see Figure 2.3), but expressed in the phoneme set of a particular recognizer and in some cases contains alterations that were found to be beneficial to ASR performance during the years of development at ISL.

As discussed in Section 2.2, it is possible to obtain a complete description of phones by composing them out of attributes, or features. These attributes can represent multi-value structures such as place and manner of articulation or binary features such as voicing or lip rounding. Still, multi-value attributes can be broken down into sets of binary attributes, e.g. manner of articulation can be described by the binary attributes plosive, nasal, fricative, and approximant. This transformation obviously induces a correlation between the attributes. Switching to binary attributes however creates an unified view of discriminatory effects in an articulatory domain.

Also, in our approach, articulatory attributes are not used to enhance the front-end. If that were the case, the constraints on the human body result in mutual dependence of feature properties which would conflict with the assumption of independent dimensions made for efficient score computation using diagonal covariance modeling, even when semi-tied full covariances or similar approaches were to be used.

The weighted combination of the scores from the HMM based models and the articulatory feature detectors as described above requires the selection of an appropriate set of weights. The weights control the influence that the individual detectors have on calculating the score and thus have a great impact on the search for the best hypothesis. The task is to find an optimal set of weights $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_M)$ that minimizes the word error rate of the recognition system. Weight estimation is being discussed in Chapter 7.

6.2 Mathematical Analysis

In mathematical terms, the state-level combination of acoustic scores in the log-likelihood domain used in this work can be derived from the log-linear interpolation formulated in Equation 4.1. Neglecting the global normalization constant C and going into the logarithmic domain to better match the dynamic range of numerical values encountered, one can write the *score* *function* g as

$$g(o_t|\Lambda, \Gamma) = -\sum_{i=0}^{M} \lambda_i \log p_i(o_t|\Gamma_i)$$

where Γ denotes the parameters of Gaussian mixture densities. As shown in Figure 6.1, Γ_0 consists of the parameters of several thousand context dependent GMMs N_j, while the $\Gamma_{i>0}$ model feature streams which only have GMMs N_a and $N_{\bar{a}}$ for *feature present* and *feature absent*. As mentioned before, introducing weighting factors λ_i manipulates the probability mass:

$$\int \sum_i p_i(o|\Lambda_i)^{\lambda_i} dx \neq 1$$

Introducing constraints, such as $\sum_i \lambda_i^K = L$ with constants K and L as suggested in [Her97] does not solve that problem. In fact, the function $g(o_t|\Lambda, \Gamma)$ is also not a probability density function (PDF) in the log domain. There are two components in a speech recognizer where the loss of normalization might have consequences:

From a decoding point of view, the Viterbi algorithm attempts to find the best hypothesis with respect to the acoustic and language models. In general, it does not matter if the scores rely on a PDF or not. Independent from the optimization criterion, the decoder searches for the word sequence with the best score.

From a training point of view, the acoustic model parameters Γ_i in the individual streams can be estimated by optimizing the ML criterion since the conditionals $p_i(o_t|\Gamma_i)$ are valid PDFs. On the other hand, the weighting factors λ_i cannot be estimated by maximizing the training likelihood without further constraints [Her97], which is why we chose to work with a discriminative criterion instead of introducing artificial constraints on the λ_i apart from a normalization requirement $\sum_i \lambda_i = 1$ to ensure the comparability of acoustic scores during search.

Section 7 discusses these problems in detail.

6.3 HMM Topology and Decision Trees

The formalism presented in the previous section allows combining several acoustic models into a single acoustic score. We therefore still need to define more formally, which models to combine in order to compute a score for a specific state or state sequence.

The acoustic models used in this work are tri-state left-to-right HMMs as shown in Figure 3.3. The acoustic model to be used for a given state is determined by evaluating a context decision tree containing the following questions:

1. Type of phone HMM state (*begin, middle*, or *end*)

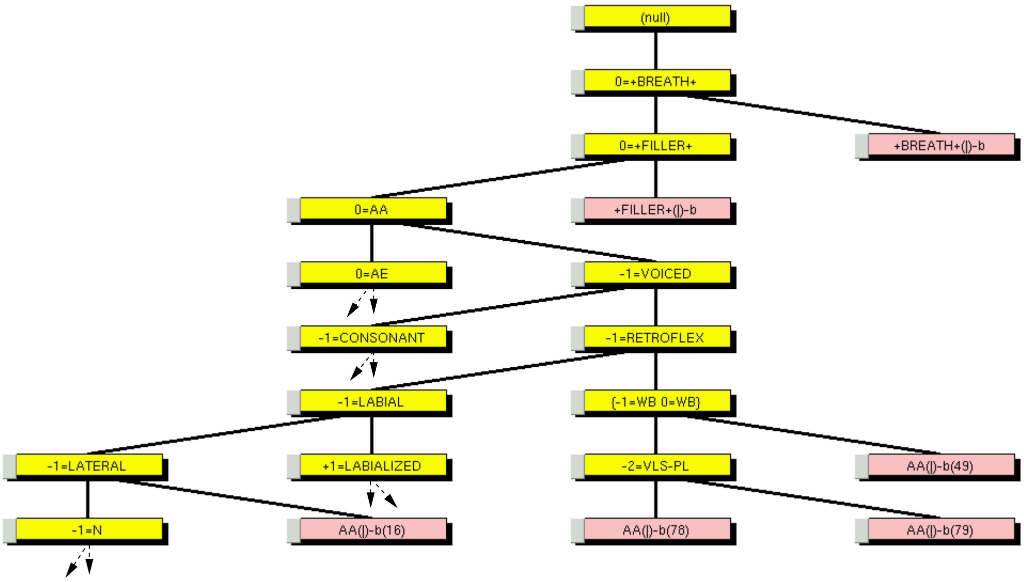


Figure 6.2: Top nodes of ESST phonetic context decision tree for *begin* states: YES answers go to the right, NO answers to the left. We see context-independent "noise" models, then questions for phone identity, linguistic class, and tags. Root node is marked "null", leafs (acoustic models) are shown in pink, tree nodes are shown in yellow.

2. Phone identity (e.g. /a/, /k/)

3. Phone identity of neighboring phones (context of ±2)

4. Tags of neighboring phones (only "word boundary" `WB` currently used)

5. Linguistic classes (Labial, Vowel) of neighboring phones (context of ±2)

The first two questions are always positioned directly after the root node of the tree, questions of type 3 and 4 have multiple occurrences, which are determined on the training data using a divisive clustering scheme based on an entropy or likelihood criterion, typical systems employ several thousand context dependent models.

Silence and noises (see Appendix B) are not treated in a context-dependent way. The first few nodes of the ESST tree for *begin* states are shown in Figure 6.2. By contrast, the complete decision tree for the Syllabic feature of the ESST system is shown in Figure 6.3: the acoustic model contains $-\log p(o_t|a)$, the "feature present" model, $-\log p(o_t|\bar{a})$, the "feature absent" model and models for non-phonetic events such as silence and noise. The Syllabic phones are defined in Appendix B.2.

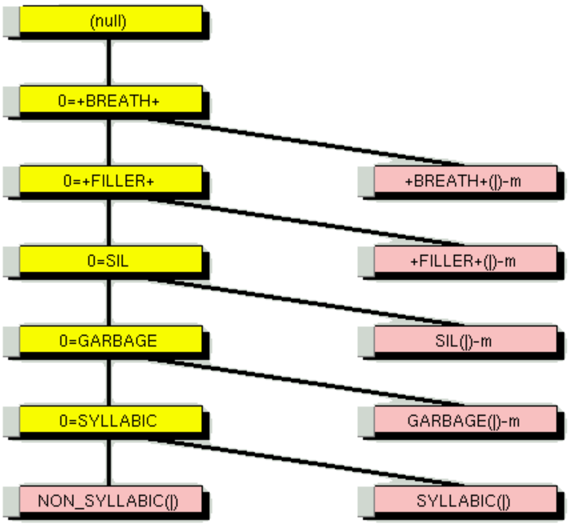


Figure 6.3: Complete ESST decision tree for the SYLLABIC feature. YES answers go to the right, NO answers to the left. The only acoustic models used (apart from dedicated "noise" and "silence" models) are the models for "feature present" (here: `SYLLABIC(|)`) and "feature absent" (here: `NON_SYLLABIC(|)`). Root node is marked "null", leafs (acoustic models) are shown in pink, tree nodes are shown in yellow.

The same decision tree is used for *begin, middle*, and *end* HMM states in the articulatory feature streams $i \neq 0$.

6.4 State Dependent Stream Weights

In the formulation so far, the stream weights λ_i were assumed to be stream-dependent (or "global", G) only, i.e. they were assumed to be equal for all HMM states s or (equivalent) equal for all leafs of the phonetic decision tree used for the "main" stream 0.

To vary the relative weighting of the streams and increase the number of parameters usable for modeling and adaptation, stream weights can be made state dependent (SD), i.e. they can vary depending on base phone identity or phonetic context. This results in a different set of weights $\lambda_{i,s}$ for every context-dependent HMM state s. Independent of the actual estimation method used to determine the stream weights $\lambda_{i,s}$, the phonetic decision tree can also be used to tie states s during re-estimation, to make sure weight updates are performed on sufficient statistics.

To ensure comparability of acoustic scores during search, $\sum_i \lambda_{i,s} = \text{const}$ has to be valid $\forall s$. As JRTk employs a divisive clustering scheme for constructing context-dependent models and uses questions based on features [FGH+97], context dependent stream weights permits modeling for example *voicing* of *end*-states of *unvoiced fricatives* before *vowels* or other related effects, which means the proposed architecture can escape the "beads-on-a-string" problem at the state level.

Chapter 7

Discriminative Combination of Knowledge Sources

The previous chapter presented the stream approach, which allows combining information from different sources in an intuitive and manageable way. The approach introduces a new set of free parameters, the so-called "stream weights" λ_i or $\lambda_{i,s}$ (for context dependent weights).

"Guessing" the weights for the feature streams is naturally unsatisfying since it will most likely provide a solution that is far from optimal. Also the fact that none of the heuristic feature selection methods tested [MW03] seemed to be clearly superior to the others, gives the impression that more improvements can be reached by better ways of selecting the stream weights. It does not seem feasible to apply rules, e.g. obtained from linguistic knowledge, in order to find an optimal set of weights, i.e. one that gives the lowest word error rate. It is therefore desirable to have a data-driven machine learning method that finds a good, if not optimal, weighting of the feature streams. In a first set of experiments, we set the weights globally, i.e. we have the same weight λ_i for a feature i, independent of the acoustic model m evaluated.

In our approach, we do not train acoustic models discriminatively on a large corpus, instead we train acoustic models using the fast and well-understood Maximum Likelihood approach and then combine these models by estimating the combination weights on a relatively small development set. In this sense, our approach can also be interpreted as discriminative adaptation. In the context independent case, we only have a few stream weights to estimate, which ensures stable convergence, while for the more powerful context dependent case more data and careful parameter selection are necessary in order to ensure convergence (see Section 7.4).

The first section of this chapter briefly presents and compares the different discriminative criteria employed in this work, while the following sections discusses them in more detail.

7.1 MLE vs MCE and MMIE criteria

The principle behind “Maximum Likelihood Estimation” (MLE) as presented in Section 3.6 is the optimization of a set of models on the training data by improving the likelihood, i.e. the average expected probability of the models, for every model separately. This approach eventually leads to an optimal Bayesian classifier, but only for the impractical case of having access to unlimited training data. As the goal in practically all ASR tasks is to minimize the word error rate (WER), it would be preferable if one could optimize the models used directly on this optimization criterion, or something more closely related. The popularity of MLE is due to its ability to produce accurate systems that can be quickly trained using the globally convergent Baum-Welch algorithm. Given that MLE’s assumptions are wrong, it is not surprising that it often leads to sub-optimal results and many researchers have employed discriminative criteria directly to acoustic model training [BBdSM86, Nor86, SMMN01, Pov05] and adaptation [PGKW03]. Discriminative training attempts to optimize the correctness of a model by formulating an objective function that in some way penalizes parameter set that are liable to confuse correct and incorrect answers.

In this work, we apply two different discriminative criteria not to the probability density functions themselves, but to the combination process represented by the stream architecture. For the first criterion, “Minimum Word Error Rate” (MWE, closely related to MCE, “Minimum Classification Error”), this has already been developed in the “Discriminative Model Combination” (DMC) approach [Bey00], while the second one can be derived from the same update rules but tries to optimize word posterior probabilities instead. This “Maximum Mutual Information Estimation” (MMIE) approach is much more practical for larger tasks, as it can easily be optimized using word lattices instead of N-best lists. In practice, while MCE works better on smaller tasks, MMIE, or further improvements such as MPE [Pov05], reach equivalent error reductions on more general tasks [SMMN01] while being easier to handle. Here, we show how MMIE stream weight estimation can improve on MCE-based stream weight estimation (i.e. DMC) when using context-dependent stream weights. By setting stream weights at the state level, the importance of individual features in the overall speech model can be set with sub-phonetic resolution, which permits modeling context dependency and asynchronous transitions for articulatory features to a certain degree

7.2 Weight Selection with DMC

First experiments to learn weights for feature streams from data were conducted using the iterative *Discriminative Model Combination* (DMC) algorithm [Bey00]. DMC was developed to integrate multiple acoustic and/ or language models into one log-linear model, i.e. it was used to automatically set the language model weight(s). In this work, we used DMC to optimize the weights of several acoustic models while leaving the language model weight unchanged. Note that only one stream (0) can discriminate between all states, while all the other streams can only discriminate between two classes (e.g. VOICED and UNVOICED), we would therefore expect to work with a high weight for stream 0 and relatively low values for the other "feature" streams, because each of them can only discriminate between some, not all, hypotheses on its own. These decisions however should be more robust than the standard models, because every feature detector is trained on more, shared data.

So, given an hypothesis W, a weight vector Λ and the feature vector o the posterior probability $p_\Psi(W|o)$ is:

$$p_\Psi(W|o) = C(\Lambda, o) \exp \left\{ \sum_i^M \lambda_i \log p_i(W|o) \right\} \tag{7.1}$$

$C(\Lambda, o)$ is a constant necessary for normalization so that $p_\Psi(W|o)$ really is a probability distribution. However since we are only interested in finding the hypothesis W with the highest probability, we ignore C for the sake of simplicity, since it does not depend on W. Note that this simplified formulation no longer represents a probability density function, but simply a "score" function which our system uses to compute similarity measures for acoustic features and we can directly use the acoustic model as an approximation to this distribution, setting $p(W|o) \propto p(o|W)$ [Bey00].

In our special case, with the combination of a standard model stream and the feature detector streams as described above, $p_0(W|o)$ is the posterior probability of W as given by the standard models, while the $p_1, \ldots, p_M$ are the posterior probabilities from the M feature detectors. Our iterative implementation [Bey00] of DMC is based on *Minimization of the Smoothed Word Error Rate (MWE)*, which in turn is based on *Generalized Probabilistic Descent (GPD)* [JCL95]. Similar approaches have been presented in [Ver00].

MWE implements a gradient descent on a numerically estimated and smoothed word error rate function that is dependent on the weight vector Λ for the combination of the models. The estimation of the error function is necessary because the real error function over Λ is not known. Even if the error function were given, since it maps the weight vector Λ, which is defined in $\mathbb{R}^n$, to the number of errors, which is defined in $\mathbb{N}$, the derivative of the function for any Λ would either be undefined or zero. Therefore it is necessary to smooth the empirical approximation of the error function.

The smoothed approximation of the error function that is used for MWE is:

$$E_{\mathrm{MWE}}(\Lambda) = \frac{1}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{W \neq W_n} \mathcal{L}(W, W_n) S(W, n, \Lambda) \tag{7.2}$$

In this equation the $W \neq W_n$ are all possible hypotheses, while the W_n $(n = 1 \dots N)$ are the N given training references for the discriminative training. $\mathcal{L}(W, W_n)$ is the Levenshtein distance. $S(W, n, \Lambda)$ is an indicator function that is used for smoothing the Levenshtein distance. If no smoothing is done, then S would be 1 if W is the hypothesis from the decoder, and 0 otherwise. In order to get a differentiable error function E_{MWE}, S is now set to be:

$$S(W, n, \Lambda) = \frac{p_\Lambda(W|o_n)^\eta}{\sum_{W'} p_\Lambda(W'|o_n)^\eta} \tag{7.3}$$

$p_\Lambda(W|o_n)$ is the posterior probability of hypothesis W, given the set of weights Λ and the internal model of the recognizer, for the feature vector o_n of the n-th training utterance. η determines the amount of smoothing that is done by S. The higher η is the more accurately S describes the decision of the recognizer, and thereby the real error function. However η should not be chosen to be too large, in order to be able to numerically compute S. After initial experiments with several values of η, we used $\eta = 3$.

For the estimation of E_{MWE}, Equation 7.2 and 7.3 take into account all possible hypotheses W. This is clearly not feasible for the numerical computation of E_{MWE}. Therefore the set of hypotheses is limited to the most likeliest ones. In our experiments, we used the hypotheses from an N-best list, where $N = 150$, that resulted from a lattice rescoring.

The derivative of E_{MWE} is now:

$$\frac{\partial}{\partial \lambda_i} E_{\mathrm{MWE}}(\Lambda) = \frac{\eta}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{W \neq W_n} S(W, n, \Lambda) \tilde{\mathcal{L}}(W, n, \Lambda) \log \frac{p_i(W|o_n)}{p_i(W_n|o_n)}$$

where

$$\tilde{\mathcal{L}}(W, n, \Lambda) = \mathcal{L}(W, W_n) - \sum_{W' \neq W_n} S(W', n, \Lambda) \mathcal{L}(W', W_n)$$

With this partial derivative it is now possible to construct a gradient descent:

$$\lambda_j^{(I+1)} = \lambda_j^{(I)} - \frac{\epsilon\eta}{\sum_{n=1}^{N} L_n} \sum_{n=1}^{N} \sum_{W \neq W_n} S(W, n, \Lambda^{(I)}) \tilde{\mathcal{L}}(W, n, \Lambda^{(I)}) \log \frac{p_j(W|o_n)}{p_j(W_n|o_n)}$$

Here ϵ is the learning rate, and has to be chosen carefully in order to adjust the change in the weights per iteration.

Also, we approximated the posterior probabilities with the likelihoods of the hypotheses that were returned by the decoder. Since in the case of the likelihoods the classification rule stays the same as with the posterior probabilities this does not change the update rules for the gradient descent.

7.3 MMIE-based Weight Selection

For tasks which include long utterances or highly spontaneous language, which leads to many similar hypotheses, the N-best lists quickly become very large. Beyerlein set $N = 800$, while in our experiments $N = 150$ gave the best compromise between training time and model power. If however, as is usually the case for systems built with JRTk, many of these hypotheses only differ in one word, which like `A` (pronounced /ə/) and `A(2)` (pronounced /eɪ/) have different lexicon entries, but do not carry different meaning, the list length cannot be increased sufficiently in order to still capture enough variability for discrimination. N-best lists therefore are not a “dense” representation of knowledge gained through the speech recognition process. A better representation is given by word lattices, which are defined as directed, a-cyclic graphs of words as described in Section 3.9. This choice also leads to a new optimization criterion, which can be computed efficiently on lattices.

MMIE (Maximum Mutual Information Estimation) can best be developed from an information theoretic point of view [Bro87]. Given an observation sequence O, a speech recognizer should choose a word sequence W such that there is a minimal amount of uncertainty about the correct answer. While this is still not the same as directly optimizing the word error rate, it is a related principle. In other words, by asking to model the data so that we can pick a hypothesis with “the minimum amount of uncertainty”, we want to minimize the entropy, which is a measure of uncertainty. The entropy of a discrete random variable W is defined as

$$H(W) = -\sum_{w} P(W = w) \log P(W = w)$$

In speech recognition, we therefore want to minimize the conditional entropy H of W given O

$$H_\Psi(W|O) = -\sum_{w,o} P(W = w, O = o) \log P_\Psi(W = w|O = o)$$

which gives the uncertainty in the random event W (the word hypothesis) given another random event O (our observation). The subscript Ψ denotes the dependence on the model parameters.

From this equation, it is easy to see that by minimizing the conditional entropy, the probability of the word sequence given the observation must increase. In speech recognition, this corresponds to the uncertainty in knowing what words were spoken given access to the alternative hypotheses present in O. The amount of information provided by O about W can then be defined as the difference between the two entropies above, i.e. the entropy of W not knowing O minus the conditional entropy of W given O. The mutual information $I(W;O)$ between W and O can now be written as:

$$I(W;O) = H(W) - H(W|O) \quad \text{or} \quad H(W|O) = H(W) - I(W;O)$$

Since $I(W;O) = I(O;W)$ this is known as the *mutual information* between W and O. Thus, if our goal is to minimize $H(W|O)$, then we can try and minimize $H(W)$ or maximize $I(W;O)$, which is the goal of MMIE training. The minimization of $H(W)$ would be called "minimum entropy language modeling", which is a difficult problem as the probabilities of all possible word sequences must be estimated. In this work, as in most other work, the language model which defines $H(W)$ is therefore kept fixed.

Using the expressions for entropy, conditional entropy and the above equations, it can be shown [Bro87] that maximizing the mutual information on a set of observations $O = \{O_1, \ldots, O_R\}$ requires choosing the parameter set Ψ to maximize the function

$$F_{\text{MMIE}}(\Psi) = \sum_{r=1}^{R} \log \frac{p_\Psi(O_r|W_r)P(W_r)}{\sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w})} \tag{7.4}$$

where W_r enumerates the (correct) transcriptions, $P(W)$ is the probability of the word sequence W as determined by the language model, and the denominator $\hat{w}$ sums over all possible word sequences.

To maximize Equation 7.4, the numerator must be increased while the denominator must be decreased. The numerator is identical to the MLE objective function. The difference now is the denominator term, which can be made small by reducing the probabilities of other possible (competing) word sequences. Thus MMIE attempts to both make the correct hypothesis more probable, while at the same time making incorrect hypotheses less probable.

As MMIE estimation uses local (i.e. frame level) posterior probabilities instead of global (sentence level) estimates of word error to update the model, it is possible to compute MMIE updates using word lattices and confidence measures as estimates of the posterior probability. In our work, we use the γ confidence measure [KS97].

Following 7.4 we can now write $F_{\text{MMIE}}(\Psi)$ as a difference of HMM likelihoods [Pov05, SMMN01]:

$$F_{\text{MMIE}} = \sum_{r=1}^{R} \log p_\Psi(O_r|\mathcal{R}_r) - \log p_\Psi(O_r|\mathcal{S}_r)$$

where $\mathcal{R}_r$ represents the Hidden-Markov-Model for the correct transcription of the utterance r and $\mathcal{S}_r$ is an HMM containing all possible transcriptions (for example derived from a decoder lattice) of r. Both encode the full acoustic and language model information used for recognition. Since the "denominator HMM" $\mathcal{S}_r$ includes all possible word sequences (including the correct one), the objective function has a maximum value of zero. Following Equation 7.1, we write the likelihoods as

$$\log p_\Psi(O_{r,t}|s) := \sum_{i=0}^{M} \lambda_i \log p_i(O_{r,t}|s)$$

and taking the partial derivative of this expression we can now write

$$\frac{\partial F}{\partial \lambda_i} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} (\gamma_{r,t}(s; W_r) - \gamma_{r,t}(s)) \log p_i(O_{r,t}|s)$$

Here, we have used Forward-Backward probabilities $\gamma_{r,t}$ [KS97, SMMN01], which can easily be computed from the lattice. s enumerates all states of HMM $\mathcal{S}_r$ and $\gamma_{r,t}$ is an estimate for the a-posteriori probability associated with this state. Now it is straightforward to update the stream weights λ_i using gradient descent with a re-estimation equation of the form $\lambda_i^{(I+1)} = \lambda_i^{(I)} + \epsilon \frac{\partial}{\partial \lambda_i} F(\lambda)$ according to the following rule:

$$\lambda_i^{(I+1)} = \lambda_i^{(I)} + \epsilon(\Phi_i^{\text{NUM}} - \Phi_i^{\text{DEN}}) \tag{7.5}$$

where the statistics Φ can be collected for the numerator or the denominator lattice as

$$\Phi_i^{\text{NUM}} := \sum_{r=1}^{R} \sum_{s \in \mathcal{R}_r} \gamma_{r,t}(s; W_r) \log p_i(O_{r,t}|s)$$
$$\Phi_i^{\text{DEN}} := \sum_{r=1}^{R} \sum_{s \in \mathcal{S}_r} \gamma_{r,t}(s) \log p_i(O_{r,t}|s)$$

In this formula, t is implicitly dependent on s. The main difference is that for Φ_j^{NUM} the sum is over $s \in S_r$, i.e. the reference for utterance r ("numerator lattice") while for Φ_j^{DEN} the index $s \in \mathcal{S}_r$ enumerates all possible HMM states for utterance r ("denominator lattice"). Obviously, $\mathcal{R}_r \subset \mathcal{S}_r$. A detailed derivation can be found in Appendix A.

The simple structure of Equation 7.5 violates the normalization requirement of a probability density function. However, Equation 7.1 already is no PDF and does not need to be. In order to ensure comparable acoustic scores needed with context-dependent stream weights during the beam search, the λ_i can be re-normalized after every iteration of update Equation 7.5 to ensure $\forall s : \sum_i \lambda_{i,s} = \text{const}$.

The update equations presented here do not guarantee convergence of the λ_i to an optimum, however as long as ϵ is small enough, Equation 7.4

will be improved. As we are eventually interested in lowering the word error rate and not the acoustic score or the mutual information, which we still use as our optimality criterion for weights re-estimation, this does not pose a problem in practice. In fact, other work [Pov05] which proves convergence of the mutual information runs into the same problem, as word error rate does not improve for later iterations in MMI training, although the optimality criterion continues to improve monotonically. Here, the mutual information is normally optimized for two iterations only, because word error rate does not improve any further although the F criterion still improves. An example of the convergence behavior of MMI re-estimation of λ_i is shown in Figure 8.1, Figure 9.1 shows an example of the evolution of F_{MMIE} during training.

Using the formulation above, the MMI re-estimation of stream weights can also easily be done separately for different HMM states, i.e. in a context dependent way. In this case there exist different tying and smoothing strategies to improve generalization:

- Tie the statistics for every phone.
- Tie (cluster) the statistics bottom up using the context decision tree, using a minimum count criterion to determine the number of classes. Alternatively, it is also possible to only update the models which have received a minimum count during training/ adaptation.
- Run state dependent MMIE on top of global MMIE, possibly with a smaller step size. This approach resembles annealing strategies in statistical physics.

In this work, the best results were reached with a combination of the second and fourth approach, although no experiments with statistical significance have been conducted.

7.4 Discussion

Most forms of discriminative training criteria suffer from three main problems:

It is difficult to maximize the objective function: the objective functions in discriminative training cannot be optimized using the conventional Baum-Welch algorithm. The only known methods that converge for MMIE and MWE are GPD [JCL95] and the extended Baum-Welch algorithm [GKNN91]. Given the high dimensionality of the parameter space, this may lead to slow convergence and require extensive parameter tuning.

As mentioned in the previous section, virtually all current implementations of MMI training, including this one, do not globally optimize the MMI objective function but instead perform very few gradient descent steps around an optimum found by other optimization functions, typical MLE. In our case, experiments with different initial values for the stream weights and different numbers of iterations confirm that the MMI-based estimation of stream weights is well-behaved in the sense that the outcome does not depend significantly on the initial values and the performance does not degrade for higher iterations, even though it does not improve any more, as long as the step size is reasonably chosen.

It is computationally expensive to maximize the objective function: the expense for computing the MMIE objective function stems from the denominator in 7.4, which requires a summation over all possible word sequences. This amounts to performing recognition on each training or adaptation utterance and for each iteration of training. For the MMIE objective function, lattices can be used, which provide a compact representation of the hypothesis space. Formulations using the MWE or similar objective function are usually based on N-best lists, which cannot handle conversational speech very well.

Poor generalization to unseen data: discriminative training techniques often perform very well on the training data, but fail to generalize well to unseen test data. This effect arises, because Equations 7.4 and 7.3 are dominated by very few paths or from the modeling of globally insignificant data. Optimizing only a few hundred stream weights however does not lead to specialization even for only a few training utterances, as it is very easy to use the context decision tree for parameter tying.

Despite these caveats, discriminative training provides significant gains in many current state-of-the-art speech recognition systems. In contrast to other systems, which use discriminative criteria to update acoustic models directly, the approach presented here uses a discriminative combination of acoustic models. This approach significantly reduces the complexity of the problem to be solved and ensures that robust estimates can be found even on little data, as only very few parameters have to be found.

Chapter 8

Experiments on Multi-lingual Speech

This section presents speech recognition experiments on combining articulatory features with standard acoustic models. The focus is on training of stream weights and selection of features, for which we compare two approaches, Discriminative Model Combination (DMC) and Maximum Mutual Information Estimation (MMIE). These experiments are conducted on the multi-lingual GlobalPhone (GP) data, to investigate the multi-lingual properties of articulatory features. Experiments using DMC were only performed on multi-lingual GlobalPhone [SW01] data, as the generation of N-best lists on spontaneous speech proved impractical on ESST and "Meeting" tasks due to the high number of spontaneous effects (e.g. the frequent occurrence of fillers such as <NOISE>, <AEHM>, etc.) in this data, which each leads to a new entry in the N-best list, but which do not capture any discriminative information. As a consequence, N-best lists have to be very long ($N > 1000$), in order to represent semantically different information, which results in very slow training.

The experiments described in this chapter were performed on the GlobalPhone corpus [SWW97]. The purpose of this corpus is to support multilingual speech recognition research by providing a corpus uniform with respect to acoustic conditions, speaking style, and task in several languages. The main motivation for multilingual speech recognition is the desire to be able to share acoustic training data across languages for training or bootstrapping of recognizers in languages for which no, or very little, training data exists. In order to allow for uniform data to be collected cross languages, texts from international newspapers available on the World Wide Web with national, international political and economic topics were collected. Native speakers read these texts in an otherwise quiet room and were recorded through high-quality close-talking microphones.

For the experiments in this work, we used the Chinese (CH), German

(GE), Japanese (JA), and Spanish (SP) languages from GlobalPhone together with the English (EN) Wall Street Journal ("WSJ0", LDC93S6A) corpus, after which GlobalPhone is modelled. These languages were selected because they display a variety of different characteristics such as the set of sounds and features that they cover, or traits such as tonality [Stü03]. Well-trained baseline systems were also available for these languages. In this thesis, we will present results on an English baseline system being tested on the GlobalPhone Corpus in the following three distinct:

Mono-lingual case: evaluate feature detectors on the same language, on which they were trained, i.e. only use EN feature detectors.

Cross-lingual case: evaluate feature detectors from other languages. In cases where not all features can be used in other languages (e.g. tonality), these were discarded. The MM4 models used for tests with the EN baseline were trained on GE, CH, SP, and JA.

Multi-lingual case: use the feature detectors that were trained on all available languages. In this work we use the MM5 feature detectors trained on CH, EN, GE, JA, and SP.

[Stü03] also evaluates a Chinese baseline system, giving similar results. Appendix B.1 gives an overview of the size of the training, development and evaluation sets for these five languages as well as the size of the English dictionary and language model. Every word in the dictionary is tagged with the language it belongs to, so that it can be distinguished from words in other languages that might share the same orthography. The multi-lingual paradigm is based on the assumption that the articulatory representation of phonemes across different languages is so similar that phonemes can be seen as units independent of the underlying language. Thus the language specific sets of phonemes Υ_{L_i} of languages L_i $(i = 1 \dots n)$ can be combined into a single language independent phoneme set $\Upsilon = \Upsilon_{L_1} \cup \Upsilon_{L_2} \cup \ldots \cup \Upsilon_{L_N}$. This concept had first been proposed by the International Phonetic Association (IPA) [Int99]. Different language independent notation schemes for human sounds exist, such as Sampa [Wel89] or Worldbet [Hie93].

In this work, the definition of the global phoneme set is based on the IPA chart. In this global phoneme set sounds from different languages that share the same IPA symbol share the same unit. The global phoneme set covers 162 symbols taken from twelve languages. 83 of them are shared between languages, while 79 only occur in one language only. The English phone set used in our recognizer is shown in Appendix B.1, more details can be found in [Stü03, SW01].

DMC and MMIE weight estimation on the GlobalPhone corpus are presented in Sections 8.2 and 8.3, a comparison of the two approaches is given in Section 8.4.

System	Meeting	EN (GlobalPhone)
dev	19.6%	13.1%
eval	20.8%	16.1%

Table 8.1: Word error rates on the GlobalPhone development and evaluation sets. The "Meeting" number refers to unadapted "Meeting" acoustic models with GP language model and shows that the GP system is indeed the suitable baseline system for this task.

8.1 Baseline System

For the GlobalPhone baseline system, acoustic models were initialized using a fast and efficient bootstrapping algorithm with the help of a four-lingual phoneme pool [SW97]. The acoustic models for each language consist of a fully continuous HMM system with 3000 quinphone models. Each Gaussian mixture model contains 32 Gaussians with diagonal covariances.

The feature vector is made up of 13 Mel-scale cepstral coefficients plus first and second order derivatives as well as power and zero crossing rate. After cepstral mean subtraction the feature vector is reduced to 32 dimensions by a linear discriminant analysis (LDA). Note that this is the same feature extraction that we used for the training of the articulatory feature detectors.

The sub-polyphone models were created with the use of a decision tree clustering procedure that uses an entropy gain based distance measure defined over the mixture weights of the Gaussians [FR97]. The set of available questions consists of linguistically motivated questions about the phonetic context of a model. English acoustic models were trained with 4 iterations of label training on 15h of training data. The English trigram language model was trained on CSR data, the perplexity is 252 with an OOV rate of 0.1% on the development set using a 9k vocabulary. A summary of the system description is available in Appendix B.1.

The language model parameters used for decoding were optimized on the development sets. Using these parameters, the final evaluation of the recognizers was done on the corresponding evaluation set. Table 8.1 shows the word error rate (WER) for the English recognizers with the optimized language model parameters on their development ("dev") and evaluation set ("eval").

8.2 Experiments using DMC

With the methods for integrating the trained feature detectors with HMM based recognition systems and finding stream weights described in the pre-

English data	EN		GE	
	Dev	Eval	Dev	Eval
Baseline	13.1%	16.1%	13.1%	16.1%
DMC adapted weights	11.7%	14.4%	11.9%	15.1%
Best rel. reduction	10.8%	10.6%	9.2%	6.2%

Table 8.2: WER when decoding English data using AF streams in a mono-lingual (EN) and cross-lingual (GE) scenario and DMC adapted weights.

English data	MM4		MM5	
	Dev	Eval	Dev	Eval
Baseline	13.1%	16.1%	13.1%	16.1%
DMC adapted weights	11.8%	14.8%	11.9%	14.5%
Best rel. reduction	9.9%	8.1%	9.2%	9.9%

Table 8.3: WER on EN data with AF streams in a cross-lingual (MM4) and multi-lingual (MM5) scenario and DMC adapted weights.

vious chapters we performed a series of experiments [SSMW03, SMSW03, Stü03]. Our experiments show that porting AF detectors from one language to another can result in WER reductions comparable to using detectors from the original language.

8.2.1 Decoding using AF and Adapted Stream Weights

Using Discriminative Model Combination (see Section 7.2), we calculated stream weights for the different scenarios as described in the last section using the respective articulatory feature streams. For the calculation of the smoothed word error rate function E_{MWE} the hypotheses from an N-best list were used. The N-best lists contained $N = 150$ hypotheses and was obtained from a lattice rescoring. The smoothing factor was experimentally set to $\eta = 3.0$. Higher η led to numerical instability despite normalizations and double precision calculations due to the high dynamic range of $S(W, n, \Lambda)$, lower values resulted in slow convergence.

The step width ϵ for the gradient descent was selected so that the maximum change of a single stream weight equaled a constant δ. For the mono-lingual case δ was initially set to $\delta = 0.01$; for the cross- and multi-lingual case we chose $\delta = 0.005$. The smaller δ compensates for the higher average scores that the feature detectors gave when used across languages. As soon as the weight estimation was fluctuating for several iterations around a local minimum, δ was decreased, and further iterations were calculated until no further improvements were seen. A maximum of 30 iterations was trained.

More details about the weights training using DMC can be found in [Stü03].

The utterances from the development set served as training set for the DMC. In order to see how well the weights found for the development set generalize, we decoded the evaluation set using the stream weights calculated on the development set.

Mono-lingual case

The mono-lingual case yields error reductions around 10% relative, as shown in column "EN" of Table 8.2. It was possible to heuristically find a set of stream weights, which resulted in the same reduction in word error rate [Stü03], although using less features (these were Postalveolar, Palatal, Glottal, Affricate, Labiodental, Lateral-Approximant, Nasal, Round, and Open). These features have also received high weights during DMC training, Postalveolar being the most important feature and Glottal being the third-important feature (see Table C.9). DMC however does not try to reduce the number of streams to a minimum, so that different weights will arise.

Cross-lingual case

Cross-lingual training of articulatory feature detectors also resulted in a reduction in word error rate. As an example, the results when using German as a second stream are shown in column "GE" of Table 8.2. German was chosen as an example, because German feature detectors were best at classifying English features (see Table 5.2).

Combining English standard models with German feature streams leads to a word error rate of 11.9% on the development set, a relative reduction of 9.2% compared to the baseline. Using the MM4 feature detectors, which were trained on German, Chinese, Japanese, and Spanish, the word error rate was reduced to 11.8%, a relative reduction of 9.9%.

Multi-lingual case

Adapting the weights for the MM5 streams with DMC also showed improvements. On the English development set the word error rate was reduced to 11.9%, a relative reduction of 9.2% (see column "MM5" of Table 8.3). The difference to the cross-lingual case is statistically insignificant.

Complete Detector Set

In Section 5.2.2 we showed that combining the feature detectors from many languages can improve the average classification accuracy, still, it can be better to pick individual feature detectors from a specific language (cross-lingual approach) instead of merging them with data from another language

All AF detectors		
	Dev	Eval
Baseline	13.1%	16.1%
DMC adapted weights	11.5%	14.1%
Best rel. reduction	12.2%	12.4%

Table 8.4: WER when decoding the EN data using all AF detectors as additional knowledge sources (streams) in a multi-lingual scenario and DMC adapted weights.

(multi-lingual approach). In order to see whether it is possible to utilize this effect for the combination of the standard models with the feature detectors we presented the feature detectors from all languages and the standard models from the English recognizer to the DMC.

The results are shown in Table 8.4: it is possible to get a relative reduction in WER of 12.2%. This is the best reduction that we were able to achieve so far using weights trained with Discriminative Model Combination.

8.2.2 Weights Learned

Appendix C.2 shows the feature weights as learned by the DMC for the different combinations of standard models and feature detectors. In these tables only features with a weight greater or equal than 10^{-5} are shown.

For the "complete detector set", only Chinese and Spanish feature detectors are chosen, when the English standard models and the feature detectors from all languages are presented to the DMC. Neither the English detectors, which show the best classification accuracy on English, nor the German detectors, which show the best cross-lingual performance on English, are selected. It seems that Spanish and Chinese detectors provide the "most complementary" information at locations, where the standard models make mistakes.

As shown in Table C.9, DMC usually selects the same features independent on which language(s) they have been trained on (provided they exist in both languages). Among the 24 features that were selected when combining English standard models and English feature detectors, 18 are also among the selected German detectors, 17 among the MM4, and also 17 among the selected MM5 detectors (see Table 5.3). AFFRICATE or GLOTTAL for example receive a high relative probability in all languages, while LATERAL-APPROXIMANT or ALVEOLAR don't help for classification. BILABIAL seems to be a good indicator when trained on EN audio data, while it does not contribute significantly when trained on GE or MM4 (which is CH, GE, JA, and SP). In the MM5 setup, which includes EN training data, its weight however is significantly increased.

8.2.3 Conclusion

Using DMC it is possible to find suitable weights for the stream based approach described in Section 6.1 in a data-driven way for the mono-, cross-, and multi-lingual setup.

The stream weights found on the development data generalize well, as the improvement in word error rate is nearly as high on the evaluation data as on the development data. There is good correlation between features selected from different languages, which indicates that the underlying property, i.e. a language-independent articulatory feature, carries useful information. Only some features, e.g. BILABIAL, seem to be very specific to English, as the English bilabial sounds / p b m w / are indeed produced differently (e.g. with aspiration) in other languages, so that sharing is not possible.

8.3 Experiments using MMIE

Discriminative Model Combination (DMC) based on the MWE (Minimum Word Error) rate criterion is desirable, as it directly optimizes the optimality criterion for speech recognition, the word error rate (WER). Using the settings given in Section 8.2, which were necessary in order to guarantee a stable update, the experiments however required up to 30 (sometimes even more) decoding runs and generation of N-best list (typically: $150 < N < 800$) over the adaptation data in order for the estimation to converge. This approach is therefore not feasible for

- more data, i.e. larger tasks, which increases training time to the order of days even on today's machines. A large part of training time is spent on the alignment and computation of acoustic scores for the N-best lists.

- spontaneous tasks, which increase the required size of N-best lists, because pronunciation variants are used extensively, which do not add any semantic meaning and do not influence the WER. The N-best lists however need to contain semantically different sentences in order to be useful for MWE training.

We first conducted a couple of experiments to generate N-best lists from confusion networks (CNs) [MBS00] instead of lattices and directly work on CNs instead of N-best lists. These present a more compact representation of the hypothesis space and also allow excluding homo-phones, i.e. words that have identical phonetic transcriptions and cannot be distinguished by the acoustic model, from the N-best list generation, but they can only partially alleviate the second problem. Because it can be efficiently computed on lattices, the Maximum Mutual Information (MMI) criterion is better suited for this task, as discussed in Section 7.3.

Iteration	Lattice density			
		dev		eval
	5	10	20	10
0	12.7%	12.7%	12.7%	15.6%
1	12.4%	12.4%	13.0%	14.7%
2	12.3%	12.4%	13.0%	14.3%
3	11.7%	12.3%	13.0%	14.2%
4	11.6%	11.9%	13.0%	14.3%
5	11.5%	11.7%	13.0%	14.1%
6	11.4%	11.6%	13.0%	14.1%
7	11.9%	11.5%	13.0%	14.1%
8	12.2%	11.3%	13.1%	14.4%
9	12.8%	11.7%	13.1%	14.5%
10	13.1%	11.9%	13.1%	14.4%

Table 8.5: WER for global (G) stream weight training on GlobalPhone "dev" and "eval" data. Weights λ_i are carried over from "dev" to "eval". The best relative improvement in word error rate is 11% on "dev" and 8% "eval" for the 8th iteration and $d = 10$.

For comparison between DMC and MMIE, we ran multi-lingual experiments on the English GlobalPhone data using CH and SP feature detectors as in the the multi-lingual DMC setup. This permits comparing the DMC and MMIE criteria on the best-performing setting for DMC training.

The results of a training of global (G, i.e. state independent) stream weights for Spanish and English using the English GlobalPhone system for different lattice densities[1] d are shown in Table 8.5.

The step size was set to $\epsilon = 2 \cdot 10^{-7}$ for the MMI training after initial experiments, the initial stream weight was $\lambda^0_{i \neq 0} = 1 \cdot 10^{-4}$ A further parameter to set is the lattice density d (comparable to the length of the N-best lists), which influences the posterior probabilities γ used during the MMI update. Experiments lead to an optimal value of $d = 10$, which proved to be stable across tasks.

A comparison of performance reached with state independent MMI training on the development and evaluation set is also shown in Figure 8.1. While the performance on the development data increases monotonously up to a certain point, performance on the evaluation data increases also, but reaches saturation earlier and starts to fluctuate. Nonetheless, generalization to the evaluation data is good: on the training set, using the weights trained in the eighth iteration, MMIE stream weight estimation reduces the error rate by

[1] In the Ibis [SMFW02] framework, lattice density is defined as number of nodes measured without language model information, i.e. without linguistic poly-morphism, over the length of the best path.

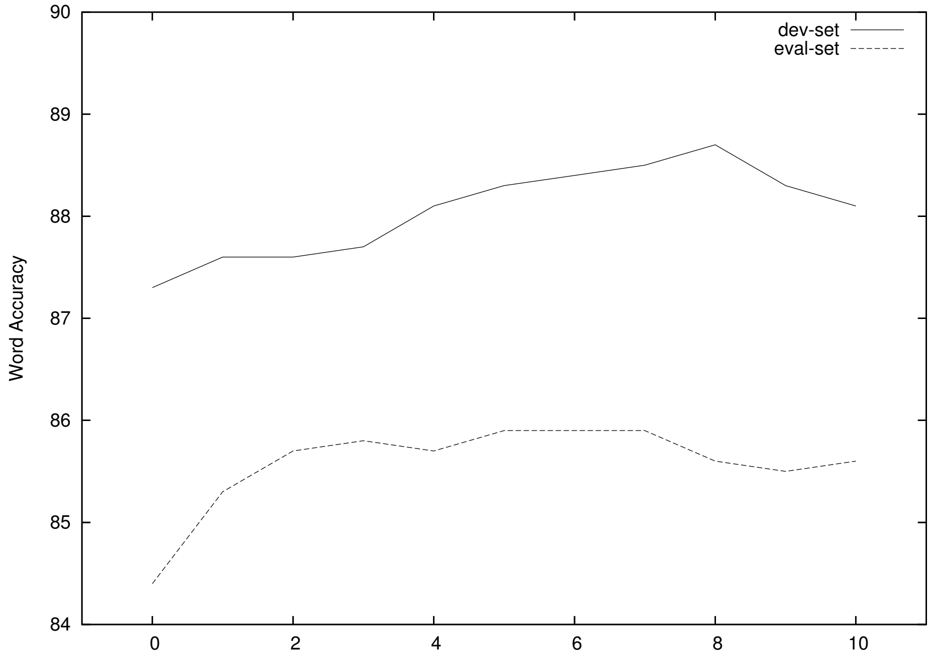


Figure 8.1: Convergence of G-MMI stream weight estimation on GlobalPhone development and evaluation data.

1.4% absolute (which is 11% relative). On the disjoint evaluation data, the error rate is still reduced by 1.2% (8% relative) when using the best parameters as determined on the development set. The decoding experiments using MMI were always run with a wider beam than the DMC experiments, so the correct baseline for the DMC experiments has a WER of 12.7%/ 15.6% (dev/ eval) instead of 13.1%/ 16.1%. Further widening of the beam does not improve the WER.

When performing context dependent training using the MMIE criterion, the algorithm is able to nearly reduce to half the error rate on the development set (from 12.7% to 7.6%) when using state dependent (SD) stream weights as shown in Figure 8.2. While this is of course meaningless in practice, as the improvement does not carry over to the evaluation set at all, it shows the learning capabilities of the algorithm. In this setup, we are adapting $75 \cdot 3'000 \approx 225k$ weight parameters $\lambda_{i,s}$, while the original (main stream) acoustic models contain $6'144k$ parameters and each feature stream uses about $16k$ parameters.

To improve generalization, we reduced the learning rate and set $\epsilon_{SD} = 0.1 \cdot \epsilon_G$, starting the SD training in the spirit of an annealing scheme with the best performing global weights λ_i on the development data (iteration 8 at $d = 10$) and perform one iteration of context dependent training using state tying with a minimum count of 100 for performing a state dependent

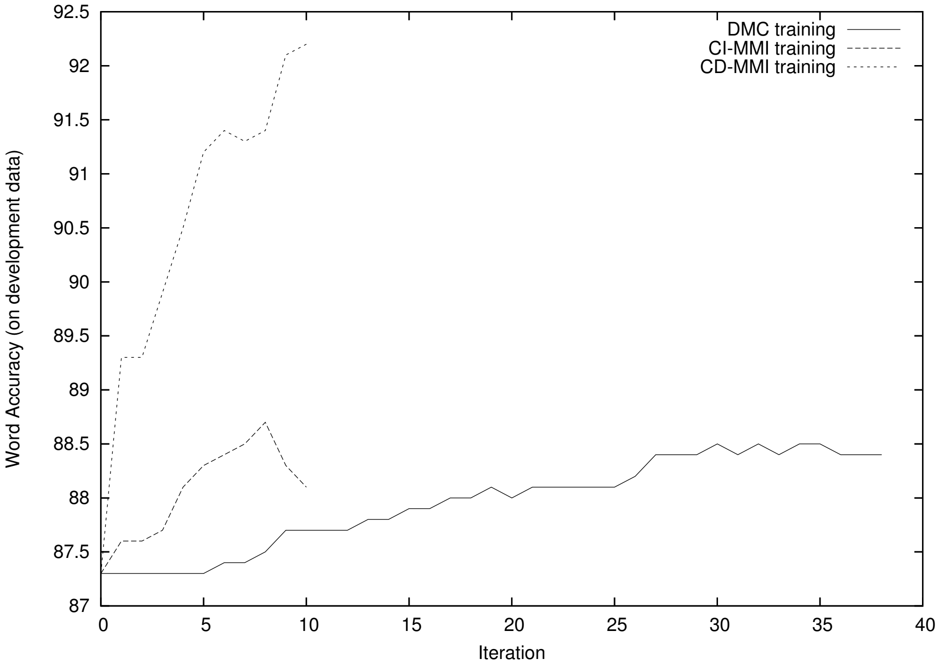


Figure 8.2: Convergence of DMC, G-MMI, and SD-MMI training on GlobalPhone EN development data. DMC needs 30 iterations to reach saturation, MMIE is computationally much less expensive.

update. This results in a word error rate of 10.9% on the development data and 13.9% on the evaluation data. Iterating context dependent training continues to improve performance on the development data, but does not transfer to the evaluation data because of over-specialization.

8.4 Comparison of DMC and MMIE Results

The convergence behavior of DMC, G-MMI, and SD-MMI are plotted in Figure 8.2. G-MMI training can be made to converge on the development and evaluation data by using slightly less aggressive settings. DMC training is numerically instable for more aggressive settings.

The results of DMC-driven and MMIE-driven weight estimation are compared in Table 8.6. While the error reductions are comparable for both DMC and MMIE approaches, DMC reached its maximum after 23 iterations, while MMIE required just 8 iterations to reach an equivalent level of performance. In addition, every MMIE iteration requires much less time because the accumulation of statistics can be done on lattices and a confidence measure instead of an N-best list and word error rates. While no controlled timings

	dev	eval	Iterations
DMC	11.5%	14.1%	30
G-MMIE	11.3%	14.4%	8
SD-MMIE	10.9%	13.9%	8+1

Table 8.6: WER for MMIE and DMC weight estimation. State dependent (SD) MMIE estimation is reported after one iteration on top of G-MMIE.

have been computed, it is clear that MMIE estimation converges faster then DMC estimation.

We can compare the features selected by DMC (see Table C.10) and context-independent MMIE (Table C.11). These seem to be roughly similar, although not strictly identical. The languages English and Spanish get similar average weights, although the weights of individual features vary. For example, NASAL is almost universally switched off or reduced in all languages and for both DMC and MMIE when testing on English, while OPEN receives a higher weight in English.

Chapter 9

Experiments on Spontaneous Speech

To investigate the performance of Articulatory Features on spontaneous speech, we tested the feature detectors built on ESST data on the ESST task. Training of feature weights was achieved with MMIE, which shows comparable performance to DMC while significantly reducing the computational effort. The baseline (i.e. non-AF) system was the best system available to us on the ESST task.

9.1 MMIE Experiments on Spontaneous Speech

The ESST speech data was collected during the Verbmobil project [WSS+00] with close-talking microphones in 16kHz/16bit quality. The participants were in separate rooms, talking over a telephone, but could usually see each other. Many also knew their conversation partner. Training data for the phone models and the non-feature baseline system consisted of 32h from the ESST corpus, which was merged with 66h Broadcast News '96 data, for which speaker labels are available, for robustness. A system trained on ESST only reaches comparable performance on the ESST test set, but performs worse on other data. The system is trained using ML and uses 4000 acoustic models. The parameters of the training and test sets used in this work are shown in Table 9.1. Further details about the system can be found in Appendix B.2. The ESST evaluation data used in this work was recorded during the second phase of the Verbmobil project (VM-II) and is different from the VM-I evaluation data used in other work [Fri00].

The ESST test vocabulary contains 9400 words including pronunciation variants (7100 words without pronunciation variants) while the language model perplexity is 43.5 with an OOV rate of 1%. The language model is a tri-gram model trained on ESST data containing manually annotated semantic classes for most proper names (persons, locations, numbers, etc.).

Data Set	Training		Test		
	BN	ESST	1825	ds2	xv2
Duration	66h	32h	2h25	1h26	0h59
Utterances	22’700	16’400	1’825	1’150	675
Speakers	175	248	16	9	7
Recordings	6’473	2’208	58	32	26

Table 9.1: The English Spontaneous Scheduling Task .

ESST Test Set	1825	ds2	xv2	# Gaussians
WER no LM rescoring	26.3%	25.5%	27.2%	128k
WER baseline	25.0%	24.1%	26.1%	128k
WER 24 Gaussians	25.6%	25.0%	26.3%	96k
WER 44 Gaussians	24.9%	24.4%	25.4%	176k
WER 5.2k models	25.0%	24.3%	25.8%	166k

Table 9.2: Baseline WER on the ESST task using a system trained on ESST and BN ’96, with and without language model rescoring. The 44 Gaussians and 5.2k models systems use the same number of parameters as the AF based system (WER with rescoring).

Generally, systems run in less than 4 RTF on Pentium4-class machines.

The baseline results on the ESST test set 1825 of 1825 sentences divided into a development test set ds2 (1150 utterances) and an evaluation set xv2 (675 utterances) are shown in Table 9.2.

The ESST test set is suitable to test speaker-specific properties of articulatory features, because it contains 16 speakers in 58 different recordings. One recording consists of one side of a dialog by one speaker. There are at least two recordings for every speaker. The system performance was optimized on the ESST development set ds2.

As the stream weight estimation process introduces a scaling factor for the acoustic model, we verified as in the GlobalPhone data that the baseline system can not be improved by widening the beam or by readjusting the weight of the language model vs. the acoustic model. In the experiments presented here, the total weight of the acoustic model is slightly increased, as the “rough” feature models produce on average a higher score as the “main” models, so the beam is even effectively narrowed a little bit.

To improve turnaround times, the settings for MMIE AF weight estimation on the ESST task have been optimized, so that one iteration of AF statistics accumulation and a following update result in a significant improvement, although a second step would then decrease the word error rate, as the step size used is too large to guarantee convergence of the discrimi-

	ESST Test set		
AFs adapted on	1825	ds2	xv2
No AF training	25.0%	24.1%	26.1%
1825	23.7%	22.8%	24.9%
ds2	23.6%	22.6%	24.9%

Table 9.3: WER on the ESST task using global stream weights when adapting on test sets 1825 and ds2.

native update. Results after one iteration of weight estimation on the 1825 and ds2 data sets (which is a sub-set of 1825) using step size $\epsilon = 4 \cdot 10^{-8}$, initial stream weight $\lambda^0_{i \neq 0} = 3 \cdot 10^{-3}$, and lattice density $d = 10$ are shown in Table 9.3:

While adaptation generally works slightly better when adapting and testing on the same corpus (1825 and ds2), there is only a 0.1% loss in accuracy on xv2 when adapting the weights ds2 instead of 1825, which has no speaker overlap with xv2, so generalization on unseen test data is good.

As ESST provides between 2 and 8 dialogs per speaker, it is now possible to adapt the system to individual speakers in a round-robin experiment, i.e. it is possible to decode every test dialog with weights adapted on all remaining dialogs from that speaker in the 1825 test set. Using speaker-specific global weights computed with the above settings, the resulting WER is 21.5%. The improvements from using speaker-dependent global AF stream weights are therefore from 25.0% to 21.5%.

The training parameters for the results shown in Table 9.3 were chosen to display improvements after the first iteration without convergence. Consequently, training a second iteration of global weights does not improve the performance of the speaker adapted system. It is however possible to compute state dependent (SD) feature weights on top of the global (G) weights using the experimentally determined smaller learning rate of $\epsilon_{SD} = 0.2 \cdot \epsilon_G$. In this case, context dependent AF stream weights can further reduce the word error rate to 19.8%. These are the lowest numbers reported on the ESST test set reported so far.

To show the correspondence between improvements in optimization criterion F_{MMIE} and word error rate, an experiment was run with lower settings of the learning rate ϵ. Figure 9.1 shows that the optimization criterion F_{MMIE} indeed improves with training and that improvements in F generally correspond with an improved Word Accuracy, although there is no direct correspondence as discussed in Section 7.3.

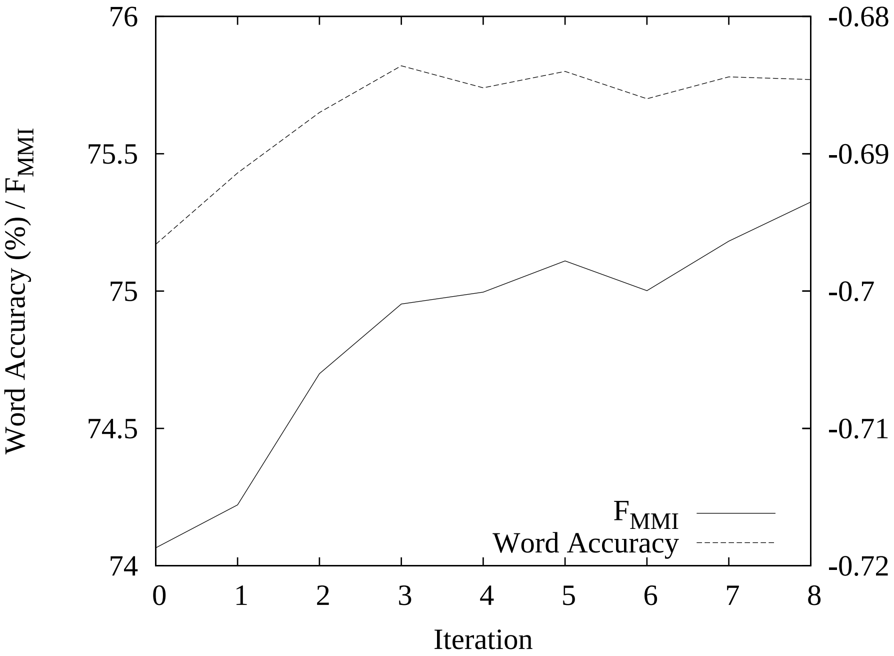


Figure 9.1: Correspondence between Maximum Mutual Information optimization criterion F_{MMIE} and Word Accuracy (WA) in %. Settings: step size $\epsilon = 2 \cdot 10^{-8}$, initial stream weight $\lambda^0_{i \neq 0} = 1 \cdot 10^{-4}$, lattice density $d = 10$.

9.2 Analysis of ESST MMIE Results

9.2.1 Constant Feature Probabilities

Interpolating the standard models and the feature models in a stream architecture amounts to smoothing the standard models, depending on the weight of the main stream. If we replace the feature detectors with an "average" feature detector, which always outputs an average value for each feature, determined on the test data, we can reach a word accuracy of 24.6% on ds2, which is still an improvement over the baseline (25.0%), but clearly behind the trained feature weights (23.3%). The improvement here comes from a slight re-adjustment of the relative weights between language model and acoustic model in the first decoding pass.

9.2.2 Phone Recognizer as Second Stream

Another approach would be to combine the information with a context independent (CI) recognizer. This recognizer would normally be used during construction of the context decision tree. The CI acoustic models are trained in exactly the same way as the standard models, however there is no context decision tree and the number of Gaussians is $143 * 60 = 8580$, which is approximately the same number of parameters as a 16-stream feature model. The baseline performance of this system is 38.2% WER on `1825` (37.9% on `ds2` and 38.5% on `xv2`).

Building a two-stream system "CD+CI" of CD and CI models, similar to [SZH+03], although we are using state likelihoods instead of phone posteriors here, allows training the weights of the two streams on `ds2` using the MMIE criterion as for the feature streams. Doing this results in a best performance of 23.3% on the `ds2` data set after a maximum of 4 iterations of training, which compares to 25.0% for the baseline system and 24.6% for the system with constant "average" feature detectors. The trained context-independent articulatory feature detectors reach 22.8% WER.

On the `xv2` evaluation set, the respective numbers are 26.1% for the baseline, 26.7% for the system with constant feature weights, 25.5% for the CD+CI system, and 24.9% for the SD-AF system. The training of this system is shown in Figure 9.2. For the CD+CI (and the SD-AF systems), the final weights and the performance attained after training are practically independent of the starting weights $\lambda_{0,i}^{(0)}$, which shows the numerical stability of the algorithm.

9.2.3 Adaptation Experiments

When we trained speaker-dependent Articulatory Feature weights in Section 9.1, we were effectively performing supervised speaker adaptation using Articulatory Features. It is therefore interesting to compare the performance of AFs to other approaches to speaker adaptation. We therefore adapted the ESST acoustic models to the test data using supervised constrained MLLR [Gal97], which exhibits a comparable number of free parameter as an adaptation approach.

The results in Table 9.4 show that AF adaptation performs quite well when compared to supervised C-MLLR adaptation, particularly for the speaker-specific case. Supervised C-MMLR reaches a WER of 22.8% when decoding every ESST dialog with acoustic models adapted to the other (between 1 and 7) dialogs available for this speaker. AF-based adaptation reaches a number of 21.5% for the global (G) case and a number of 19.8% for the state dependent (SD) case. The number of free parameters is 40*40=1.6k for the C-MLLR case and 69 for the G-AF case. The SD-AF case has 69*4000=276k free parameters (equivalent to an extra 4k Gaussians), but decision-tree based tying using a minimum count reduces these to 4.3k per speaker. Full MLLR (adapting the means only) on a per-speaker basis uses 4.7k parameters in the transformation matrix on average per speaker, but performs worse than AF-based adaptation by about 1% absolute.

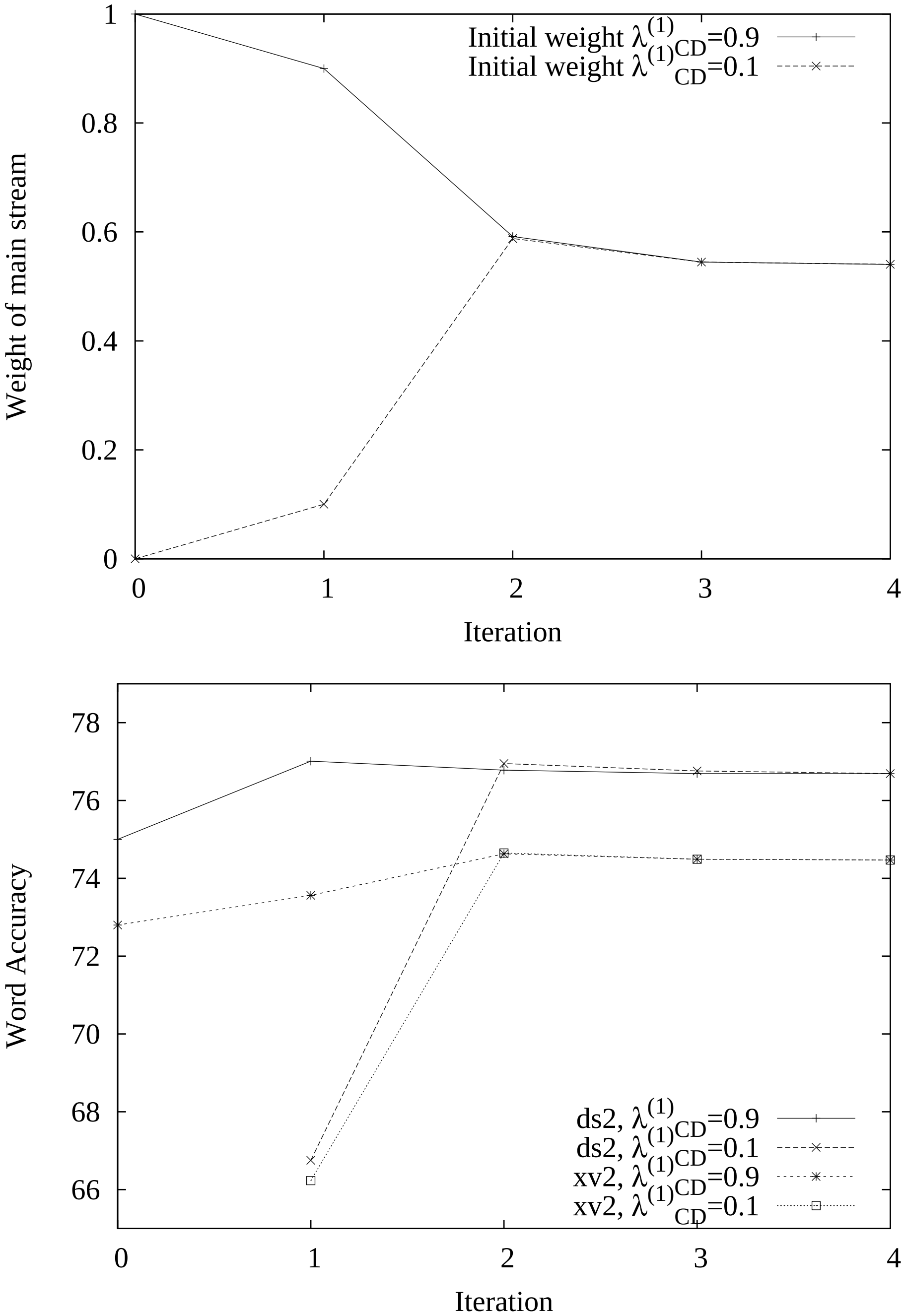


Figure 9.2: Four iterations of MMI training of feature weights for a two-stream "CD+CI" system for initial values of $\lambda^0_{CD} = 0.1$ and $\lambda^0_{CD} = 0.9$. The learned weights (top) and the word accuracy (on ds2 and xv2, bottom) do not depend on initial values $\lambda^{(0)}$.

Adaptation Type	1825	ds2	xv2
None	25.0%	24.1%	26.1%
C-MLLR on ds2		22.5%	25.4%
C-MLLR on speaker	22.8%	21.6%	24.3%
MLLR on speaker	20.9%	19.8%	22.4%
AF on ds2 (G)		22.8%	24.9%
AF on ds2 (SD)		22.5%	26.5%
AF per speaker (G)	21.5%	20.1%	23.6%
AF per speaker (SD)	19.8%	18.6%	21.7%

Table 9.4: Word error rates on the ESST task using different kinds of adaptation: The first three adaptations use C-MLLR, "on speaker" refers to adaptation on all dialogs of the speaker, except the one currently decoded ("round-robin", "leave-one-out" method). Speaker-based AF adaptation outperforms speaker adaptation based on C-MLLR.

9.2.4 Weights Learned

The combination of the "main" stream with the "feature" streams uses different weights for different features, depending on how "important" these streams are for discrimination. Features that help to avoid specific mistakes (phonetic confusions) the main stream makes, will have a high weight, while streams that do not contribute discriminative information will be reduced to have a low weight by the iteration procedure. The resulting stream weights therefore represent a measure of how "important" a specific stream is.

The global feature weights learned by MMI training on ESST data are shown in Table C.12 in Appendix C.4. The most important questions are for the Vowel/ Consonant distinction and then for vowel qualities (Low-Vow, Cardvowel, Back-Vow, Round-Vow, Lax-Vow). These are followed by questions on point (Bilabial, Palatal) and manner (Stop) of articulation. The least important questions are for voicing and consonant groups, which span several points of articulation (Apical, Vls-Pl, Vls-Fr), particularly Sibilants and similar features (Strident, Alveolar). Similar (Consonant, Consonantal and Round, Round-Vow) features receive similar weights while complementary (Vowel, Consonant and Voiced, Unvoiced) features receive (nearly) identical weights.

9.3 Comparison of Speaking Styles

To analyze the influence of the speaking style on the selected features for a specific speaker, the features trained on different kinds of speech (e.g. read and spontaneous) can be compared. Data is available for "Rob Malkin", who is speaker "RGM" in the ESST test set and also a speaker in the "ReadBN" database (see Appendix B.3). This database consists of sentences from the Broadcast News corpus re-read in a quiet anchor-speech like setting. We therefore have this speaker's speech in "read" (ReadBN) and "spontaneous" (ESST) style, though we did not perform formal analysis of the speaking styles present in the different recordings.

Comparing the stream weights from ReadBN and ESST in Table 9.5 we find that for spontaneous speech the feature streams place more weight on the identification of vowel classes such as Cardvowel, Low-Vow, High-Vow, Front-Vow, and Lax-Vow as well as generic classes such as Fricative, Plosive, and Continuant, while read speech requires feature streams to help with the recognition of Diphthongs, lip rounding (Rounded) and sounds introduced into the pronunciation lexikon to model Reduced realizations. Both speaking styles do not need feature streams for classes such as Voiced, Obstruent, or Strident.

While this study on the only "found" data available for this experiment is statistically insignificant, the results are consistent with the findings in [Esk93], which concludes that generally the articulatory targets of vowels are not normally reached in sloppy speech, so that a "feature" recognizer, that tries to detect more general vowel classes in spontaneous speech, seems a sensible strategy found by the weight training algorithm.

Feature		Rank	
	ReadBN	ESST	difference
NASAL	55	1	-54
ALVEOPALATAL	63	11	-52
AFFRICATE	58	8	-50
DEL-REL	57	7	-50
ALV-FR	53	5	-48
LATERAL	65	17	-48
REDUCED-CON	69	23	-46
REDUCED	67	24	-43
ROUND-DIP	56	13	-43
SIBILANT	73	32	-41
APICAL	52	18	-34
MH-DIP	39	6	-33
ALVEOLAR	70	39	-31
W-DIP	46	16	-30
BF-DIP	59	30	-29
LH-DIP	51	25	-26
VEL-PL	40	14	-26
VCD-FR	62	38	-24
RETROFLEX	45	22	-23
DIPHTHONG	49	27	-22
ALVEOLAR-RIDGE	66	49	-17
BACK-CONS	17	4	-13
VELAR	16	3	-13
CENTRAL-VOW	14	2	-12
CORONAL	68	58	-10
VLS-FR	54	44	-10
Y-GLIDE	41	31	-10
REDUCED-VOW	20	12	-8
SONORANT	61	53	-8
LIQUID	35	28	-7
OBSTRUENT	71	65	-6
UNVOICED	72	67	-5
ROUND	22	20	-2
STRIDENT	64	62	-2
ANTERIOR	60	59	-1
VOICED	74	73	-1
LAB-FR	37	37	0
LABIODENTAL	36	36	0
W-GLIDE	42	42	0
Y-DIP	18	19	1
LAB-PL	6	10	4
MID-VOW	29	33	4
PALATAL	5	9	4
LQGL-BACK	43	48	5
BACK-VOW	19	26	7
LABIALIZED	38	45	7
LW	33	41	8
DNT-FR	48	57	9
INTERDENTAL	47	56	9
ASPIRATED	25	35	10
GLOTTAL	24	34	10
TENSE-VOW	50	61	11
HIGH-CONS	3	15	12
ALV-PL	27	40	13
CONSONANTAL	44	60	16
BILABIAL	26	43	17
ROUND-VOW	10	29	19
VCD-PL	2	21	19
VLS-PL	32	51	19
APPROXIMANT	31	55	24
LIQUID-GLIDE	30	54	24
LABIAL	21	46	25
PLOSIVE	23	52	29
FRONT-VOW	34	66	32
STOP	28	63	35
HIGH-VOW	9	47	38
LAX-VOW	7	50	43
FRICATIVE	15	68	53
SYLLABIC	13	70	57
CONSONANT	12	72	60
LOW-VOW	4	64	60
VOWEL	11	71	60
CONTINUANT	8	74	66
CARDVOWEL	1	69	68

Table 9.5: Rank for different features (1=”highest weight”, 75=”lowest weight” in read (ReadBN) and spontaneous (ESST e029) speech for speaker Rob Malkin.

Chapter 10

Robustness against Conversational Speech

By virtue of the high error rates shown in Figure 3.1, "Meeting" speech is the most difficult task in current large vocabulary speech recognition. This figure shows that, even though a-priori the recording conditions are usually better for 16kHz RT-04S "Meeting" data than for 8kHz telephony data (CTS), error rates are significantly higher even for the close-talking condition. While some part of the loss can be attributed to the overall limited amount of training data available today for the "Meeting" task, a major difficulty in acoustic modeling is the wide range of speaking styles found in meeting data. This is a result of the meeting participant's physical proximity, which allows them to interact more freely than for example during telephone conversations.

This wide range of speaking styles observed can be dealt with either by adaption, specialization, or by building a recognizer robust against variations in speaking style. While there are a number of acoustic cues to speaking style that can be computed, training data for speaking-style-specific or -adapted systems is not abundant and laborious to label. As the AF-based recognizer presented in the previous chapters has been shown to improve recognition on spontaneous speech, this chapter will evaluate the robustness of the AF-based approach against the different speaking styles found in "Meeting"-type speech recorded through close-talking microphones. In this application, the speaker is usually known, because the data has been recorded as part of a series of recordings, so that speaker-adapted systems and feature stream weights trained for a specific speaker can be used.

Meeting speech is characterized as being "Highly-Interactive/ Simultaneous Speech":[1]

> The speech found in certain forms of meetings is spontaneous

[1] From http://www.nist.gov/speech/test_beds/mr_proj/.

> and highly interactive across multiple participants. Further, meeting speech contains frequent interruptions and overlapping speech. These attributes pose great challenges to speech recognition technologies which are currently typically single-speaker/ single speech stream contextual.

10.1 The NIST RT-04S "Meeting" Task

Because of its high spontaneity, "Meeting"-type speech is therefore suitable to verify the potential of AFs for improving automatic transcription of conversational speech. The ASR system used in these experiments is trained for 16kHz/ 16bit close-talking audio data from group meeting recordings. The acoustic models were developed for and used in ISL's submission to the IPM ("Individual Personal Microphone") condition of the STT ("Speech-To-Text") part of the NIST RT-04S "Meeting" evaluation system [MJF+04, GLF04].

Training data for 16kHz acoustic models in the ISL system consisted of the close-talking parts of the "Meeting" training data merged with 180h of existing Broadcast News data from the 1996 and 1997 training sets. "Meeting" training data [NIS04a] was collected for the NIST RT-04S "Meeting" evaluation [NIS04b] and consists of "naturally occurring multi-party interaction" [JAB+04] collected in meeting rooms at ICSI, CMU, and NIST. As is was collected at different sites over a longer period of time with different recording procedures, it is not a homogeneous data set. Initial experiments on a pre-release of the official development set with un-adapted single-pass systems confirmed that merging meeting and BN data for acoustic model training is beneficial.

A comprehensive description of each data set with recording conditions and transcription conventions can be found in the literature [BS04, JAB+04, SG04b, SG04a]. For our experiments, BN data was automatically clustered for VTLN estimation and speaker-adaptive training. The parameters of the training data are tabulated in Appendix B.4, durations reported are the actual amount of data processed by the system. No training data was available for "LDC" meetings.

10.1.1 Dictionary and Language Model

Language models were trained in analogy to the ISL Switchboard system [SYM+04]. We trained a simple 3-gram LM and a 5-gram LM with ∼800 automatically introduced classes on a mixture of the Switchboard and Meeting transcriptions and also a 4-gram BN LM. All LMs were computed over a vocabulary of ∼47k words with an OOV rate of 0.6% on the development set. For the first decoding passes only the 3-gram LM was used, later decoding and CNC passes uses a 3-fold context dependent interpolation of all three LMs. The perplexity on the development set of the 3-fold interpolated LM was 112.

All tests use a dictionary extended with vocabulary from the meeting domain and the simple language model described above for decoding unless stated otherwise. Consensus lattice processing (CLP) [MBS00] and confusion network combination (CNC) was also performed in later stages using the interpolated language model (see Appendix B.4).

10.1.2 Development and Test Data

Three evaluation conditions using different amounts of information were defined for RT-04S meeting data:

MDM Multiple Distant Microphones (primary)

SDM Single Distant Microphone (optional)

IPM Individual Personal Microphone (required contrast)

The experiments with articulatory features are run on the "IPM" (i.e. close-talking) data. While the official evaluation system used automatic segmentation, the experiments described here are using manual segmentation to prevent possible interactions of AF adaptation and segmentation and to speed up experimentation. The manual segmentation was derived from the reference transcriptions, which were given as SDM files [NIS04b].

Development ("dev") data for the RT-04S evaluation consisted of 10-minute excerpts of eight meetings, two per site (CMU, ICSI, LDC, NIST), with mostly identical speakers, although some meetings were recorded on different days. Eight 11-minute excerpts of different meetings (two per site again) were used for the evaluation ("eval") data. Each meeting has between three and ten participants, recorded on individual channels. The durations reported in Appendix B.4 give the total amount of data processed by the system. There is a significant amount of overlapping speech, as the total audio duration (89m for dev and 100m for eval) is larger than the "wall-clock-time" of the meeting excerpts (approximately 80m and 90m). The data used in these experiments is documented further in [NIS04b].

10.1.3 RT-04S "IPM" Evaluation System

The ISL's entry to the "IPM" condition of the NIST RT-04S evaluation uses following acoustic models were used in the evaluation system:

PLAIN Merge-and-split training on all data followed by 2 iterations of ML Viterbi training on the "Meeting" close-talking data, global STC, no VTLN

SAT ≡ PLAIN, but trained with VTLN and 2 iterations of Viterbi feature-space speaker-adaptive training (FSA-SAT) [JMSK98] on top of ML training

Tree6.8ms "Tree6" Switchboard acoustic models [SYM+04], decoded with 8ms frame shift

Tree150.8ms "Tree150" Switchboard acoustic models [SYM+04], decoded with 8ms frame shift

SAT.8ms "SAT" models decoded with 8ms frame shift

The acoustic models in every pass were always adapted using constrained MLLR in feature space (C-MLLR) [Gal97] and model-space MLLR to the hypotheses from the previous pass, only the first pass is unadapted. The "Tree6" and "Tree150" models were taken from the ISL Switchboard system [SYM+04]:

Tree6 ML-trained, global STC, VTLN, FSA-SAT, single-pronunciation dictionary and context clustering across phones (6 trees)

Tree150 MMIE-trained, global STC, VTLN, MLLR, FSA-SAT, standard phonetic decision tree

For the SWB-trained models, meeting adaptation and test data was down-sampled to 8kHz and passed through a telephony filter. "SAT.8ms" acoustic models are the same acoustic models as in "SAT", only adapted differently and run at a frame-rate of 8ms instead of 10ms. The largest part of the gain between the two passes with "SAT" acoustic models is due to the adaptation on the Switchboard acoustic models, which make significantly different errors than the Meeting models, which results in a "cross-adaptation" effect. The word error rates reached by the different passes of the evaluation system for Manual segmentation (as used here) and automatic segmentation (as used during the evaluation [LJS04]) are shown in Table 10.1.

Comparing results achieved with both segmentations, it is clear that segmentation is one of the IPM condition's main challenges. The problem lies mainly in the number of insertion errors, which increases from 9.8% for manual segmentation to 14.7% with automatic segmentation. This is due to the large amount of overlapping speech and the physical proximity of the speakers, as a combination of these two factors results in a high amount of cross-talk and background speech from other speakers to be present in each speaker's dedicated channel. For manual segmentation, overlapping speech is still present, but to a lesser degree than for automatic segmentation, which does not achieve a clean separation of foreground and background speech.

Table 10.2 shows a breakdown of word error rates to the individual sites.

Models	Segmentation	
	Manual	IPM-SEG
PLAIN	39.6%	43.6%
SAT	33.8%	38.8%
Tree6.8ms	30.8%	35.0%
Tree150.8ms	29.9%	34.2%
SAT.8ms	30.2%	35.3%
CNC	28.0%	32.7%

Table 10.1: Results on the RT-04S development set, IPM condition for manual and automatic segmentation. Confusion Network Combination (CNC) is between the last three passes. There is a loss of ≈4% absolute when using automatic segmentation instead of manual segmentation.

	Manual	IPM-SEG
Overall	28.0%	32.7 %
CMU	39.6%	43.0 %
ICSI	16.2%	20.4 %
LDC	28.9%	33.3 %
NIST	28.2%	35.0 %

Table 10.2: Results on the RT-04S development set, IPM condition, per data site.

10.2 AF Detector Training

Feature detectors for the Meeting data were trained using the methods described in Section 5.1 using the same setup and preprocessing as for the standard 16kHz "SAT" acoustic models. The fully continuous GMMs with diagonal covariance matrices were initialized with maximum likelihood merge & split training up to a maximum of 256 components. Following the merge & split training, one iteration of label training was performed on the meeting training data to compute the distribution weights. Due to the large amount of training data, all feature models reached 256 components.

For the decoding experiments, the AF detectors are evaluated on the same 42-dimensional feature space as the normal acoustic models, which has been adapted to the current speaker using un-supervised FSA (constrained MLLR [Gal97]).

10.3 AF Experiments on Meeting Data

For the AF experiments, we worked with "SAT.8ms" 16kHz models, as they run significantly faster than the SWB models. We opted to work with manual segmentation in order to avoid problems with wrong segmentation and in order to improve turnaround times.

In order to further reduce turnaround times, training experiments were performed with a faster system that used tighter beams (1.2 instead of 1.5) and no optimization of language model weight. This system reaches a WER of 31.2% on the RT-04S development data instead of 30.2% for the "full" system.

Using context-independent speaker-dependent stream weights with optimized settings for the learning rate, a word error rate of 30.2% can be reached instead of 31.2% WER after a single iteration of MMIE training. Using context-dependent and speaker-dependent stream weights the error rate goes down to 28.7%. Using these stream weights in the fully optimized system (i.e. with wide beams), the error rate reaches 28.2%, which is an 7% relative improvement over and nearly equals the 3-way CNC step with the SWB models.

For these experiments, as in the ESST experiments reported in Section 9.2.3, we used transcribed speaker data to adapt the acoustic models to a known speaker. For the 19 of 43 speakers in the development data and 19 of 39 speakers in the evaluation data, which were only seen once, adaptation was performed on the merged data of all other speakers. These speakers, however, do not contribute much speech to the corpus. Doing this supervised speaker-specific adaptation step using MLLR we can reach a performance of 29.3%, which is clearly inferior to the AF-based adaptation as in the case of the ESST experiments reported in Section 9.2.3.

AF Model	Test on	Baseline	Adapted
CMU	CMU	43.1%	42.1%
ICSI	CMU	43.1%	42.1%
NIST	CMU	43.1%	42.3%
CMU	ICSI	18.4%	17.4%
ICSI	ICSI	18.4%	17.2%
NIST	ICSI	18.4%	17.4%
CMU	NIST	31.3%	29.0%
ICSI	NIST	31.3%	28.9%
NIST	NIST	31.3%	29.2%

Table 10.3: Results (word error rate) on the RT-04S development set; IPM condition; CMU, ICSI, and NIST parts; using AF models trained on CMU, ICSI, and NIST data and weights adapted to this data.

To evaluate the robustness of the feature approach and to quantify the influence of different model training on the performance of an AF stream system, we trained AF feature detectors on the CMU (ISL), ICSI, and NIST meeting training data only to see if the performance depends on the amount and source of training data.

The results in Table 10.3 show that performance depends very little on the type of models and adaptation (feature weight training) used. ICSI models (trained on 75h of data) are slightly better than CMU/ NIST models (trained on 11h/ 13h). NIST-trained models even perform worst on NIST data. Articulatory Features therefore can be ported robustly from one recording site and recording condition to another one. The generally better performance of ICSI detectors is due to better model training given the amount of training data and parameters, as the merge & split training did only assign around 95% of the possible Gaussians for CMU and NIST training.

The adapted 16kHz RT-04S "Meeting" evaluation system on the development data can be improved from 30.2% WER to 28.2% WER using "Meeting"-trained models alone, which is nearly as good as the confusion network combination of the "Meeting" and "SWB" system. On the evaluation data, the improvement is from 31.9% to 29.7%, which is also close to the respective performance of the combined system. AF-based speaker adaptation therefore improves ASR also for adapted systems. Table 10.4 shows a summary of results.

MLLR is using a variable number of transforms per speaker. The number of transforms used is determined by the amount of available adaptation data using a minimum count of 1500 frames optimized on development data. The average number of transforms per speaker is 7.1. AF-based adaptation uses a minimum frame count of 150 for the tree-based tying approach, also

System	Description	RT-04S Dev	RT-04S Eval
AF Baseline	Narrow beams, no LM opt	31.2%	33.5%
CI-AF	1i	30.2%	32.7%
CD-AF	2i	28.7%	31.8%
Meeting Baseline	Wide beams and LM opt	30.2%	31.9%
Full CD-AF	+CD-AF	28.2%	29.7%
Meeting+SWB	CNC-Pass with Eval settings	28.0%	29.0%
Meeting	Superv. Dialog-MLLR	26.9%	28.8%
Meeting	Superv. Speaker-MLLR	29.3%	30.5%

Table 10.4: Results on the RT-04S development and evaluation sets, IPM condition; gains through AF adaptation are 7% relative on development and evaluation data. "Superv. Dialog-MLLR" is a cheating experiment to show how much adaptation is possible using supervised adaptation on the test dialog using MLLR.

optimized on the training set.

10.4 Analysis

The stream weights learned on the "Meeting task" are largely comparable to the ones learned on the spontaneous ESST task, presented in Table C.12, i.e. the system places weight on features describing vowel qualities (place of articulation) and manner of articulation (e.g. FRICATIVE, PLOSIVE) while again the VOICED feature is not used to a large extent. The important features are usually the ones which have only few phones in their class. Many of the most prominent features have two homorganic phones in their class, which can be distinguished by voicing (e.g. DEL-REL=AFFRICATE (CH JH), ALV-FR (SH ZH), LAB-PL (P B), LAB-FR=LABIODENTAL (F V); have very similar places of articulation ALVEOPALATAL (SH ZH CH JH), VLS-FR (F TH SH), or are related to diphthongs or vowel characteristics: LH-DIP (AY AW), BF-DIP (AY OY AW OW), CENTRAL-VOW (AH AX IX), X-LMN (XL XM XN), or REDUCED (IX AX AXR).

For further analysis, we computed a phonetic confusion matrix for the Meeting data before and after adaptation with articulatory feature detectors and check the most frequent confusions. In this case, we compute a Viterbi alignment of the reference (allowing optional words and pronunciation variants) and compare it to the recognizer hypothesis.

Table 10.5 shows that the most frequent confusions are the ones between Z and S, which is an inconsistency in the ISL dictionary (this is based on LDC PronLex but has been extended using a rule-based approach), followed by confusions between vowels and vowels and/ or consonants. The "Change" column shows that the largest reductions occur in the confusion

Rank	**No AFs**		**Trained AFs**		**Change**	
	Confusion	Count	Confusion	Count	Confusion	Change
1	Z S	3872	Z S	3817	OW AX	280
2	T IH	1504	T IH	1430	EH AE	229
3	IH AX	1233	T N	1158	N AX	208
4	T AX	1188	IH AX	1116	N M	182
5	EH AE	1143	T AX	1010	R AXR	179
6	T N	1132	T S	985	T AX	178
7	T S	1122	EH AE	914	N D	166
8	T D	936	IH AE	887	AX AE	162
9	N M	930	T D	849	Z T	160
10	R AXR	919	N M	748	IY AX	158
11	IH AE	873	R AXR	740	IY IH	155
12	N D	865	N D	699	UW AX	153
13	N AX	844	N AE	689	N EH	151
14	IY IH	797	IY IH	642	T K	148
15	T K	775	N AX	636	IY EY	141
16	AX AE	746	T K	627	T S	137
17	IY EY	721	Y IY	593	L AX	136
18	Z T	710	AX AE	584	AX AH	129
19	OW AX	704	T DH	582	OW EH	118
20	N AE	678	IY EY	580	UW EH	118
21	AY AE	659	AY AE	569	Y N	118
22	Y IY	653	Z T	550	IH AX	117
23	DH AX	561	IH EY	510	DH D	109
24	N IH	560	UW IY	507	EH AX	102
25	T DH	556	N IH	502	T IY	98
26	OW L	541	N DH	484	T AXR	91
27	IH EY	524	S IH	478	AY AE	90
28	AX AH	521	DH AX	475	T D	87
29	N DH	506	OW L	457	DH AX	86
30	IH EH	501	NG N	455	N AXR	85
31	OW N	501	OW N	439	AY AXR	84
32	EH AX	489	OW AX	424	OW L	84
33	UW IY	487	IH EH	422	R ER	82
34	NG N	475	IH DH	411	AE AA	80
35	S IH	461	N IY	407	IH EH	79
36	DH D	454	AXR AX	405	DH B	77
37	AXR AX	451	UW OW	401	T CH	75
38	IH DH	451	ER AXR	395	T IH	74
39	UW AX	446	AX AH	392	L AY	72
40	EH AY	442	EH AX	387	V AX	72
41	N EH	421	EH AY	374	UW IH	71
42	N IY	415	EY AY	372	K AX	70
43	T IY	411	DH D	345	UW N	70
44	Z IH	396	OW AA	344	AY AH	68
45	OW AA	394	T AY	330	EH AH	68
46	L AX	393	Z IH	330	EH AY	68
47	UW OW	390	N AY	324	T EH	67
48	IY AX	382	T IY	313	Z IH	66
49	DH B	373	Z DH	309	T AE	65
50	N AY	365	T P	306	R AX	64
51	T AY	359	DH B	296	Y IH	64
52	R ER	351	UW AX	293	OW N	62
53	AY AX	350	AY AX	291	T R	60
54	OW EH	350	N AA	284	Y AE	60
55	ER AXR	347	Z AX	283	Y IY	60
56	Z DH	334	W L	272	AX AW	59
57	AE AA	333	N EH	270	AY AX	59
58	Z AX	322	R ER	269	IX AX	59
59	EY AY	320	W OW	269	T M	59
60	T AE	311	AO AA	267	Y AY	59
61	T P	311	IH AXR	267	N IH	58
62	Y IH	303	AH AA	264	OW AY	58
63	D AX	293	N L	261	R OW	58
64	UW N	292	L AX	257	XL UW	57
65	N AA	289	AE AA	253	OW M	56
66	IH AXR	284	D AX	251	T B	56
67	R AX	283	T AE	246	UH AX	56
68	W L	278	S F	243	V T	55
69	UW IH	277	Y IH	239	Y DH	55
70	Y N	277	UW D	237	Z S	55

Table 10.5: Influence of AFs on confusions: the left column shows the most frequent phonetic confusions of a decoding on the RT-04S development set without AFs (Z ← S and S ← Z have been merged for clarity), the middle column shows the ones with trained weights. The reduction in classified frames is given in the right column.

of vowel qualities (Reduced, Reduced-Vow, and Central-Vow), which is consistent with the high stream weights observed for vowel qualities and the observation that vowel qualities are affected in sloppy, conversational speech.

Chapter 11

Robustness against Hyper-articulated Speech

The previous chapters presented results of articulatory feature based speech recognition on conversational speech tasks. As different speaking styles are not labeled in these data sets and there is no control over speaking styles in naturally occurring human-to-human interaction, further insights into the influence of articulatory features on the word error rate of a speech recognition system can not be gained without manually annotating data.

While conversational speech mostly contains speech with reduced articulatory effort, i.e. "sloppy speech", the most important linguistic messages, i.e. utterances which contain particularly important information the speaker wants to stress, will be spoken clearly, i.e. in a clear or at least partly hyper-articulated mode. As speech recognition systems are not trained on this type of speech, their performance usually suffers on this type of speech, although a naive user would expect the opposite. A truly robust speech recognition system must therefore cope with clear speech or at least show as little degradation on clear speech as possible.

11.1 Hyper-articulated Data

Speech recognition performance on hyper-articulated or clear speech can be evaluated on a database of elicited speech collected at ISL, in which a simulated dialog system prompts speakers to produce the same word both "normally" and "clearly", i.e. in two distinct speaking modes. As outlined in Section 5.3, articulatory features now provide contrastive attributes to perceived confusions, and speakers stress these attributes to better transport their message. In this section, we evaluate the robustness of a speech recognition system to this behavior and investigate if articulatory features can improve the performance of an ASR system on this type of data.

To define two distinct speaking styles, we assume in our experiments that humans always try to get away with using minimal effort when choosing their speaking mode. It is clear that hyper-articulation, i.e. attempting to produce very un-ambiguous speech sounds, requires much more effort from the speaker. In human-human communication, hyper-articulation occurs to improve the intelligibility of spontaneous speech. It is shown in [SW98, Ovi98] that hyper-articulated speech also occurs in human-computer interaction, if users try to compensate for real or suspected recognition errors. Assuming that the manifestation of hyper-articulation is not fundamentally different between human-to-human and human-to-computer interaction types, an understanding of improvements reached on the human-to-computer domain can be ported to the human-to-human domain.

For our experiments, we collected data from users, who were told to repeat a word to a simulated automatic dialog system until it finally "understands" (i.e. displays the word the screen) them correctly. The subjects were naive users of speech technology and were not told that the system was a simulation.

The recording scenario consisted of two sessions: during the first session, the subjects used the dialog system under "normal conditions", i.e. they would not attempt to diverge from a canonical pronunciation. After that, a list of recognition errors (word confusions) from the first session was presented to the subjects, which they were told to correct, i.e. produce again in a way the system could transcribe correctly by clicking on a button wrong and repeating the same word. The recognition errors were presented as phrases, e.g. *"The word* recounting *was confused with* recounted*. Please repeat* recounting*".* A maximum of three attempts were performed to correct an error. The subjects were also asked to disambiguate the words in the other direction in order to investigate if opposite features are used to contrast word confusions.

In order to induce realistic hyper-articulated speech, we analyzed typical errors of our speech recognition systems and generated a list of frequent word confusions, which were used to generate the system responses. In most cases, recognition errors were caused by inflections and phonetically similar words. Even though the set-up presented here may look extreme in that the user is subject to artificial errors, the performance of speech recognition systems suffers greatly in many situations, justifying research on error recovery in dialog systems. The experiment presented here is also described in [SMW02] while the experimental design is described in more detail in [Sol05].

In total, the database consists of 4677 normal and 5367 hyper-articulated recordings from 45 subjects (see Table 11.1). The recordings are comparable in domain, vocabulary, microphone, and environmental noise for each speaker across different speaking styles. The corpus was divided in a training set of 34 speakers and a test set of 11 speakers. As the set of training speakers is rather small, we conducted supervised adaptation experiments using acoustic models trained on large corpora, e.g. the SWB and BN databases.

	Speakers	Utterances		Duration	
		Normal	Hyper	Normal	Hyper
Train	34	3506	3923	124min	158min
Test	11	1171	1444	34min	57min
Total	45	4677	5367	158min	215min

Table 11.1: Database for normal ("HSC-normal") and hyper-articulated ("HSC-hyper") speech.

Group	Basis
Manner	Plosive, Nasal, Fricative, Lateral, Approximant
Place	Alveolar, Bilabial, Glottal, Labiodental, Interdental, Retroflex
Vowel	High, Mid, Low, Front, Central, Back, Round
Global	Voiced, Consonantal

Table 11.2: Independent articulatory properties used for the experiments on hyper-articulated speech.

In the following experiments, the described corpus is being referred to by the name "HSC" (Hyper-articulated Speech Corpus).

11.2 Detection of Articulatory Properties

Feature detection experiments on the HSC corpus were run using feature detectors trained on the Switchboard (SWB) corpus of around 300h of conversational telephony speech [GHM92] using the set-up of and labels generated with ISL's RT-03 CTS system [SYM$^+$04] and with systems trained on the newly-collected data. The newly trained systems for normal speech from the HSC database ("HSC-normal") and hyper-articulated speech from the HSC database ("HSC-hyper") use a maximum of 48 Gaussians per feature detector during incremental growing of Gaussians in order to avoid over-training on the limited amount of training data. The preprocessing for these systems is the same as for the ESST system described in Appendix B.2.

We investigated the "independent" properties shown in Table 11.2 for consonants. The likelihood is computed using the corresponding models and anti-models for each frame as described in Section 5.3. The performance is measured as the binary classification accuracy averaged over the number of (middle) frames.

The results for the detection experiments are shown in Table 11.3. The experimental setup permits comparisons of the performance across attributes, speaking style ("normal" or "hyper"), and training corpus ("SWB", "HSC-normal", and "HSC-hyper"):

Corpus:	SWB corpus		HSC-normal		HSC-hyper	
Speaking style:	Normal	Hyper	Normal	Hyper	Normal	Hyper
Manner of Articulation Features						
Plosive	90%	83%	91%	85%	92%	88%
Nasal	88%	82%	93%	87%	93%	90%
Fricative	95%	92%	93%	91%	92%	91%
Lateral	85%	77%	89%	80%	89%	81%
Approximant	90%	85%	88%	82%	87%	85%
Place of Articulation Features						
Labial	83%	80%	88%	83%	86%	83%
Bilabial	84%	78%	87%	83%	88%	85%
Labiodental	90%	84%	80%	72%	78%	72%
Alveolar	88%	86%	87%	84%	88%	85%
Velar	82%	77%	81%	75%	84%	80%
Glottal	84%	79%	83%	81%	81%	86%
Global Features						
Voiced	96%	96%	92%	92%	86%	83%
Consonant	96%	93%	87%	83%	88%	85%
All	85%	81%	86%	81%	85%	83%

Table 11.3: Detection accuracy for different features for consonants.

Differences Between Attributes: The average classification accuracy over all attributes is 86% (Table 11.3, bottom). The detection performance for manner of articulation varied between 88% for approximants and 93% for fricatives and nasals. Classification is worse for place of articulation.

Differences Between Speaking Modes: The classification performance can be analyzed across speaking modes by comparing the fourth with the fifth column. The classification accuracy is 5% worse on hyper-articulated speech over all attributes. The impact of hyper-articulation on the detection accuracy is more or less equal for all attributes.

Differences Between Training Corpora: The detection accuracy for normal speech is independent from the training corpus. The models trained on SWB reach 85% on average, training with HSC-normal gives 86%, and 85% is also obtained by estimating the parameters on HSC-hyper. The channel mismatch for the SWB models (8kHz, telephony speech) does not seem to degrade the detection accuracy. By comparing the fifth column with the seventh column, it can be seen that hyper-articulated training data improves the performance from 81% to 83%. In particular, velar and glottal sounds profit from that data. On the other hand, the classification whether a sound is voiced or not becomes significantly worse.

Therefore, articulatory features can be detected on hyper-articulated speech with practically the same accuracy as on normal speech. Large differences between normal and clear speaking style could only be found for voicing.

11.3 Speech Recognition with AF Models

For the decoding experiments on the HSC database, the acoustic models were used together with a zero-gram language model and a search vocabulary of around 8000 words. The thresholds of the beam search algorithm were sufficiently high to avoid search errors. This experimental setup ensures that any recognition errors can be directly attributed to the acoustic models.

Initial experiments with the Switchboard models indicated significant differences between normal and hyper-articulated speech. While an error rate of 25.6% is obtained for unadapted models under “normal” conditions, there is a relative error increase of more than 60% to 41.6% under conditions of hyper-articulation on average over all test speakers. This relative error increase however strongly depends on the speaker, as it varies between 4% and 260% for the 11 speakers of the test set.

Acoustic models	Error rate		Relative error increase
	Normal	Hyper	at hyper-articulation
Baseline	25.6%	41.6%	62.5%
MLLR	21.9%	35.0%	59.8%
MAP	23.4%	37.9%	61.9%

Table 11.4: Supervised adaptation on hyper-articulated speech.

The observed speaker-dependent deterioration of word error rate suggests that the way users change their speaking style in order to disambiguate recognition errors is speaker dependent. The acoustic models, trained on conversational telephone speech, are not able to deal with hyper-articulated speech well. This experiment shows that:

- There are significantly more recognition errors under hyper-articulation
- The reaction on word confusions is a strongly speaker-dependent effect in terms of an increase in recognition errors

This is particularly remarkable, as the performance of the detectors for articulatory features did not degrade significantly between speaking styles. On other data, [GOK03] finds that WER may be more related with user frustration than with hyper-articulation and argues that hyper-articulation could be compensated for by using more speaking-mode specific training data.

To improve on the results on our data, we tried to adapt the SWB models using MLLR [LW94] and MAP [GL94] adaptation on the HSC corpus. The regression tree contains 256 nodes and the minimum occupancy threshold for the adaptation matrices is set to 1500 samples. The prior distribution for MAP is estimated on the SWB corpus using 2.6h of hyper-articulated adaptation data. The results are given in Table 11.4.

Adaptation data	Error rate		Relative error increase
	Normal	Hyper	at hyper-articulation
Baseline	25.6%	41.6%	62.5%
Normal	21.9%	36.8%	68.0%
Hyper	21.9%	35.0%	59.8%
Normal+Hyper	21.4%	35.3%	64.9%

Table 11.5: Supervised MLLR on different training sets.

While significant reductions in word error rate can be achieved with standard likelihood based approaches, these fail to improve the performance on hyper-articulated data, even when adapting on hyper-articulated data. The relative increase in error rate between normal and hyper-articulated speech stays at around 60%. To investigate if channel effects may be masking speaking style adaptation, we repeated these experiments with the ESST+BN models described in Chapter 9. The result of this experiment on the HSC data is shown in Table 11.6.

Initially, the SWB models provide better WER. After MLLR adaptation on the HSC-normal data however, the ESST+BN models give significantly better results. The adaptation is more effective for the ESST+BN models, resulting in an error rate of 18.9% for normal speech and 29.9% for hyper-articulated speech. The degradation on hyper-articulated speech however is still nearly 60% relative. The adapted ESST+BN models however set a baseline for the multi-stream AF approach, since they provide a "harder" baseline for ASR experiments.

To see if particular phonological properties are affected differently by

Adapted on	SWB models		ESST+BN models	
HSC-normal	Normal	Hyper	Normal	Hyper
No	25.6%	41.6%	32.7%	46.3%
Yes	21.9%	36.8%	18.9%	29.9%

Table 11.6: Comparison of ESST+BN with SWB models and supervised adaptation.

hyper-articulation, we partitioned the set of articulatory features into four sub-spaces shown in Tables 11.2 and 11.7. The total number of Gaussians in the feature models is 1216. The number of additional parameters introduced with the feature models is therefore negligible when compared to the phone models.

A separate system was built for each sub-space in a first step investigating the capabilities of each feature attribute group. The baseline is the standard phone-based model set. The full vector space uses all attributes. Stream weights were set equally for all feature streams with the main stream receiving a weight of $\lambda_0 = 0.5$ and the sum of the feature streams set to 0.5.

Acoustic Models	Speaking Style	
	Normal	Hyper
Phone-based Models	18.9%	29.9%
+ Manner AF models	17.3%	22.2%
+ Place AF models	17.5%	22.3%
+ Vowel AF models	17.4%	22.4%
+ Global AF models	18.2%	23.2%
Full AF Models	17.8%	21.5%

Table 11.7: Recognition experiments with AF stream architecture.

The results in Table 11.7 demonstrate the advantages of using articulatory features for robust recognition of hyper-articulated speech. The error rate is reduced from 29.9% with the phone models to 21.5% using the available detectors for articulatory features. This is an improvement of more than 28% relative. Moreover, this improvement on hyper-articulated speech does not cost performance for normal speech. The phone based models have an error rate of 18.9% for normal speech, but the vector models achieve 17.8%. More important, by re-assigning part of the overall decision from specialized phone models to generic detectors for articulatory features, the performance of the speech recognition system is improved dramatically on hyper-articulated speech, on which articulatory feature detectors are much more reliable than standard phone models, as there is nearly no degradation in feature classification rate. Also, for hyper-articulated speech, a combination of all features is better than the selection of feature sub-groups.

The performance on the articulatory sub-spaces is rather good. The spaces formed by manner or place(s) of articulation give most of the gain. This suggests that only a limited number of contrastive attributes are needed to correct a recognition error. Vowel and consonants appear to be well separated, but apart from that there is no indication that one of the sub spaces is more important than any other for compensating hyper-articulation. The results for all sub spaces are comparable.

Attributes changed as predicted	51.2%
Attributed changed in the wrong direction	14.8%
Attributes did not change	34.0%
At least one correct prediction per phone	78.6%

Table 11.8: Prediction of contrastive attributes.

11.4 Analysis of Contrastive Attributes

Using the contrastive attributes (CAs) introduced in Section 5.3 should allow us to predict changes in articulation when changing from “normal” to “clear” speaking mode. Earlier individual examples of predictable changes (see Figures 5.5, 5.6, 5.7, and Table 5.5) however need to be verified by a comprehensive analysis of the changes occurring in real speech.

To arrive at feature change candidates, we aligned the phone sequences of the confused words: for example, if `bitter` was uttered and `better` was recognized, we aligned the sequences [bɪtər] and [betər] (and their pronunciation variants) with the acoustic evidence, i.e. the “normal” and “hyper” utterances using the Viterbi algorithm, at the same time selecting the best pronunciation variant. The alignment procedure produces a set of insertions, deletions, and substitutions at the phonetic level. The phone errors can then be represented in the articulatory formulation as the *activation* of one (or more) features (Near-Close and Near-Front in the above example) and the *de-activation* of others (Close-Mid and Front), i.e. the hyper-articulated [ɪ] will be “more closed” than the standard sound to disambiguate it from the “more mid” [e]. On average, phone confusions led to an average of 3.5 feature changes per affected segment, while correct phones did not generate any changes.

A CA is now correctly predicted, if the average score difference

$$\Delta_g = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \Delta(o_t, a)$$

per frame is larger (lower) for the activated (deactivated) feature in the hyper-articulated realization than in the normal realization of the same word over the region with different attributes. For the above example, this value was plotted over time in Figures 5.5, 5.6, and 5.7 (cf. Table 5.5).

Table 11.8 shows how often contrastive attributes are correctly predicted by the detectors for articulatory features. A wrong prediction does not necessarily mean that the predictor models were not able to detect the attribute change. Instead, it is also possible that the attribute change did not occur. For example, there are 3.5 predicted changes per phone on average and it might also be possible that humans use only a limited number of attribute changes for disambiguation between the true and mis-recognized

Contrastive Attributes	Speaking style	
	Normal	Hyper
AF models	17.8%	21.5%
Enforced attributes (ref)	17.8%	17.0%
Enforced attributes (hyp)	17.8%	19.4%

Table 11.9: Enforcing contrastive attributes.

word. Keeping this in mind, the results can be interpreted only as a correlation between predicted and observed changes and not as a correctness of the predictor models. Indeed it is our initial assumption that phone models are too coarse a model to accurately described the articulatory changes during spontaneous or hyper-articulated speech.

The results in Table 11.8 show that 51.2% of all predictions occurred, while 14.8% negative attribute changes were observed. In 34.1% of cases, the attributes did not change. Furthermore, at least one attribute change per phone is correctly predicted in 78.6% of all phone occurrences. In other words, the probability for observing a contrastive attribute in a hyper-clear speaking mode is 78.6%.

Given the predictions, a recognition experiment can be performed by enforcing the contrastive attributes. The idea is to increase or decrease the weighting factors of the contrastive attributes in the acoustic score computation. This recognition run is a kind of a "cheating experiment" since the contrastive attributes are obtained by an alignment of the confused words, i.e. with knowledge of the true hypothesis. The result of this experiment is shown in Table 11.9. The error rate improves from 21.5% to 17.0% on the hyper-articulated data. Instead of using transcripts to obtain contrastive attributes, hypotheses from the corresponding normal utterances can be used. As shown in Table 11.9, enforcing attributes based on hypotheses leads to a recognition performance of 19.4% error rate, which represents an improvement of 9.8% relative and is only 0.5% worse than the baseline performance (without articulatory features) on the "normal" data.

The analysis presented in this section gives evidence that changes due to a hyper-clear speaking mode can be explained by the concept of contrastive attributes based on articulatory features. There is a correlation between the observed and predicted attribute changes. Enforcing contrastive attributes improves the recognition performance significantly. There is no need to train feature detectors on hyper-articulated data, as these can be trained on normal data.

Articulatory features therefore play an important role in how humans produce speech in different situations, as they seem to particularly stress contrastive attributes when trying to discriminate between confusion pairs. It sounds plausible that a similar effect can be observed in sloppy speech, where only those features are retained, which are still needed for discrimination in the current context. The discriminatively trained multi-stream approach to ASR therefore improves automatic speech recognition by incorporating information about competing hypotheses in a linguistic space, instead of a likelihood-based adaptation, which is mostly based on incorrect model assumptions.

Chapter 12

Conclusions

This thesis presented an automatic speech recognition system particularly suited for conversational speech, which allows combining standard context dependent acoustic models with detectors of broad, phonologically motivated "articulatory features" (AFs) such as VOICED or ROUNDED. Combining these two types of classifiers in a multi-stream approach with discriminatively trained stream weights for the individual articulatory features allows adapting the recognizer to the articulatory characteristics of an individual speaker, his or her speaking style in a particular situation, e.g. in a meeting, or a particular task better than existing approaches. The streams of the approach presented model different articulatory features, which the discriminatively trained stream weighting mechanism combines into an overall decision. Because the feature streams directly model phonologically motivated broad, distinctive categories, the multi-stream approach can capture articulatory changes and characteristics occurring in individual speakers better than a purely phone based approach. Adaptation is achieved by setting the combination weights appropriately for each phonetic context. Experiments on hyper-articulated data show that the proposed system improves the performance on "clear", i.e. "important", parts of speech by more than 25%. Overall improvements on conversational speech show improvements between 7% and 20% relative.

12.1 Thesis Results

This thesis evaluated a multi-stream approach to knowledge combination using well-trained context-dependent phone models and models based on articulatory features on a number of tasks and derived new formulae for the computation of stream weights suitable for the combination of asymmetric (i.e. differently salient) knowledge sources.

The experiments presented in this thesis show improvements over previous automatic speech recognition systems on several tasks:

- The WER on English GlobalPhone data could be reduced from 15.6% to 13.9%. Similar improvements could also be reached using feature detectors from different languages (cf. Section 8.3, "multi-lingual" data) or on other languages, which shows that articulatory features are indeed a language-independent property of human speech production and perception. On this data set, the new MMIE based stream weight estimation was shown to outperform previous DMC based estimation techniques in terms of computational effort while reaching comparable reductions in error rate.

- On spontaneous speech (ESST), the WER could be reduced to 21.5% using global (G) AF speaker adaptation. State-dependent (SD) AF speaker adaptation reaches 19.8% WER while MLLR speaker adaptation using a comparable number of parameters reaches 20.9%. Baseline performance using ML-trained models without adaptation is 25.0% WER (cf. Section 9.1). Using speaker-independent AF weights trained on the development test set, the WER on the evaluation set can be reduced from 26.1% to 24.9% while MLLR adaptation gives 50% of that gain.

- On the ESST task, the proposed algorithm allows computing a weight using a phone recognizer as a second stream instead of the feature streams. This improves the WER by 0.6% absolute while the AF system improves the performance by 1.2%. This shows that the algorithm works robustly and can be applied to the combination of other types of knowledge sources as well (cf. Section 9.2).

- The RT-04S "Meeting" system can be improved from 30.2% WER to 28.2% WER on the development data using "Meeting" models alone in the last decoding pass, which is nearly as good as the combined "Meeting" and "SWB" system (28.0%). The best non-AF single pass system reaches 29.9%. On the evaluation data, the system is improved from 31.9% to 29.7% (cf. Section 10.1).

- Experiments on "hyper-articulated" speech confirm that modeling speech using articulatory features can improve the performance over a standard phone based recognizer particularly for cases when users change their speaking mode in order to speak "more clearly". When enforcing distinctive attributes for confusable word pairs on hyper-articulated data, the WER is reduced from 29.9% without AFs to 19.4%, which is nearly as good as the performance of the non-AF system on clean data (18.9%) (cf. Chapter 11).

Table 12.1 summarizes the improvements in word error rate reached on the different tasks considered in this thesis.

Experiment	WER	Improvement	
		absolute	relative
Hyper-articulated data (HSC)			
Normal Baseline	18.9%		
Normal AF	17.8%	-1.1%	-6%
Hyper Baseline	29.9%		
Hyper AF	21.5%	-8.4%	-28%
GlobalPhone data (GP)			
EN Baseline Dev	12.7%		
G-MMI-AF	11.3%	-1.4%	-11%
SD-MMI-AF	10.9%	-1.8%	-14%
EN Baseline Eval	15.6%		
G-MMI-AF	14.4%	-1.2%	-7%
SD-MMI-AF	13.9%	-1.7%	-11%
English Spontaneous Scheduling Task data (ESST)			
1825 Baseline	25.0%		
Speaker AFs (G)	21.5%	-3.5%	-14%
Speaker AFs (SD)	19.8%	-5.2%	-21%
"Meeting" data (RT-04S)			
Dev Baseline (fast)	31.2%		
G-AF	30.2%	-1.0%	-3%
SD-AF	28.7%	-2.5%	-8%
Dev Full Baseline	30.2%		
SD-AF	28.2%	-2.0%	-7%
Eval Full Baseline	31.9%		
SD-AF	29.7%	-2.2%	-7%

Table 12.1: WER improvements achieved with articulatory features and a multi-stream HMM architecture on different tasks.

12.2 Thesis Contributions

The experiments presented in this thesis allow the following conclusions, which contribute to improved automatic speech recognition:

1. Articulatory features in combination with existing well-trained acoustic models are helpful for recognizing spontaneous, conversational speech. Even though they are simpler in structure, they perform better than a context independent phone recognizer added as a second stream. This supports the notion that atomic "articulatory" properties such as VOICED or ROUNDED can help to discriminate between words, while phones, which represent a whole bundle of these features, are a convenient short-hand notation of the articulatory process, but do not necessarily play a major role in perception.

2. Articulatory features can increase the robustness of a recognizer as they are particularly suitable for recognizing "emphasized" or "hyper-articulated" speech. While automatic speech recognition systems can be made to recognize "standard" speech rather well, they tend to fail as soon as users are excited, angry, or otherwise under pressure and try to speak "hyper-clear", i.e. with emphasis. People hyper-articulate, as they expect this type of speech to be more easily understandable for a human, however the effect when using a machine is usually just the opposite. AFs have been shown to be effective at reducing that degradation, therefore contributing to the applicability of dialog systems and other "end-user" products. AFs are particularly suitable for building systems adapted to a specific speaker.

3. The MMI based discriminative training approach developed as part of this thesis can successfully combine specific models (standard context-dependent phone based acoustic models) with generic ones (context-independent models based on phonetic classes) or other acoustic models, e.g. a context-independent phone recognizer.

4. Articulatory features can be reliably detected from speech by using well known standard acoustic modeling techniques. Articulatory features can also be recognized across languages, i.e. the phonetic assumption that phonological features are "universal" seems justified.

5. As the phonetic properties the detectors are built on can be detected reliably and these are more portable across languages than phones, articulatory features can help to improve speech recognition in languages with sparse data or can help to bootstrap systems in new languages, therefore enabling speech recognition to be made available to more users more easily.

12.3 Recommendations for Future Work

The present work represents significant improvements in terms of word error rate over well-trained state-of-the-art baseline systems on several tasks, yet still there are many open issues within the field of ASR that the approach presented here could contribute to.

The approach was shown to be working well when adaptation has been performed on data from the same speaker. As AFs have also been used for speaker verification, it is interesting to investigate the suitability of articulatory features for an integrated approach to speaker verification and speech recognition. Also, the dependency of stream weights on other factors such as speaking style, emotion, or dialect should be investigated further in order to improve the selection of weights for unknown conditions. Another area of research would be the portability of speaker specific stream weights across different tasks or channels.

Also, the language-independent properties of articulatory features lend themselves to further research on multi-lingual speech recognition or bootstrapping of systems in new languages. Recognition of non-native speech might also profit from the use of articulatory feature detectors ported to the foreign language recognizer from the speaker's mother tongue.

Articulatory features performed particularly well on extreme, i.e. hyperarticulated or clear, speech, without the need to train extra models. Reliable detection of extreme speech would allow improving the performance of a speech recognition system particularly on these important parts of speech by appropriately including articulatory features.

Next, existing work on noise-robust speech recognition could be replicated with the proposed multi-stream approach and the change of optimum feature stream weights with noise condition could be investigated. This would permit to quickly adapt the recognition system to different noise conditions.

Also, the dependency of word error rate improvements on the accuracy of the underlying feature detectors has not been studied extensively in this work. It seems plausible that some feature detectors could be improved by computing their likelihoods not on MFCCs, but on other features (power, PLP coefficients), which should in turn improve word error rate. Finally, first experiments on speech recognition based on feature streams alone (without pairing them with the baseline models) did not lead to improved performance on the tasks covered in this work, but careful training of state-dependent stream weights should allow for a feature-only speech recognition system.

Appendix A

MMIE Update Equation

Derivation of Equation 7.5 starting from the MMIE criterion 7.4:

$$\begin{aligned}
F_{\mathrm{MMIE}}(\Psi) &= \sum_{r=1}^{R} \log \frac{p_\Psi(O_r|W_r)P(W_r)}{\sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w})} \\
&= \sum_{r=1}^{R} \left(\log p_\Psi(O_r|W_r)P(W_r) - \log \sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w}) \right)
\end{aligned}$$

where W_r is the correct transcription of utterance r and $\hat{w}$ enumerates all possible transcriptions of r with a non-zero likelihood given the lattice produced using the acoustic model PDF p_Ψ and language model probabilities P. Formally deriving F_{MMIE} with respect to λ_i (Ψ comprises the full parameter set $\{\lambda_{i,s}, \mu_l, c_l, \Sigma_l\}$ for all streams i and all Gaussians l, independent of their state s) gives:

$$\frac{\partial F}{\partial \lambda_i} = \sum_{r=1}^{R} \left(\frac{\partial}{\partial \lambda_i} \log p_\Psi(O_r|W_r)P(W_r) - \frac{\partial}{\partial \lambda_i} \log \sum_{\hat{w}} p_\Psi(O_r|\hat{w})P(\hat{w}) \right)$$

Let $\mathcal{S}$ denote all possible states s contained in the possible hypotheses $\hat{w}$. Using the Markov property of any state sequence through $\mathcal{S}$, we can write the partial derivatives with respect to the weights $\lambda_{i,s}$ in the time range 1 to T_r

$$\frac{\partial}{\partial \lambda_{i,s}} \log p(O_r|W_r) = \sum_{t=1}^{T_r} p(s_t = s|O_r, W_r) \frac{\partial}{\partial \lambda_{i,s}} \log p(O_{r,t}|s)$$

Now introducing the *Forward-Backward probabilities*

$$\begin{aligned}
\gamma_{r,t}(s; W_r) &:= p_\lambda(s_t = s|O_r, W_r) \text{ and} \\
\gamma_{r,t}(s) &:= p_\lambda(s_t = s|O_r)
\end{aligned}$$

we can write

$$\frac{\partial F}{\partial \lambda_i} = \sum_{r=1}^{R}\sum_{t=1}^{T_r}(\gamma_{r,t}(s;W_r)-\gamma_{r,t}(s))\,\frac{\partial}{\partial \lambda_{i,s}}\log p_\Psi(O_{r,t}|W_{r,t})$$

As in our case (independent of the state s)

$$\begin{aligned}\frac{\partial}{\partial \lambda_i}\log p_\Psi(O_r|W_r) &= \frac{\partial}{\partial \lambda_i}\sum_j \lambda_j \log p_j(O_r|W_r)\\ &= \log p_i(O_r|W_r)\end{aligned}$$

we can write

$$\frac{\partial F}{\partial \lambda_i} = \sum_{r=1}^{R}\sum_{t=1}^{T_r}(\gamma_{r,t}(s;W_r)-\gamma_{r,t}(s))\log p_i(O_{r,t}|s)$$

Defining

$$\begin{aligned}\Phi_i^{\text{NUM}} &:= \sum_{r=1}^{R}\sum_{s\in\mathcal{R}}\gamma_{r,t}(s;W_r)\log p_i(O_{r,t}|s)\\ \Phi_i^{\text{DEN}} &:= \sum_{r=1}^{R}\sum_{s\in\mathcal{S}}\gamma_{r,t}(s)\log p_i(O_{r,t}|s)\end{aligned}$$

the update equation can now be written as follows:

$$\begin{aligned}\lambda_i^{(I+1)} &= \lambda_i^{(I)} + \epsilon\frac{\partial}{\partial\lambda}F(\lambda)\\ &= \lambda_i^{(I)} + \epsilon(\Phi_i^{\text{NUM}} - \Phi_i^{\text{DEN}})\end{aligned}$$

Here, the enumeration $s \in \mathcal{R}$ is over all reference states ("numerator lattice") and $s \in \mathcal{S}$ is over all states given by the recognizer output ("denominator lattice").

q.e.d.

A more detailed discussion of some of the steps, particularly those involving the exploitation of the Markov chain and the definition of the FB probabilities, can be found in [Mac98] and [SMMN01].

Appendix B

System descriptions

B.1 GlobalPhone Systems

Training: GlobalPhone corpus recorded in 16kHz/ 16bit with close-talking microphones and quiet environment [SWW97] ("Wall-Street-Journal" (WSJ) style); EN data taken from WSJ-0 corpus

Language	# Speakers	Duration (h)	Utterances
CH	132	27	8663
EN	103	15	7137
GE	77	17	9259
JA	144	24	9234
SP	100	18	5426

Test: test data was recorded under the same conditions as training data

Development test set			
Language	# Speakers	Duration (h)	Utterances
CH	10	0.7	250
EN	10	0.4	144
GE	3	0.4	199
JA	11	0.7	250
SP	10	0.7	250

Evaluation set			
Language	# Speakers	Duration (h)	Utterances
CH	10	0.7	240
EN	10	0.4	152
GE	3	0.4	250
JA	5	0.7	250
SP	8	0.7	250

Pre-processing: the same pre-processing was applied to all languages: 10ms frame-shift, ML-VTLN, per-utterance CMS, 32-dimensional feature space after LDA computed on MFCCs with Δ, ΔΔ, zero-crossing-rate and power coefficients

Acoustic models: 3000 fully continuous models trained with 4 iterations of maximum likelihood label training, 32 Gaussians per model, diagonal covariances (for all languages)

Dictionary and LM: English

Dictionary: 17k vocabulary

Language model: 9k unigrams, 1.6M bigrams, 6.6M trigrams, PPT=252, OOV=0.1% trained on CSR 1994 LM (200M words)

Phone Set: English

```
; -------------------------------------------------------
;  Name           : ps
;  Type           : PhonesSet
;  Number of Items : 2
;  Date           : Sat Oct 28 13:48:39 1995
; -------------------------------------------------------
PHONES          @ M_+QK/EN M_+hGH/EN M_il/EN M_ip/EN M_ae/EN M_ale/EN M_i2/EN M_etu/EN M_ov/EN M_u/EN
                M_vst/EN M_oc/EN M_ab/EN M_eI/EN M_aIp/EN M_ocI/EN M_aVs/EN M_oVs/EN M_l/EN M_r9/EN M_j/EN
                M_w/EN M_r3/EN M_ETr/EN M_m/EN M_n/EN M_ng/ENM_tS/EN M_dZ/EN M_D/EN M_b/EN M_d/EN
                M_rfd/EN M_g/EN M_ph/EN M_th/EN M_kh/EN M_z/EN M_Z/EN M_v/EN M_f/EN M_T/EN M_s/EN M_S/EN M_h/EN SIL
NOISES          M_+QK/EN M_+hGH/EN
SILENCES        SIL
CONSONANT       M_ph/EN M_b/EN M_f/EN M_v/EN M_T/EN M_D/EN M_th/EN M_d/EN M_s/EN M_z/EN M_S/EN M_Z/EN M_tS/EN M_dZ/EN
                M_kh/EN M_g/EN M_h/EN M_m/EN M_n/EN M_ng/EN M_r9/EN M_j/EN M_w/EN M_l/EN M_r3/EN M_rfd/EN M_ETr/EN
CONSONANTAL     M_ph/EN M_b/EN M_f/EN M_v/EN M_T/EN M_D/EN M_th/EN M_d/EN M_s/EN M_z/EN M_S/EN M_Z/EN M_tS/EN
                M_dZ/EN M_kh/EN M_g/EN M_h/EN M_m/EN M_n/EN M_ng/EN M_rfd/EN
OBSTRUENT       M_ph/EN M_b/EN M_f/EN M_v/EN M_T/EN M_D/EN M_th/EN M_d/EN M_s/EN M_z/EN M_S/EN M_Z/EN M_tS/EN
                M_dZ/EN M_kh/EN M_g/EN
SONORANT        M_m/EN M_n/EN M_ng/EN M_r9/EN M_j/EN M_w/EN M_l/EN M_r3/EN M_ETr/EN M_rfd/EN
SYLLABIC        M_aIp/EN M_ocI/EN M_eI/EN M_il/EN M_aVs/EN M_oVs/EN M_ae/EN M_ip/EN M_oc/EN M_ale/EN
                M_ab/EN M_ov/EN M_u/EN M_vst/EN M_i2/EN M_etu/EN M_r3/EN M_ETr/EN
VOWEL           M_aIp/EN M_ocI/EN M_eI/EN M_il/EN M_aVs/EN M_oVs/EN M_ae/EN M_ip/EN M_oc/EN M_ale/EN
                M_ab/EN M_ov/EN M_u/EN M_vst/EN M_i2/EN M_etu/EN
DIPHTHONG       M_aIp/EN M_ocI/EN M_eI/EN M_aVs/EN M_oVs/EN
CARDVOWEL       M_il/EN M_ip/EN M_ae/EN M_ale/EN M_ab/EN M_ov/EN M_oc/EN M_vst/EN M_u/EN M_i2/EN M_etu/EN
STIMMHAFT       M_b/EN M_d/EN M_g/EN M_dZ/EN M_v/EN M_D/EN M_z/EN M_Z/EN M_m/EN M_n/EN M_ng/EN M_w/EN M_r9/EN
                M_j/EN M_l/EN M_r3/EN M_aIp/EN M_ocI/EN M_eI/EN M_il/EN M_aVs/EN M_oVs/EN M_ae/EN M_ip/EN M_oc/EN
                M_ale/EN M_ab/EN M_ov/EN M_u/EN M_vst/EN M_rfd/EN M_ETr/EN M_i2/EN M_etu/EN M_b/EN M_v/EN M_D/EN
                M_d/EN M_z/EN M_Z/EN M_dZ/EN M_g/EN M_m/EN M_n/EN M_ng/EN M_j/EN M_w/EN M_l/EN M_ETr/EN
VOICED          M_b/EN M_d/EN M_g/EN M_dZ/EN M_v/EN M_D/EN M_z/EN M_Z/EN M_m/EN M_n/EN M_ng/EN M_w/EN M_r9/EN M_j/EN
                M_l/EN M_r3/EN M_aIp/EN M_ocI/EN M_eI/EN M_il/EN M_aVs/EN M_oVs/EN M_ae/EN M_ip/EN M_oc/EN M_ale/EN
                M_ab/EN M_ov/EN M_u/EN M_vst/EN M_rfd/EN M_ETr/EN M_i2/EN M_etu/EN
UNVOICED        M_ph/EN M_f/EN M_T/EN M_th/EN M_s/EN M_S/EN M_tS/EN M_kh/EN
CONTINUANT      M_f/EN M_T/EN M_s/EN M_S/EN M_v/EN M_D/EN M_z/EN M_Z/EN M_w/EN M_r9/EN M_j/EN M_l/EN M_r3/EN
DEL-REL         M_tS/EN M_dZ/EN
LATERAL         M_l/EN
ANTERIOR        M_ph/EN M_th/EN M_b/EN M_d/EN M_f/EN M_T/EN M_s/EN M_S/EN M_v/EN M_D/EN M_z/EN M_Z/EN M_m/EN
                M_n/EN M_w/EN M_j/EN M_l/EN M_rfd/EN
CORONAL         M_th/EN M_d/EN M_tS/EN M_dZ/EN M_T/EN M_s/EN M_S/EN M_D/EN M_z/EN M_Z/EN M_n/EN M_l/EN M_r9/EN M_rfd/EN
APICAL          M_th/EN M_d/EN M_n/EN M_rfd/EN
HIGH-CONS       M_kh/EN M_g/EN M_ng/EN M_w/EN M_j/EN
BACK-CONS       M_kh/EN M_g/EN M_ng/EN M_w/EN
LABIALIZED      M_r9/EN M_w/EN M_r3/EN M_ETr/EN
STRIDENT        M_tS/EN M_dZ/EN M_f/EN M_s/EN M_S/EN M_v/EN M_z/EN M_Z/EN
SIBILANT        M_s/EN M_S/EN M_z/EN M_Z/EN M_tS/EN M_dZ/EN
BILABIAL        M_ph/EN M_b/EN M_m/EN M_w/EN
LABIODENTAL     M_f/EN M_v/EN
LABIAL          M_ph/EN M_b/EN M_m/EN M_w/EN M_f/EN M_v/EN
INTERDENTAL     M_T/EN M_D/EN
ALVEOLAR-RIDGE  M_th/EN M_d/EN M_n/EN M_s/EN M_z/EN M_l/EN M_rfd/EN
ALVEOPALATAL    M_S/EN M_Z/EN M_tS/EN M_dZ/EN
ALVEOLAR        M_th/EN M_d/EN M_n/EN M_s/EN M_z/EN M_l/EN M_S/EN M_Z/EN M_tS/EN M_dZ/EN M_rfd/EN
RETROFLEX       M_r9/EN M_r3/EN M_ETr/EN
```

```
PALATAL         M_j/EN
VELAR           M_kh/EN M_g/EN M_ng/EN M_w/EN
GLOTTAL         M_h/EN
ASPIRATED       M_h/EN
STOP            M_ph/EN M_b/EN M_th/EN M_d/EN M_kh/EN M_g/EN M_m/EN M_n/EN M_ng/EN
PLOSIVE         M_ph/EN M_b/EN M_th/EN M_d/EN M_kh/EN M_g/EN
FLAP            M_rfd/EN
NASAL           M_m/EN M_n/EN M_ng/EN
FRICATIVE       M_f/EN M_v/EN M_T/EN M_D/EN M_s/EN M_z/EN M_S/EN M_Z/EN M_h/EN
AFFRICATE       M_tS/EN M_dZ/EN
APPROXIMANT     M_r9/EN M_l/EN M_j/EN M_w/EN
LAB-PL          M_ph/EN M_b/EN
ALV-PL          M_th/EN M_d/EN
VEL-PL          M_kh/EN M_g/EN
VLS-PL          M_ph/EN M_th/EN M_kh/EN
VCD-PL          M_b/EN M_d/EN M_g/EN
LAB-FR          M_f/EN M_v/EN
DNT-FR          M_T/EN M_D/EN
ALV-FR          M_S/EN M_Z/EN
VLS-FR          M_f/EN M_T/EN M_S/EN
VCD-FR          M_v/EN M_D/EN M_Z/EN
ROUND           M_oc/EN M_oVs/EN M_vst/EN M_u/EN M_ocI/EN M_aVs/EN M_oVs/EN
HIGH-VOW        M_il/EN M_ip/EN M_vst/EN M_u/EN M_i2/EN
MID-VOW         M_ae/EN M_ov/EN M_etu/EN
LOW-VOW         M_ab/EN M_ale/EN M_oc/EN
FRONT-VOW       M_il/EN M_ip/EN M_ae/EN M_ale/EN
CENTRAL-VOW     M_ov/EN M_etu/EN M_i2/EN
BACK-VOW        M_ab/EN M_oc/EN M_vst/EN M_u/EN
TENSE-VOW       M_il/EN M_u/EN M_ale/EN
LAX-VOW         M_ip/EN M_ab/EN M_ae/EN M_ov/EN M_vst/EN
ROUND-VOW       M_oc/EN M_vst/EN M_u/EN
REDUCED-VOW     M_i2/EN M_etu/EN
REDUCED-CON     M_ETr/EN
REDUCED         M_i2/EN M_etu/EN M_ETr/EN
LH-DIP          M_aIp/EN M_aVs/EN
MH-DIP          M_ocI/EN M_oVs/EN M_eI/EN
BF-DIP          M_aIp/EN M_ocI/EN M_aVs/EN M_oVs/EN
Y-DIP           M_aIp/EN M_ocI/EN M_eI/EN
W-DIP           M_aVs/EN M_oVs/EN
ROUND-DIP       M_ocI/EN M_aVs/EN M_oVs/EN
LIQUID-GLIDE    M_l/EN M_r9/EN M_w/EN M_j/EN
W-GLIDE         M_u/EN M_aVs/EN M_oVs/EN M_w/EN
LIQUID          M_l/EN M_r9/EN
LW              M_l/EN M_w/EN
Y-GLIDE         M_il/EN M_aIp/EN M_eI/EN M_ocI/EN M_j/EN
LQGL-BACK       M_l/EN M_r9/EN M_w/EN
```

B.2 English Spontaneous Scheduling Task System

Training: mixture of BN'96 and Verbmobil I+II (ESST) data

BN'96 training set: 66h, 6467 manually labeled speaker clusters

ESST training set: 32h, 2208 speakers consisting of Verbmobil `CDs 6, 8, 9, 10, 13, 23, 28, 31, 32, 39, 42, 43, 47, 50, 51, 52, 55, 56` unless dialog marked as test data

Total: 98h in 16kHz/ 16bit quality, varying acoustic conditions, Verbmobil corpus is close-talking, spontaneous speech in tourism and scheduling domain

Test: test data is taken from Verbmobil II corpus only.

Development test data `ds2`: `AHS_e056ach1, AHS_e057ach1, BJC_e125ach1, BJC_e126ach1, BJC_e127ach1, BJC_e128ach1, CLW_e044ach1, CLW_e045ach1, DRC_e125ach2, DRC_e126ach2, DRC_e127ach2, DRC_e128ach2, JLF_e100ach1, JLF_e101ach1, JLF_e102ach1,` JLF_e115ach2, MBB_e044ach2, MBB_e045ach2, SNC_e094ach1, SNC_e095ach1, SNC_e096ach1, SNC_e097ach1, SNC_e100ach2, SNC_e101ach2, SNC_e102ach2, SNC_e115ach1, VNC_e094ach2, VNC_e095ach2, VNC_e096ach2, VNC_e097ach2, WJH_e056ach2, WJH_e057ach2 (32 dialogs, 9 speakers)

Evaluation data xv2**:** BAT_e116ach1, BAT_e117ach1, BAT_e118ach1, BAT_e119ach1, BAT_e123ach2, BAT_e124ach2, BMJ_e120ach1, BMJ_e121ach1, BMJ_e122ach1, DNC_e029ach2, DNC_e030ach2, DNC_e031ach2, DNC_e032ach2, JDH_e116ach2, JDH_e117ach2, JDH_e118ach2, JDH_e119ach2, KRA_e123ach1, KRA_e124ach1, RGM_e029ach1, RGM_e030ach1, RGM_e031ach1, RGM_e032ach1, TAJ_e120ach2, TAJ_e121ach2, TAJ_e122ach2 (26 dialogs, 7 speakers)

Pre-processing: 10ms frame shift, ML-VTLN, per-dialog (-cluster) CMS/ CVN, 40-dimensional feature space after LDA computed on ±3 frames context window, global STC matrix

Acoustic models: 4000 fully continuous models trained with 6 iterations of maximum likelihood label training, 32 Gaussians per model, diagonal covariances, global STC matrix

Dictionary and LM:

Training: 40k words, 49k pronunciation variants (BN'96 and ESST merged)

Testing: 7100 words, 9400 pronunciation variants

Language model: 7k unigrams, 39k bigrams, 119k trigrams, PPT=43, OOV rate=1% trained on ESST training data

Phone Set:

```
; -------------------------------------------------------
;  Name           : ps
;  Type           : PhonesSet
;  Number of Items : 2
;  Date           :
;  Remarks: DX->T, add XL/XM/XN
;    removed DX from SONORANT & VOICED, added X-LMN class
; -------------------------------------------------------
PHONES          PAD IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER AXR M N
                NG CH JH DH B  D  G  P  T K Z ZH V  F  TH S  SH HH XL XM XN SIL GARBAGE
                +FILLER+ +BREATH+
HUMANSND        IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER AXR M N
                NG CH JH DH B  D  G  P  T K Z ZH V  F  TH S  SH HH XL XM XN
VOLATILE        AO EY AY OY AW OW L R Y W ER AXR M N NG CH JH DH B D G P T K Z ZH V F
                TH S SH HH XL XM XN
NOISES          GARBAGE +BREATH+ +FILLER+
FILLERS         +FILLER+
BREATH          +BREATH+
SILENCES        SIL
CONSONANT       P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG R Y W L ER AXR XL XM XN
CONSONANTAL     P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG XL XM XN
OBSTRUENT       P B F V TH DH T D S Z SH ZH CH JH K G
SONORANT        M N NG R Y W L ER AXR XL XM XN
```

```
SYLLABIC        AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX ER AXR XL XM XN
VOWEL           AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX
DIPHTHONG       AY OY EY AW OW
CARDVOWEL       IY IH EH AE AA AH AO UH UW IX AX
VOICED          B D G JH V DH Z ZH M N NG W R Y L ER AY OY EY IY AW OW EH IH AO AE AA AH
                UW UH AXR IX AX XL XM XN
UNVOICED        P F TH T S SH CH K
CONTINUANT      F TH S SH V DH Z ZH W R Y L ER XL
LATERAL         L XL
ANTERIOR        P T B D F TH S SH V DH Z ZH M N W Y L XM XN
CORONAL         T D CH JH TH S SH DH Z ZH N L R XL XN
APICAL          T D N
HIGH-CONS       K G NG W Y
BACK-CONS       K G NG W
LABIALIZED      R W ER AXR
STRIDENT        CH JH F S SH V Z ZH
SIBILANT        S SH Z ZH CH JH
BILABIAL        P B M W
LABIAL          P B M W F V
ALVEOLAR-RIDGE  T D N S Z L
ALVEOPALATAL    SH ZH CH JH
ALVEOLAR        T D N S Z L SH ZH CH JH
RETROFLEX       R ER AXR
PALATAL         Y
GLOTTAL         HH
STOP            P B T D K G M N NG
PLOSIVE         P B T D K G
NASAL           M N NG XM XN
FRICATIVE       F V TH DH S Z SH ZH HH
AFFRICATE       CH JH
APPROXIMANT     R L Y W
LAB-PL          P B
ALV-PL          T D
VEL-PL          K G
VLS-PL          P T K
VCD-PL          B D G
LAB-FR          F V
DNT-FR          TH DH
ALV-FR          SH ZH
VLS-FR          F TH SH
VCD-FR          V DH ZH
ROUND           AO OW UH UW OY AW OW
HIGH-VOW        IY IH UH UW IX
MID-VOW         EH AH AX
LOW-VOW         AA AE AO
FRONT-VOW       IY IH EH AE
CENTRAL-VOW     AH AX IX
BACK-VOW        AA AO UH UW
TENSE-VOW       IY UW AE
LAX-VOW         IH AA EH AH UH
ROUND-VOW       AO UH UW
REDUCED-VOW     IX AX
REDUCED-CON     AXR
REDUCED         IX AX AXR
LH-DIP          AY AW
MH-DIP          OY OW EY
BF-DIP          AY OY AW OW
Y-DIP           AY OY EY
W-DIP           AW OW
ROUND-DIP       OY AW OW
W-GLIDE         UW AW OW W
LIQUID          L R
LW              L W
Y-GLIDE         IY AY EY OY Y
LQGL-BACK       L R W
X-LMN           XL XM XN
```

B.3 ReadBN System

Training, system setup, and pre-processing identical to ESST system described in Appendix B.2. Test data consists of 198 sentences re-read under F0-like conditions by 2 speakers:

Rob Malkin this speaker also appears as "RGM" in the ESST database (part of `xv2` evaluation data)

Michael Bett

Total duration is 17min.

B.4 Meeting System

Training: mixture of BN'96, BN'97 and RT-04S "Meeting" corpus

BN data: 137h after segmentation, 3912 automatically determined speaker clusters from BN'96 and BN'97 corpora

RT-04S "Meeting" data: close-talking data from NIST training set recorded in different meeting rooms using different setups

Site	Duration	Meetings	Speakers
CMU	11h	21	93
ICSI	72h	75	455
NIST	13h	15	77

Test: official NIST RT-04S "Meeting" development and evaluation test data

Dev data: 2090 segments marked in STM reference file, 43 speakers and 8 meetings (2 each from CMU, ICSI, LDC, NIST), total 89 minutes

Eval data: 2502 segments resulting from STM reference file, 40 speakers and 8 meetings (2 each from CMU, ICSI, LDC, NIST), total 100 minutes

Pre-processing: 10ms frame shift in training, 8ms frame shift during test; ML-VTLN, per-utterance CMS/ CVN, 42-dimensional feature space after LDA on ±3 frames context window, global STC matrix

Acoustic models: 24000 semi-continuous HMM states tied over 6000 models, up to 64 Gaussians per codebook, 300k Gaussians in total trained with merge & split training and 2 iterations of Viterbi training

Dictionary and LM:

Training: 47k words, 55k pronunciation variants

Testing: 55k vocabulary

Language model: 47k vocabulary, PPT=112, OOV rate=1%, 3-fold interpolation consisting of

3-gram LM trained on SWB+Meeting data (252k 3-grams)

4-gram LM trained on BN (3.3M 4-grams)

5-gram LM trained on SWB+Meeting (800 automatically clustered classes, 200k 5-grams)

Phone Set:

```
; -------------------------------------------------------
;  Name            : ps
;  Type            : PhonesSet
;  Number of Items : 2
;  Date            :
;  Remarks: DX->T, add XL/XM/XN
;    removed DX from SONORANT & VOICED, added X-LMN class
; -------------------------------------------------------
PHONES          PAD IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER A
                XR M N  NG CH JH DH B  D  G  P  T K Z ZH V  F  TH S  SH HH XL XM XN SIL GARBAGE
                +FILLER+ +BREATH+
HUMANSND        IY IH EH AE IX AX AH UW UH AO AA EY AY OY AW OW L R Y W ER AXR M
                N  NG CH JH DH B  D  G  P  T K Z ZH V  F  TH S  SH HH XL XM XN
VOLATILE        AO EY AY OY AW OW L R Y W ER AXR M N NG CH JH DH B D G P T K Z Z
                H V F TH S SH HH XL XM XN
NOISES          GARBAGE +BREATH+ +FILLER+
FILLERS         +FILLER+
BREATH          +BREATH+
SILENCES        SIL
CONSONANT       P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG R Y W L ER AXR
                XL XM XN
CONSONANTAL     P B F V TH DH T D S Z SH ZH CH JH K G HH M N NG XL XM XN
OBSTRUENT       P B F V TH DH T D S Z SH ZH CH JH K G
SONORANT        M N NG R Y W L ER AXR XL XM XN
SYLLABIC        AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX ER AXR XL XM XN
VOWEL           AY OY EY IY AW OW EH IH AO AE AA AH UW UH IX AX
DIPHTHONG       AY OY EY AW OW
CARDVOWEL       IY IH EH AE AA AH AO UH UW IX AX
VOICED          B D G JH V DH Z ZH M N NG W R Y L ER AY OY EY IY AW OW EH IH AO
                AE AA AH UW UH AXR IX AX XL XM XN
UNVOICED        P F TH T S SH CH K
CONTINUANT      F TH S SH V DH Z ZH W R Y L ER XL
DEL-REL         CH JH
LATERAL         L XL
ANTERIOR        P T B D F TH S SH V DH Z ZH M N W Y L XM XN
CORONAL         T D CH JH TH S SH DH Z ZH N L R XL XN
APICAL          T D N
HIGH-CONS       K G NG W Y
BACK-CONS       K G NG W
LABIALIZED      R W ER AXR
STRIDENT        CH JH F S SH V Z ZH
SIBILANT        S SH Z ZH CH JH
BILABIAL        P B M W
LABIODENTAL     F V
LABIAL          P B M W F V
INTERDENTAL     TH DH
ALVEOLAR-RIDGE  T D N S Z L
ALVEOPALATAL    SH ZH CH JH
ALVEOLAR        T D N S Z L SH ZH CH JH
RETROFLEX       R ER AXR
PALATAL         Y
VELAR           K G NG W
GLOTTAL         HH
ASPIRATED       HH
STOP            P B T D K G M N NG
PLOSIVE         P B T D K G
NASAL           M N NG XM XN
FRICATIVE       F V TH DH S Z SH ZH HH
AFFRICATE       CH JH
APPROXIMANT     R L Y W
LAB-PL          P B
ALV-PL          T D
VEL-PL          K G
VLS-PL          P T K
VCD-PL          B D G
LAB-FR          F V
DNT-FR          TH DH
ALV-FR          SH ZH
VLS-FR          F TH SH
VCD-FR          V DH ZH
ROUND           AO OW UH UW OY AW OW
HIGH-VOW        IY IH UH UW IX
MID-VOW         EH AH AX
LOW-VOW         AA AE AO
FRONT-VOW       IY IH EH AE
CENTRAL-VOW     AH AX IX
```

```
BACK-VOW         AA AO UH UW
TENSE-VOW        IY UW AE
LAX-VOW          IH AA EH AH UH
ROUND-VOW        AO UH UW
REDUCED-VOW      IX AX
REDUCED-CON      AXR
REDUCED          IX AX AXR
LH-DIP           AY AW
MH-DIP           OY OW EY
BF-DIP           AY OY AW OW
Y-DIP            AY OY EY
W-DIP            AW OW
ROUND-DIP        OY AW OW
LIQUID-GLIDE     L R W Y
W-GLIDE          UW AW OW W
LIQUID           L R
LW               L W
Y-GLIDE          IY AY EY OY Y
LQGL-BACK        L R W
X-LMN            XL XM XN
```

Appendix C

Result Tables

C.1 Feature Classification Rates

C.1.1 ReadBN and ESST Classification Rates

Feature/ Task	ReadBN		ESST
Test on Frames	Middle	All	All
UNVOICED	91.0%	84.5%	80.8%
ROUND	89.6%	88.5%	87.9%
STOP	87.3%	78.9%	74.6%
VOWEL	84.6%	77.2%	76.2%
LATERAL	95.0%	94.3%	95.0%
NASAL	94.2%	91.8%	90.1%
FRICATIVE	92.1%	86.2%	84.0%
LABIAL	90.2%	90.2%	85.7%
CORONAL	78.3%	72.0%	70.5%
PALATAL	96.7%	96.6%	96.2%
GLOTTAL	98.8%	97.9%	97.3%
HIGH-VOW	87.6%	85.7%	86.3%
MID-VOW	83.7%	80.4%	85.6%
LOW-VOW	90.3%	89.9%	91.4%
FRONT-VOW	84.8%	81.2%	84.8%
BACK-VOW	91.4%	90.8%	91.8%
RETROFLEX	95.9%	94.1%	94.7%
OBSTRUENT	90.6%	81.3%	79.6%
ALV-FR	99.1%	98.9%	99.3%
OVERALL	90.8%	87.8%	87.3%

Table C.1: Feature classification accuracy for selected features on the ReadBN and ESST tasks (English language).

C.1.2 GlobalPhone Classification Rates

Feature	CH	EN	GE	JA	SP
LABIODENTAL	98.46%	96.77%	93.09%	96.36%	98.02%
VOICED	97.73%	83.16%	85.66%	89.51%	84.79%
APPROXIMANT	97.53%	91.80%	95.12%	92.46%	93.02%
TONAL5	97.02%	—	—	—	—
LATERAL-APPROXIMANT	96.72%	92.39%	92.90%	92.42%	88.44%
BACK	96.39%	90.12%	95.28%	72.63%	90.87%
FRICATIVE	96.31%	88.83%	90.95%	93.07%	85.30%
PLOSIVE	96.27%	88.04%	90.75%	89.63%	84.89%
OPEN	95.64%	95.45%	92.96%	90.53%	89.88%
ASPIRATED	95.46%	90.79%	—	—	—
BILABIAL	94.95%	91.05%	86.65%	88.63%	90.90%
CONSONANT	94.87%	85.03%	87.23%	85.23%	71.20%
VOWEL	94.81%	84.65%	87.51%	84.83%	70.42%
NASAL	94.78%	91.53%	90.27%	87.37%	79.65%
ROUND	94.70%	93.48%	93.53%	87.85%	90.19%
AFFRICATE	94.58%	88.47%	92.49%	91.19%	86.94%
UNVOICED	94.51%	80.66%	83.26%	85.53%	76.32%
PALATAL	94.19%	87.48%	91.35%	86.79%	87.77%
CLOSE	94.10%	92.88%	89.16%	84.40%	91.65%
OPEN-MID	93.31%	84.94%	88.29%	—	—
RETROFLEX	91.57%	83.69%	—	—	—
VELAR	91.29%	88.48%	82.84%	82.79%	82.70%
TONAL3	90.39%	—	—	—	—
FRONT	89.71%	78.98%	80.98%	79.67%	70.23%
UNROUND	89.04%	78.52%	79.06%	76.29%	69.26%
TONAL2	88.30%	—	—	—	—
ALVEOLAR	87.62%	70.96%	71.83%	78.35%	65.23%
TONAL1	87.54%	—	—	—	—
TONAL4	84.37%	—	—	—	—

Table C.2: Chinese AF Detectors.

Feature	EN	CH	GE	JA	SP
POSTALVEOLAR	99.25%	—	98.67%	96.38%	94.92%
PALATAL	99.00%	89.90%	96.13%	97.16%	96.85%
GLOTTAL	98.84%	—	97.22%	96.64%	—
FLAP	98.84%	—	—	—	94.50%
AFFRICATE	98.63%	91.01%	96.74%	96.21%	99.44%
LABIODENTAL	97.99%	98.98%	94.26%	98.08%	97.95%
LATERAL-APPROXIMANT	97.39%	91.74%	91.06%	90.74%	88.90%
DENTAL	97.04%	—	—	—	91.65%
NASAL	96.66%	90.82%	94.49%	93.19%	91.76%
ROUND	95.55%	90.72%	91.09%	85.59%	88.70%
OPEN	95.54%	92.87%	94.94%	89.69%	88.07%
VELAR	95.48%	87.86%	91.27%	91.59%	92.18%
RETROFLEX	95.28%	90.14%	—	—	—
FRICATIVE	94.71%	91.88%	89.12%	91.93%	90.59%
BILABIAL	93.86%	94.81%	89.41%	91.63%	92.08%
ASPIRATED	93.81%	90.99%	—	—	—
APPROXIMANT	93.79%	92.22%	92.08%	93.88%	93.69%
CLOSE	93.40%	87.57%	88.07%	85.02%	88.48%
PLOSIVE	92.99%	89.00%	89.74%	89.29%	84.09%
BACK	91.38%	85.65%	85.36%	76.49%	86.64%
VOWEL	91.08%	86.24%	87.34%	86.76%	78.81%
CONSONANT	91.03%	85.64%	87.44%	85.47%	78.34%
OPEN-MID	90.92%	87.32%	87.59%	—	—
UNVOICED	90.46%	83.69%	83.88%	86.35%	84.72%
CLOSE-MID	89.87%	—	83.14%	76.59%	80.45%
FRONT	89.65%	78.35%	82.99%	85.19%	77.67%
VOICED	89.31%	81.36%	83.52%	86.23%	84.54%
CENTRAL	88.81%	—	86.50%	—	—
ALVEOLAR	87.43%	70.59%	72.14%	71.78%	73.97%
UNROUND	86.76%	76.29%	84.10%	79.84%	78.49%

Table C.3: English AF Detectors.

Feature	GE	CH	EN	JA	SP
POSTALVEOLAR	99.46%	—	98.20%	96.60%	94.29%
APPROXIMANT	98.86%	96.87%	93.02%	95.60%	95.97%
AFFRICATE	98.31%	90.33%	97.45%	95.59%	98.80%
PALATAL	98.20%	90.98%	97.83%	94.52%	95.13%
GLOTTAL	97.90%	—	96.06%	94.72%	—
OPEN-MID	97.11%	89.50%	92.17%	—	—
OPEN	95.36%	88.64%	93.77%	88.40%	85.68%
BACK	95.00%	93.99%	89.69%	75.11%	90.35%
LABIODENTAL	94.39%	97.92%	95.13%	94.95%	96.23%
NASAL	94.11%	89.54%	91.14%	89.90%	86.22%
LATERAL-APPROXIMANT	93.97%	94.71%	95.60%	90.23%	89.01%
FRICATIVE	93.94%	90.75%	82.82%	92.09%	85.99%
PLOSIVE	93.81%	92.40%	87.67%	87.69%	78.22%
ROUND	93.79%	92.26%	94.29%	88.25%	90.08%
TRILL	93.46%	—	—	—	85.13%
VOICED	92.36%	83.02%	75.46%	83.75%	75.73%
UNVOICED	91.77%	84.79%	73.99%	82.81%	74.79%
VOWEL	91.77%	86.98%	75.09%	77.23%	63.79%
CONSONANT	91.06%	85.75%	73.07%	77.86%	65.53%
VELAR	90.66%	87.05%	91.74%	87.20%	84.90%
FRONT	90.41%	77.31%	81.27%	83.85%	70.62%
CENTRAL	89.88%	—	91.63%	—	—
BILABIAL	89.27%	95.43%	93.15%	92.52%	91.32%
UNROUND	89.15%	77.20%	77.70%	76.73%	70.00%
CLOSE	89.01%	87.29%	88.78%	81.21%	83.10%
CLOSE-MID	86.74%	—	90.98%	78.46%	76.15%
ALVEOLAR	79.54%	75.84%	67.63%	69.89%	57.36%

Table C.4: German AF Detectors. Note that the German detectors for Central perform better on the English data than the English detectors.

Feature	JA	CH	EN	GE	SP
LABIODENTAL	99.23%	98.70%	95.88%	94.50%	98.23%
PALATAL	97.96%	89.47%	97.23%	94.09%	95.03%
POSTALVEOLAR	97.71%	—	95.00%	96.11%	91.15%
GLOTTAL	97.50%	—	96.15%	91.85%	—
OPEN	97.23%	87.19%	86.99%	91.12%	91.66%
APPROXIMANT	96.99%	94.68%	91.69%	94.43%	94.89%
AFFRICATE	96.80%	93.88%	94.50%	96.60%	97.51%
ROUND	96.61%	88.25%	87.74%	90.00%	91.32%
LATERAL-APPROXIMANT	96.59%	96.02%	92.83%	92.26%	91.40%
UVULAR	96.24%	—	—	—	—
FRICATIVE	95.40%	91.96%	89.00%	90.41%	87.17%
FRONT	95.25%	77.47%	85.31%	79.78%	76.62%
BILABIAL	94.94%	96.61%	92.70%	91.80%	93.07%
NASAL	94.84%	90.63%	93.35%	92.74%	89.70%
PLOSIVE	94.72%	92.63%	87.44%	87.65%	90.82%
VOICED	94.68%	84.94%	83.86%	83.96%	87.82%
VELAR	94.40%	85.78%	91.35%	88.71%	90.92%
UNVOICED	94.00%	85.81%	83.89%	84.35%	87.58%
CLOSE-MID	93.78%	—	83.61%	82.46%	83.44%
CONSONANT	93.73%	85.48%	81.48%	80.37%	81.68%
VOWEL	93.53%	83.95%	81.81%	79.50%	80.53%
BACK	93.37%	68.05%	74.97%	71.58%	76.77%
CLOSE	92.56%	82.74%	88.92%	82.17%	82.53%
UNROUND	92.48%	73.12%	77.01%	77.02%	77.10%
ALVEOLAR	89.92%	81.86%	70.95%	69.08%	72.97%

Table C.5: Japanese AF Detectors.

Feature	SP	CH	EN	GE	JA
AFFRICATE	99.33%	91.45%	95.56%	96.76%	95.05%
LABIODENTAL	98.84%	98.78%	96.35%	94.14%	97.67%
TRILL	98.12%	—	—	91.43%	—
APPROXIMANT	97.05%	89.84%	90.41%	92.69%	95.26%
DENTAL	96.49%	—	96.68%	—	—
PALATAL	96.43%	84.35%	94.56%	91.63%	94.04%
VOICED	95.84%	87.00%	83.76%	78.70%	90.18%
UNVOICED	94.92%	86.04%	83.28%	77.77%	88.08%
CLOSE	94.39%	88.64%	91.52%	87.79%	81.48%
NASAL	94.07%	87.59%	91.04%	87.75%	91.45%
FRICATIVE	93.91%	88.71%	87.07%	83.03%	89.26%
PLOSIVE	93.43%	86.39%	84.43%	77.19%	83.12%
BACK	93.31%	88.31%	85.43%	86.87%	80.85%
ROUND	93.31%	87.13%	88.28%	85.99%	92.89%
LATERAL-APPROXIMANT	93.26%	89.81%	89.94%	88.63%	92.49%
FLAP	93.07%	—	87.90%	—	—
POSTALVEOLAR	93.07%	—	86.59%	86.19%	90.79%
OPEN	92.87%	86.90%	83.30%	89.93%	95.42%
BILABIAL	92.81%	90.89%	90.06%	84.81%	84.64%
VELAR	92.00%	81.29%	85.86%	81.00%	82.86%
CONSONANT	90.76%	73.45%	76.43%	70.06%	82.68%
VOWEL	90.47%	70.13%	77.53%	67.37%	82.75%
UNROUND	90.42%	72.13%	79.05%	74.11%	84.17%
FRONT	90.42%	71.88%	75.88%	77.97%	80.94%
CLOSE-MID	87.98%	—	82.91%	78.22%	86.65%
ALVEOLAR	83.34%	79.18%	75.33%	69.41%	75.71%

Table C.6: Spanish AF Detectors.

	Test Set					
Languages	CH	DE	EN	JA	SP	Train
CH_DE_EN_JA	91,82%	90,05%	91,04%	91,45%	87,06%	92,66%
CH_DE_EN_SP	91,56%	89,87%	90,68%	88,02%	88,48%	91,99%
CH_DE_JA_SP	91,07%	89,30%	88,27%	90,99%	88,16%	91,93%
CH_EN_JA_SP	91,28%	88,04%	90,81%	91,65%	89,05%	92,58%
DE_EN_JA_SP	89,51%	90,05%	91,78%	92,25%	89,49%	92,37%

Table C.7: MM4 Detectors

	Test Set					
Languages	CH	DE	EN	JA	SP	Train
CH_DE_EN_JA_SP	90,36%	89,00%	90,22%	90,77%	88,29%	91,32%

Table C.8: MM5 Detectors

C.2 DMC GlobalPhone Stream Weights

EN Feature	Weight	GE Feature	Weight
AFFRICATE	0.02061	AFFRICATE	0.00811
APPROXIMANT	0.01613	ALVEOLAR	0.00003
BACK	0.02765	APPROXIMANT	0.00561
BILABIAL	0.03270	ASPIRATED	0.00011
CENTRAL	0.01757	BACK	0.00391
CLOSE	0.00058	BILABIAL	0.00020
CLOSE-MID	0.00879	CLOSE	0.00704
CONSONANT	0.00391	CLOSE-MID	0.00067
DENTAL	0.04785	CONSONANT	0.01118
FLAP	0.02847	DENTAL	0.00407
GLOTTAL	0.05009	FLAP	0.00304
LABIODENTAL	0.01890	FRICATIVE	0.00320
LATERAL-APPROXIMANT	0.01549	FRONT	0.00001
NASAL	0.00191	GLOTTAL	0.01057
OPEN	0.02349	LABIODENTAL	0.02340
OPEN-MID	0.02227	LATERAL-APPROXIMANT	0.00011
PALATAL	0.03478	NASAL	0.00015
PLOSIVE	0.03056	OPEN-MID	0.00445
POSTALVEOLAR	0.06919	PALATAL	0.00139
ROUND	0.02823	PLOSIVE	0.00086
UNVOICED	0.05961	POSTALVEOLAR	0.00233
VELAR	0.03079	RETROFLEX	0.00470
VOICED	0.02356	ROUND	0.01235
VOWEL	0.02314	VELAR	0.00539

MM4 Feature	Weight	MM5 Feature	Weight
AFFRICATE	0.05515	AFFRICATE	0.02780
ALVEOLAR	0.00145	ALVEOLAR	0.00062
APPROXIMANT	0.01678	APPROXIMANT	0.01645
BILABIAL	0.01435	BILABIAL	0.01719
CENTRAL	0.00004	CLOSE	0.00773
CLOSE	0.00812	CLOSE-MID	0.00496
CLOSE-MID	0.00700	DENTAL	0.00007
DENTAL	0.00318	FLAP	0.01933
FLAP	0.03879	FRONT	0.00811
FRONT	0.00737	GLOTTAL	0.03064
GLOTTAL	0.02548	LABIODENTAL	0.04350
LABIODENTAL	0.03969	LATERAL-APPROXIMANT	0.00726
LATERAL-APPROXIMANT	0.00715	OPEN	0.00031
OPEN-MID	0.01898	OPEN-MID	0.00925
PALATAL	0.03780	PALATAL	0.02197
PLOSIVE	0.01157	PLOSIVE	0.00574
POSTALVEOLAR	0.03209	POSTALVEOLAR	0.02024
RETROFLEX	0.00525	ROUND	0.02471
ROUND	0.02358	VELAR	0.01071
VELAR	0.00153		

Table C.9: Feature weights as learned by DMC on English (EN) data.

Feature	Weight	Feature	Weight
AFFRICATE_CH	0.00764	AFFRICATE_SP	0.01316
ALVEOLAR_CH	0.00614	ALVEOLAR_SP	0.01398
APPROXIMANT_CH	0.00491	APPROXIMANT_SP	0.01101
ASPIRATED_CH	0.00655		
BACK_CH	0.00927	BACK_SP	0.01465
BILABIAL_CH	0.00778	BILABIAL_SP	0.01249
CLOSE_CH	0.00794	CLOSE_SP	0.01073
		CLOSE-MID_SP	0.01253
CONSONANT_CH	0.00537	CONSONANT_SP	0.01093
		DENTAL_SP	0.01463
		FLAP_SP	0.01329
FRICATIVE_CH	0.00625	FRICATIVE_SP	0.01267
FRONT_CH	0.00325	FRONT_SP	0.00788
LABIODENTAL_CH	0.00537	LABIODENTAL_SP	0.01273
LATERAL-APPROXIMANT_CH	0.00969	LATERAL-APPROXIMANT_SP	0.01523
NASAL_CH	0.00527	NASAL_SP	0.00649
OPEN_CH	0.01075	OPEN_SP	0.01343
OPEN-MID_CH	0.00655		
PALATAL_CH	0.00577	PALATAL_SP	0.01258
PLOSIVE_CH	0.00451	PLOSIVE_SP	0.01150
		POSTALVEOLAR_SP	0.01284
RETROFLEX_CH	0.00920		
ROUND_CH	0.00560	ROUND_SP	0.01233
UNROUND_CH	0.00442	UNROUND_SP	0.00787
UNVOICED_CH	0.00666	UNVOICED_SP	0.01568
VELAR_CH	0.00751	VELAR_SP	0.01425
VOICED_CH	0.00224	VOICED_SP	0.00958
VOWEL_CH	0.00556	VOWEL_SP	0.01105

Table C.10: Feature selection and weighting as learned by DMC on English when using the feature detectors from all languages.

C.3 MMIE GlobalPhone Stream Weights

Feature	Weight	Feature	Weight
AFFRICATE_CH	0.006076	AFFRICATE_SP	0.012825
ALVEOLAR_CH	0.005061	ALVEOLAR_SP	0.013173
APPROXIMANT_CH	0.006575	APPROXIMANT_SP	0.012937
ASPIRATED_CH	0.007076		
BACK_CH	0.005404	BACK_SP	0.013021
BILABIAL_CH	0.006678	BILABIAL_SP	0.013696
CLOSE_CH	0.008719	CLOSE_SP	0.014257
		CLOSE-MID_SP	0.012735
CONSONANT_CH	0.007173	CONSONANT_SP	0.012575
		DENTAL_SP	0.013653
		FLAP_SP	0.013350
FRICATIVE_CH	0.006383	FRICATIVE_SP	0.013784
FRONT_CH	0.004371	FRONT_SP	0.009640
LABIODENTAL_CH	0.005948	LABIODENTAL_SP	0.013738
LATERAL-APPROXIMANT_CH	0.006127	LATERAL-APPROXIMANT_SP	0.013434
NASAL_CH	0.006905	NASAL_SP	0.008883
OPEN_CH	0.009671	OPEN_SP	0.013744
OPEN-MID_CH	0.006135		
PALATAL_CH	0.005710	PALATAL_SP	0.012878
PLOSIVE_CH	0.005780	PLOSIVE_SP	0.012823
		POSTALVEOLAR_SP	0.012923
RETROFLEX_CH	0.006396		
ROUND_CH	0.005016	ROUND_SP	0.012910
UNROUND_CH	0.005563	UNROUND_SP	0.008286
UNVOICED_CH	0.005670	UNVOICED_SP	0.014720
VELAR_CH	0.006298	VELAR_SP	0.013158
VOICED_CH	0.005004	VOICED_SP	0.010538
VOWEL_CH	0.007204	VOWEL_SP	0.012349

Table C.11: Feature weights as learned by MMIE on English when using CH and SP feature detectors (all languages).

C.4 MMIE ESST Stream Weights

Feature	Weight	Feature	Weight
VOWEL	0.016926	DNT-FR	0.006808
CONSONANT	0.016926	LQGL-BACK	0.006802
LOW-VOW	0.016866	ANTERIOR	0.006784
CARDVOWEL	0.016134	HIGH-CONS	0.006690
SYLLABIC	0.015692	BACK-CONS	0.006616
BACK-VOW	0.014194	REDUCED-CON	0.006576
ROUND-VOW	0.013140	SONORANT	0.006552
ROUND	0.011844	REDUCED	0.006524
CONSONANTAL	0.010746	VEL-PL	0.006450
BILABIAL	0.010330	ROUND-DIP	0.006436
LAX-VOW	0.010242	BF-DIP	0.006216
CONTINUANT	0.010060	TENSE-VOW	0.006128
LAB-PL	0.009762	APPROXIMANT	0.006006
STOP	0.009570	AFFRICATE	0.005970
VCD-PL	0.009354	ALV-PL	0.005796
Y-DIP	0.008552	GLOTTAL	0.005742
LABIAL	0.008416	RETROFLEX	0.005732
PALATAL	0.008348	ALV-FR	0.005580
DIPHTHONG	0.008288	HIGH-VOW	0.005562
NASAL	0.008232	STRIDENT	0.005484
MID-VOW	0.008020	ALVEOPALATAL	0.005406
FRICATIVE	0.007938	LIQUID	0.005220
Y-GLIDE	0.007872	APICAL	0.005214
CENTRAL-VOW	0.007760	LAB-FR	0.005194
MH-DIP	0.007694	LATERAL	0.005038
W-GLIDE	0.007428	LH-DIP	0.004840
LW	0.007418	VLS-PL	0.004692
REDUCED-VOW	0.007412	VLS-FR	0.003932
OBSTRUENT	0.007340	CORONAL	0.002360
PLOSIVE	0.007226	ALVEOLAR-RIDGE	0.002260
W-DIP	0.007146	ALVEOLAR	0.002068
FRONT-VOW	0.007134	UNVOICED	0.002002
VCD-FR	0.006886	VOICED	0.002000
LABIALIZED	0.006832	SIBILANT	0.001212

Table C.12: Feature weights as learned by MMIE on the ESST data: most important questions are for vowel qualities, least important are questions for specific points of articulation and voicing.

Bibliography

[ABK$^+$00] Sebastian Albrecht, Jan Busch, Martin Kloppenburg, Florian Metze, and Paul Tavan. Generalized radial basis function networks for classification and novelty detection: self-organization of optimal bayesian decision. *Neural Networks*, 13:1075–1093, May 2000.

[ADL99] Martine Adda-Decker and Lori Lamel. Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29(2-4):83–98, November 1999.

[AKC94] Andreas Andreou, Terri Kamm, and Jordan Cohen. Experiments in vocal tract normalization. In *Proceedings of the CAIP Workshop*, 1994.

[Aub02] Xavier Aubert. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech and Language*, 16:89–114, 2002.

[BBdSM86] Lalit R. Bahl, Peter F. Brown, Peter V. de Souza, and Robert L. Mercer. Maximum mutual information estimation of Hidden Markov Model parameters for speech recognition. In *Proc. ICASSP*, volume 1, pages 49–52, Tokyo; Japan, May 1986. IEEE.

[BDR96] Hervé Bourlard, Stéphane Dupont, and Christophe Ris. Multi-Stream Speech Recognition. Technical report, Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny; Switzerland, December 1996. IDIAP-RR 96-07.

[Bey00] Peter Beyerlein. *Diskriminative Modellkombination in Spracherkennungssystemen mit großem Wortschatz.* PhD thesis, Rheinisch-Westfälisch-Technische Hochschule Aachen (RWTH), October 2000. In German.

[Bla96] Charles Simon Blackburn. *Articulatory Methods for Speech Production and Recognition.* PhD thesis, Trinity College & CU Engineering Department, December 1996.

[Bro87] Peter F. Brown. *The Acoustic Modeling Problem in Automatic Speech Recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1987.

[BS85] Sheila E. Blumstein and Kenneth N. Stevens. On some issues in the pursuit of acoustic invariance in speech: A reply to Lisker. *JASA*, 77(3):1203–1204, 1985.

[BS04] Susanne Burger and Zachary A. Sloan. The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions. In *Proc. ICASSP-2004 Meeting Recognition Workshop*, Montreal; Canada, May 2004. NIST.

[Cam06] Cambridge Dictionary of American English. http://dictionary.cambridge.org/, 2006.

[Cat77] John C. Catford. *Fundamental problems in phonetics.* Indiana University Press, Bloomington, IN, 1977.

[CH68] Noam Chomsky and Morris Halle. *The Sound Pattern of English.* Harper and Row, New York; USA, 1968.

[Cha02] Shuangyu Chang. *A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech Recognition.* PhD thesis, University of California, Berkeley, 2002.

[Che00] Marilyn Y. Chen. Nasal detection module for a knowledge-based speech recognition system. In *Proc. ICSLP-2000*, volume 4, Beijing; China, 2000. ISCA.

[CWG05] Shuangyu Chang, Mirjam Wester, and Steven Greenberg. An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. *Speech Communication*, 47:290–311, 2005.

[CY95] John Clark and Colin Yallop. *An introduction to phonetics and phonology.* Blackwell Publishers, 1995. 2nd ed.

[Den97] Li Deng. Integrated Multilingual Speech Recognition using Universal Phonological Features in a Functional Speech Production Model. In *Proc. ICASSP 97*, München; Bavaria, 1997. IEEE.

[Den98] Li Deng. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, Vol. 30, 1998.

[DMW94] Paul Duchnowski, Uwe Meier, and Alex Waibel. See me, hear me: Integrating automatic speech recognition and lipreading. In *Proc. Int. Conference on Spoken Language Processing*, pages 547–550, Yokohama, Japan, 1994. IEEE.

[DS94] Li Deng and Don X. Sun. A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech Units Constructed from Overlapping Articulatory Features. *JASA*, 95(5):2702–2719, May 1994.

[Dus01] Sorin Dusan. Methods for integrating phonetic and phonological knowledge in speech inversion. In Vitaly V. Kluev and Nikos E. Mastorakis, editors, *Advances in Signal Processing, Robotics and Communications, Electrical and Computer Engineering Series*, pages 194–199. WSES Press, 2001.

[EB92] Kjell Elenius and Mats Blomberg. Comparing phoneme and feature based speech recognition using artificial neural networks. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 1279–1282, Banff; Canada, 1992. ISCA.

[EB00] Dan Ellis and Jeff Bilmes. Using mutual information to design feature combinations. In *Proc. ICSLP 2000*, Beijing; China, October 2000. ISCA.

[EF96] Kevin Erler and George H. Freeman. An HMM-based speech recognizer using overlapping articulatory features. *JASA*, 100(4):2500–2513, 1996.

[Eid01] Ellen Eide. Distinctive Features For Use in an Automatic Speech Recognition System. In *Proc. EuroSpeech 2001 - Scandinavia*, Aalborg; Denmark, September 2001. ISCA.

[Ell97] Tania Ellbogen. Phonetik Seminar. Internet, http://www.phonetik.uni-muenchen.de/MUSE/Seminare/PHON_Einf/anatomie, 1997.

[ERGM93] Ellen Eide, J. Robin Rohlicek, Herbert Gish, and Sanjoy Mitter. A linguistic feature representation of the speech waveform. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 483–486. IEEE, 1993.

[Esk93] Maxine Eskénazi. Trends in speaking styles research. In *Proc. EuroSpeech 1993*, Berlin; Germany, 1993. ISCA.

[EW94] Carol Y. Espy-Wilson. A feature-based semivowel recognition system. *JASA*, 96(1):65–72, 1994.

[Fan60] Gunnar Fant. *Acoustic Theory of Speech Production.* Mouton & Co., Den Haag; NL, 1960.

[FFKW99] Michael Finke, Jürgen Fritsch, Detlef Koll, and Alex Waibel. Modeling and efficient decoding of large vocabulary conversational speech. In *Proc. Eurospeech 1999*, Budapest; Hungary, September 1999. ISCA.

[FGH$^+$97] Michael Finke, Petra Geutner, Herrmann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The Karlsruhe Verbmobil Speech Recognition Engine. In *Proc. ICASSP 97*, München; Germany, April 1997. IEEE.

[Fis97] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates. In *Proc. ASRU 1997*, Santa Barbara, CA; USA, 1997. IEEE.

[FK05] Joe Frankel and Simon King. A hybrid ANN/DBN approach to articulatory feature recognition. In *Proc. EuroSpeech*, Lisbon, September 2005.

[Fla65] James L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer, New York, 1965.

[FR97] Michael Finke and Ivica Rogina. Wide context acoustic modeling in read vs. spontaneous speech. In *Proc. 1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, München; Bavaria, 1997. IEEE.

[Fri00] Jürgen Fritsch. *Hierarchical Connectionist Acoustic Modeling for Domain-Adaptive Large Vocabulary Speech recognition.* PhD thesis, Universität Karlsruhe (TH), 2000.

[FT87] Uli H. Frauenfelder and Lorraine K. Tyler. The process of spoken word recognition: An introduction. *Cognition*, 25(1):1–20, 1987.

[FW97a] Michael Finke and Alex Waibel. Flexible Transcription Alignment. In *Proc. ASRU 1997*, Santa Barbara, CA; USA, December 1997. IEEE.

[FW97b] Michael Finke and Alexander Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proc. Eurospeech 1997*, Rhodes; Greece, 1997. ISCA.

[GA03] Stephen D. Goldinger and Tamiko Azuma. Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, 31:305–320, 2003.

[Gal97] Mark J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical report, Cambridge University, Cambridge; UK, May 1997. CUED/F-INFENG/TR 291.

[Gal99] Mark J. F. Gales. Semi-Tied Covariance Matrices for Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, May 1999.

[Gal02] Mark J.F. Gales. Transformation streams and the HMM error model. *Computer Speech and Language*, 16:225–243, 2002.

[GAPN02] Guillaume Gravier, Scott Axelrod, Gerasimos Potamianos, and Chalapathy Neti. Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR. In *Proc. ICASSP 2002*, Orlando, FL; USA, April 2002. IEEE.

[GEWB$^+$99] Frank H. Guenther, Carol Y. Espy-Wilson, Suzanne E. Boyce, Melanie L. Matthies, Majid Zandipour, and Joseph S. Perkell. Articulatory tradeoffs reduce acoustic variability during american english /r/ production. *JASA*, 105:2854–2865, 1999.

[GHE96] Steven Greenberg, Joy Hollenback, and Dan Ellis. Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *Proc. ICSLP 1996*. ISCA, 1996.

[GHM92] John J. Godfrey, Edward Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, San Francisco, CA, 1992.

[GKNN91] Ponani S. Gopalakrishnan, Dimitry Kanevsky, Arthur Nádas, and David Nahamoo. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory*, 37:107–113, 1991.

[GL94] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, April 1994.

[Gla03] James R. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech & Language*, 17:137–152, 2003.

[GLF04] John S. Garofolo, Christophe D. Laprun, and Jonathan G. Fiscus. The rich transcription 2004 spring meeting recognition evaluation. In *Proc. ICASSP 2004 Meeting Recognition Workshop*, Montreal; Canada, May 2004. NIST. http://www.nist.gov/speech/test_beds/mr_proj/documents/icassp/papers/P01.pdf.

[GOK03] Julie Goldberg, Mari Ostendorf, and Katrin Kirchhoff. The impact of response wording in error correction subdialogs. In *Proc. ISCA Workshop on Error Handling in Spoken Dialog Systems*. ISCA, September 2003.

[Gop98] Ramesh Gopinath. Maximum likelihood modeling with Gaussian distributions for classification. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[GPN02] Guillaume Gravier, Gerasimos Potamianos, and Chalapathy Neti. Asynchrony modeling for audio-visual speech recognition. In *Proc. Human Language Technology Conf. (HLT)*. ACL, March 2002.

[GSBB04] John N. Gowdy, Amarnag Subramanya, Chris Bartels, and Jeff Bilmes. DBN based multi-stream models for audio-visual speech recognition. In *Proc. ICASSP 2004*, Montreal; Canada, May 2004. IEEE.

[Her97] Javier Hernando. Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.

[Hes03] Wolfgang Hess. Grundlagen der Phonetik. http://www.ikp.uni-bonn.de/dt/lehre/materialien/grundl_phon/gph_3f.pdf, 2003. Institut für Kommunikationsforschung und Phonetik (IKP).

[HHB89] Mark A. Huckvale, Ian S. Howard, and William J. Barry. Automatic phonetic labeling of continuous speech. In *Proc. EuroSpeech 1989*, Paris; France, 1989. ISCA.

[Hie93] James L. Hieronymus. ASCII Phonetic Symbols for the World's Languages: Worldbet. *Journal of the International Phonetics Association*, 23, 1993.

[HJa05] Mark Hasegawa-Johnson and al. Landmark-based speech recognition: Report of the 2004 Johns-Hopkins summer workshop. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, May 2005. IEEE.

[Int99] International Phonetic Association. *Handbook of the International Phonetic Association.* Cambridge University Press, 1999.

[JAB$^+$04] Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. The ICSI Meeting Project: Resources and Research. In *Proc. ICASSP-2004 Meeting Recognition Workshop*, Montreal; Canada, May 2004. NIST.

[JCL95] Biing-Hwang Juang, Wu Chou, and Chin-Hui Lee. *Statistical and Discriminative Methods for Speech Recognition and Coding – New Advances and Trends.* Springer Verlag, Berlin-Heidelberg, 1995.

[Jel69] Frederick Jelinek. Fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development*, 13(6), 1969.

[Jel98] Frederick Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Boston, 1998.

[JEM99] Adam Janin, Dan Ellis, and Nelson Morgan. Multi-stream speech recognition: Ready for prime time. In *Proc. EuroSpeech 1999*, Budapest; Hungary, September 1999. ISCA.

[JEW03] Amit Juneja and Carol Y. Espy-Wilson. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 675–679. IEEE, 2003.

[JFH52] Roman Jakobson, Gunnar Fant, and Morris Halle. Preliminaries to speech analysis. Technical Report 13, MIT Acoustics Lab, Cambridge, MA; USA, 1952.

[JMSK98] Hubert Jin, Spyros Matsoukas, Rich Schwartz, and Francis Kubala. Fast Robust Inverse Transform SAT and Multi-stage Adaptation. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA; USA, 1998.

[JR05] Biing-Hwang Juang and Lawrence R. Rabiner. Automatic speech recognition–a brief history of the technology. In Keith

Brown, editor, *Elsevier Encyclopedia of Language and Linguistics*. Elsevier, Oxford, 2nd edition, 2005.

[JSW05] Szu-Chen Jou, Tanja Schultz, and Alex Waibel. Whispery speech recognition using adapted articulatory features. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA; USA, 2005. IEEE.

[Jun04] Amit Juneja. *Speech Recognition based on Phonetic Features and Acoustic Landmarks*. PhD thesis, University of Maryland, College Park; MD, 2004.

[KFS00] Katrin Kirchhoff, Gernot A. Fink, and Gerhard Sagerer. Conversational Speech Recognition using Acoustic and Articulatory Input. In *Proc. ICASSP 2000*, Istanbul; Turkey, June 2000. IEEE.

[Kir98] Katrin Kirchhoff. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proc. ICSLP 98*, Sydney, NSW; Australia, December 1998. IEEE.

[Kir99] Katrin Kirchhoff. *Robust Speech Recognition Using Articulatory Information*. PhD thesis, Technische Fakultät der Universität Bielefeld, Bielefeld; Germany, June 1999.

[Kir00] Katrin Kirchhoff. Integrating Articulatory Features into Acoustic Models for Speech Recognition. In *Proceedings of the Workshop on Phonetics and Phonology in ASR: Parameters and Features, and their Implications (Phonus 5)*, Saarbrücken, Germany, March 2000. Institute of Phonetics, Universität des Saarlandes.

[Kit98] Josef Kittler. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[KS97] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *Proc. EuroSpeech 97*, Rhodes; Greece, 1997.

[KT00] Simon King and Paul Taylor. Detection of phonological features in continuous speech using neural networks. *Computer Speech & Language*, 14(4):333–353, 2000.

[KTFR00] Simon King, Paul Taylor, Joe Frankel, and Korin Richmond. Speech recognition via phonetically-featured syllables. In

Proc. Workshop on Phonetics and Phonology in ASR "Phonus 5", Saarbrücken; Germany, 2000. Institute of Phonetics.

[Lad82] Peter Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, 2nd edition, 1982.

[Lav94] John Laver. *Principles of Phonetics*. Cambridge University Press, Cambridge; UK, May 1994.

[LDGP94] Steve Lowe, Anne Demedts, Larry Gillick, and Mark Mandeland Barbara Peskin. Language identification via large vocabulary speaker independent continuous speech recognition. In *Proceedings of the workshop on Human Language Technology*, pages 437–441, Plainsboro, NJ; USA, 1994. ACL.

[Lee04] Chin-Hui Lee. From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition. In *Proc. ICSLP 2004*. ISCA, October 2004.

[Leg95] Chris J. Leggetter. *Improving Acoustic Modelling for HMMs using Linear Transforms*. PhD thesis, Cambridge University, England, 1995.

[Lem99] Sami Lemmetty. Review of speech synthesis technology. Master's thesis, Helsinki University of Technology, Finland, http://www.acoustics.hut.fi, 1999.

[LG04] Karen Livescu and James Glass. Feature-based pronunciation modeling with trainable asynchrony probabilities. In *Proc. Interspeech ICSLP-2004*, Jeju Island; Korea, October 2004. ISCA.

[LG05] Lori Lamel and Jean-Luc Gauvain. Alternate phone models for conversational speech. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, March 2005. IEEE.

[LGB03] Karen Livescu, James Glass, and Jeff Bilmes. Hidden feature models for speech recognition using dynamic Bayesian networks. In *Proc. EuroSpeech 2003*, Geneva; Switzerland, September 2003. ISCA.

[Li05] Xiang Li. *Combination and Generation of Parallel Feature Streams for Improved Speech Recognition*. PhD thesis, ECE Department; Carnegie Mellon University, Pittsburgh, PA; USA, February 2005.

[Liu96] Sharlene A. Liu. Landmark detection for distinctive feature-based speech recognition. *JASA*, 100(5):3417–3430, 1996.

[LJS04] Kornel Laskowski, Qin Jin, and Tanja Schultz. Cross-correlation–based Multispeaker Speech Activity Detection. In *Proc. Interspeech ICSLP-2004*, Jeju; Korea, October 2004. ISCA.

[LL05] Jinyu Li and Chin-Hui Lee. On designing and evaluating speech event detectors. In *Proc. Interspeech 2005*. ISCA, October 2005.

[LMSK05] Ka-Yee Leung, Man-Wai Mak, Manhung Siu, and Sun-Yuan Kung. Speaker verification using adapted articulatory feature-based conditional pronunciation modeling. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, 2005. IEEE.

[LS03] Ka-Yee Leung and Manhung Siu. Phone-level confidence measure using articulatory features. In *Proc. ICASSP 2003*, Hong Kong; China, April 2003. IEEE.

[LTL05] Jinyu Li, Yu Tsao, and Chin-Hui Lee. A study on knowledge source integration for candidate rescoring in automatic speech recognition. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, March 2005. IEEE.

[LW94] Chris J. Leggetter and Phil C. Woodland. Speaker adaptation of HMMs using linear regression. Technical report, Cambridge University, England, 1994.

[Mac98] Wolfgang Macherey. Implementierung und Vergleich diskriminativer Verfahren für Spracherkennung bei kleinem Vokabular. Master's thesis, Lehrstuhl für Informatik VI der RWTH Aachen, 1998. In German.

[MBS00] Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks. *Computer, Speech and Language*, 14(4):373–400, 2000.

[MC92] Yeshwant Muthusamy and Ronald A. Cole. Automatic segmentation and identification of ten languages using telephone speech. In *Proc. ICSLP 1992*, 1992.

[MFP$^+$04] Florian Metze, Christian Fügen, Yue Pan, Tanja Schultz, and Hua Yu. The ISL RT-04S Meeting Transcription System. In *Proceedings NIST RT-04S Evaluation Workshop*. NIST, May 2004.

[MFPW05] Florian Metze, Christian Fügen, Yue Pan, and Alex Waibel. Automatically Transcribing Meetings Using Distant Microphones. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, March 2005. IEEE.

[MHB01] Andrew Morris, Astrid Hagen, and Hervé Bourlard. MAP combination of multi-stream HMM or HMM/ANN experts. In *Proc. EuroSpeech 2001*, Aalborg; Denmark, September 2001. ISCA.

[MHMSW05] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. Session independent non-audible speech recognition using surface electromyography. In *Proc. ASRU 2005*, Cancun; Mexico, November 2005. IEEE.

[MJF+04] Florian Metze, Qin Jin, Christian Fügen, Kornel Laskowski, Yue Pan, and Tanja Schultz. Issues in Meeting Transcription – The ISL Meeting Transcription System. In *Proc. INTERSPEECH2004-ICSLP*. ISCA, October 2004.

[MKS+00] Florian Metze, Thomas Kemp, Thomas Schaaf, Tanja Schultz, and Hagen Soltau. Confidence measure based Language Identification. In *Proc. ICASSP 2001*, Istanbul, April 2000. IEEE.

[MM99] Nikki Mirghafori and Nelson Morgan. Sooner or later: Exploring asynchrony in multi-band speech recognition. In *Proc. EuroSpeech 1999*, Budapest; Hungary, 1999. ISCA.

[MND04] Konstantin Markov, Satoshi Nakamura, and Jianwu Dang. Integration of articulatory dynamic parameters in HMM/BN based speech recognition system. In *Proc. Interspeech ICSLP-2004*, Jeju Island; Korea, October 2004. ISCA.

[Moo03] Roger Moore. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proc. Eurospeech 2003*, pages 2582–2584, Geneva; Switzerland, 9 2003. ISCA.

[MSW+04] Nikki Mirghafori, Andreas Stolcke, Chuck Wooters, Tuomo Pirinen, Ivan Bulyko, David Gelbart, Martin Graciarena, Scott Otterson, Barbara Peskin, and Mari Ostendorf. From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System. In *Proc. INTERSPEECH2004 – ICSLP*, Jeju Island; Korea, October 2004. ISCA.

[MVBT95] Kevin G. Munhall, Eric Vatikiotis-Bateson, and Yoh'ichi Tohkura. X-ray film database for speech research. *JASA*, 98(2):1222–1224, 1995.

[MW02] Florian Metze and Alex Waibel. A Flexible Stream Architecture for ASR using Articulatory Features. In *Proc. ICSLP 2002*, Denver, CO; USA, September 2002. ISCA.

[MW03] Florian Metze and Alex Waibel. Using Articulatory Features for Speaker Adaptation. In *Proc. ASRU 2003*, St. Thomas, US VI, 2003. IEEE.

[MWW94] William Marslen-Wilson and Paul Warren. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, 101(4):653–675, 1994.

[NBR99] Patha Niyogi, Chris Burges, and Padma Ramesh. Distinctive feature detection using support vector machines. In *Proc. ICASSP 1999*, Phoenix, AZ; USA, 1999. IEEE.

[NIS04a] NIST. Proceedings NIST ICASSP 2004 Meeting Recognition Workshop. http://www.nist.gov/speech/test_beds/mr_proj/icassp_program.html, May 2004.

[NIS04b] NIST. Rich Transcription 2004 Spring Meeting Recognition Evaluation. http://www.nist.gov/speech/tests/rt/rt2004/spring/, May 2004.

[Nor86] Yves Normandin. Maximum mutual information estimation of hidden Markov models. In Chin-Hui Lee, Frank K. Song, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, pages 57–81. Kluwer Academic Publishers, Norwell, MA, 1986.

[NY02] Harriet Nock and Steve Young. Modelling asynchrony in automatic speech recognition using loosely coupled hidden Markov models. *Cognitive Science*, 26:283–301, 2002.

[Ost99] Mari Ostendorf. Moving Beyond the 'Beads-on-a-String' Model of Speech. In *Proc. ASRU 1999*, Keystone, CO; USA, December 1999. IEEE.

[Ovi98] Sharon Oviatt. The CHAM model of hyperarticulate adaptation during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[PBBB88] Eric D. Petajan, Bradford Bischoff, David Bodoff, and N. Michael Brooke. An improved automatic lipreading system to enhance speech recognition. In *Proc. CHI 1988*, pages 19–25, 1988.

[PDB85] Michael A. Picheny, Nathaniel I. Durlach, and Louis D. Braida. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of speech and hearing research*, 28:96–103, 1985.

[PDB86] Michael A. Picheny, Nathaniel I. Durlach, and Louis D. Braida. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of speech and hearing research*, 29:434–446, 1986.

[Per69] Joseph S. Perkell. *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study.* MIT Press, 1969.

[Per97] Joseph S. Perkell. Articulatory processes. In William J. Hardcastle and John Laver, editors, *The Handbook of Phonetics.* Blackwell Publishing, 1997.

[PG98] Gerasimos Potamianos and Hans-Peter Graf. Discriminative training of HMM stream exponents for audio-visual speech recognition. In *Proc. ICASSP 1998*, Seattle, WA; USA, 1998. IEEE.

[PGKW03] Dan Povey, Mark J.F. Gales, Do Y. Kim, and Phil C. Woodland. MMI-MAP and MPE-MAP for acoustic model adaptation. In *Proc. Eurospeech 2003*, Geneva; Switzerland, September 2003. ISCA.

[PHT^{+}92] George Papcun, Judith Hochberg, Timothy R. Thomas, François Laroche Jeff Zacks, and Simon Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data. *JASA*, 92(2):688–700, 1992.

[PK03] Sonia Parandekar and Katrin Kirchhoff. Multi-stream language identification using data-driven dependency selection. In *Proc. ICASSP 2003*, Hong Kong, 2003. IEEE.

[PNIH01] Guillaume Potamianos, Chalapathy Neti, Giri Iyengar, and Eric Helmuth. Large-vocabulary audio-visual speech recognition by machines and humans. In *Proc. EuroSpeech*, Aalborg, Denmark, September 2001. ISCA.

[Pov05] Daniel Povey. *Discriminative Training for Large Vocabulary Speech Recognition.* PhD thesis, Peterhouse College & CU Engineering Departement, 2005.

[PUB94] Karen L. Payton, Rosalie M. Uchanski, and Louis D. Braida. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.*, 1994.

[RA97] Sam Roweis and Abeer Alwan. Towards articulatory speech recognition: Learning smooth maps to recover articulator information. In *Proc. of EuroSpeech 1997*, volume 3, pages 1227–1230, Rhodes; Greece, 1997. ISCA.

[RBD03] Matthew Richardson, Jeff Bilmes, and Chris Diorio. Hidden-articulator Markov models for speech recognition. *Speech Communication*, 41(2):511–529, 2003.

[Ree00] Henning Reetz. Underspecified phonological features for lexical access. In *Proc. Workshop on Phonetics and Phonology in ASR "Phonus 5"*, Saarbrücken; Germany, 2000. Institute of Phonetics.

[RKT03] Korin Richmond, Simon King, and Paul Taylor. Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech & Language*, 2003.

[Rog97] Ivica Rogina. *Parameterraumoptimierung für Diktiersysteme mit unbeschränktem Vokabular.* PhD thesis, Fakultät für Informatik der Universität Karlsruhe (TH), Karlsruhe, Germany, 1997.

[Rog05] Ivica Rogina. *Sprachliche Mensch-Maschine-Kommunikation.* Universität Karlsruhe (TH), 2005.

[RW94] Ivica Rogina and Alex Waibel. Learning state-dependent stream weights for multi-codebook HMM speech recognition systems. In *Proc. ICASSP 94*, Adelaide; Australia, 1994. IEEE.

[SAB$^+$05] Andreas Stolcke, Xavier Anguera, Kofi Boakye, Özgür Çetin, Frantisek Grezl, Adam Janin, Arindam Mandal, Barbara Peskin, Chuck Wooters, and Jing Zeng. Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system. In *Proc. 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2005)*, Edinburgh; Scotland, July 2005. Springer.

[SB81] Kenneth N. Stevens and Sheila E. Blumstein. The search for invariant acoustic correlates of phonetic features. In Peter D. Eimas and Joanne L. Miller, editors, *Perspectives on the study of speech*, pages 1–38. Hillsdale, NJ, 1981.

[Sch89] Otto Schmidbauer. Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations. In *1989 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 616–619, Glasgow, Scotland; UK, 1989. IEEE.

[SG04a] Vincent Stanford and John Garofolo. Beyond Close-talk – Issues in Distant speech Acquistion, Conditioning Classification, and Recognition. In *Proc. ICASSP-2004 Meeting Recognition Workshop*, Montreal; Canada, May 2004. NIST.

[SG04b] Stephanie Strassel and Meghan Glenn. Shared Linguistic Resources for Human Language Technology in the Meeting Domain. In *Proc. ICASSP-2004 Meeting Recognition Workshop*, Montreal; Canada, May 2004. NIST.

[SISHB04] Hemant Misra Shajith Ikbal, Sunil Sivadas, Hynek Hermansky, and Hervé Bourlard. Entropy based combination of tandem representations for noise robust ASR. In *Proc. of the INTERSPEECH-ICSLP-04*, Jeju Island; Korea, October 2004. ICSLP.

[SK97] Thomas Schaaf and Thomas Kemp. Confidence measures for spontaneous speech. In *Proc. ICASSP 97*, München; Bavaria, April 1997. IEEE.

[SK00] Murat Saraçlar and Sanjeev Khudanpur. Properties of pronunciation change in conversational speech recognition. In *Proc. 2000 Speech Transcription Workshop*, University of Maryland, May 2000. NIST.

[SKM$^+$05] Hagen Soltau, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon, and Geoffrey Zweig. The IBM 2004 Conversational Telephony System for Rich Transcription. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA; USA, March 2005. IEEE.

[SMFW02] Hagen Soltau, Florian Metze, Christian Fügen, and Alex Waibel. Efficient Language Model Lookahead through Polymorphic Linguistic Context Assignment. In *Proc. ICASSP 2002*, Orlando, FL; USA, 2002. IEEE.

[SMMN01] Ralf Schlüter, Wolfgang Macherey, Boris Müller, and Hermann Ney. Comparison of discriminative training criteria and optimization methods for speech recognition. *Speech Communication*, 34, 2001.

[SMSHL92] Kenneth N. Stevens, Sharon Y. Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu. Implementation of a model for lexical access based on features. In *Proc. ICSLP 1992*, pages 499–503, Edmonton; Canada, 1992. ISCA.

[SMSW03] Sebastian Stüker, Florian Metze, Tanja Schultz, and Alex Waibel. Integrating Multilingual Articulatory Features into Speech Recognition. In *Proc. EuroSpeech 2003*, Geneva; Switzerland, 2003. ISCA.

[SMW02] Hagen Soltau, Florian Metze, and Alex Waibel. Compensating for Hyperarticulation by Modeling Articulatory Properties. In *Proc. ICSLP 2002*. ISCA, September 2002.

[SO72] Joanne D. Subtelny and N. Oya. Cineradiographic study of sibilants. *Folia Phoniatrica*, 24(1):30–50, 1972. Basel; CH.

[Sol05] Hagen Soltau. *Compensating Hyperarticulation for Automatic Speech Recognition*. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, February 2005.

[SPSF00] Kemal Sönmez, Madelaine Plauché, Elizabeth Shriberg, and Horacio Franco. Consonant discrimination in elicited and spontaneous speech: A case for signal-adaptive front ends in asr. In *Proc. 2000 Speech Transcription Workshop*, University of Maryland, May 2000. NIST.

[SSMW03] Sebastian Stüker, Tanja Schultz, Florian Metze, and Alex Waibel. Multilingual Articulatory Features. In *Proc. ICASSP 2003*. IEEE, April 2003.

[Ste98] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, 1998.

[Ste02] Kenneth N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *JASA*, 111(4), April 2002.

[Stü03] Sebastian Stüker. Multilingual articulatory features. Master's thesis, Universität Karlsruhe (TH), Fakultät für Informatik, April 2003.

[SW97] Tanja Schultz and Alex Waibel. Fast Bootstrapping of LVSCR Systems with Multilingual Phoneme Sets. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, volume 1, pages 371–374, Rhodes, Greece, September 1997.

[SW98] Hagen Soltau and Alex Waibel. On the influence of hyperarticulated speech on the recognition performance. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[SW01] Tanja Schultz and Alex Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. *Speech Communication*, 35(1-2):31–51, August 2001.

[SWM+04] Andreas Stolcke, Chuck Wooters, Nikki Mirghafori, Tuomo Pirinen, Ivan Bulyko, Dave Gelbart, Martin Graciarena, Scott Otterson, Barbara Peskin, and Mari Ostendorf. Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system. In *Proc. NIST 2004 Spring Evaluation Workshop*, Montreal; Canada, 2004. National Institute of Standards and Technology.

[SWW97] Tanja Schultz, Martin Westphal, and Alex Waibel. The GlobalPhone Project: Multilingual LVCSR with JANUS-3. In *Proceedings of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, Pilzen, Czech Republic, 1997.

[SYM+04] Hagen Soltau, Hua Yu, Florian Metze, Christian Fügen, Qin Jin, and Szu-Chen Jou. The 2003 ISL Rich Transcription System for Conversational Telephony Speech. In *Proc. ICASSP 2004*, Montreal; Canada, 2004. IEEE.

[SZH+03] Georg Stemmer, Viktor Zeissler, Christian Hacker, Elmar Nöth, and Heinrich Niemann. A Phone Recognizer Helps to Recognize Words Better. In *Proc. ICASSP 2003*, volume 1, pages 736–739, Hong Kong, 2003.

[TSZ03] Min Tang, Stephanie Seneff, and Victor Zue. Two-stage continuous speech recognition using feature-based models: A preliminary study. In *Proc. ASRU 2003*, St. Thomas; USVI, 2003. IEEE.

[UCB+96] Rosalie M. Uchanski, Sunkyung S. Choi, Louis D. Braida, Charlotte M. Reed, and Nathaniel I. Durlach. Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of speech and hearing research*, 39:494–509, 1996.

[UNGH98] Naonori Ueda, Ryohei Nakano, Zoubin Gharamani, and Geoffrey E. Hinton. Split and merge EM algorithm for improving Gaussian mixture density esimates. *Neural Networks for Signal Processing*, pages 274–283, 1998.

[Ver00] Dimitra Vergyri. *Integration of Multiple Knowledge Sources in Speech Recognition using Minimum Error Training*. PhD thesis, Johns Hopkins University, Baltimore, MA; USA, 2000.

[Wel89] John C. Wells. Computer-Coded Phonemic Notation of Individual Languages of the European Community. *Journal of the International Phonetic Association*, 19:32–54, 1989.

[Wes03] Mirjam Wester. Pronunciation modeling for ASR - knowledge-based and data-derived methods. *Computer Speech & Language*, 17:69–85, 2003.

[WFK04] Mirjam Wester, Joe Frankel, and Simon King. Asynchronous articulatory feature recognition using dynamic Bayesian networks. In *Proc. IEICI Beyond HMM Workshop*, December 2004.

[Wie00] Richard Wiese. *The Phonology of German*. Oxford University Press, 2000.

[Wik05] Wikipedia. Swiss German. http://en.wikipedia.org/wiki/Swiss_German, October 2005.

[WL90] Alex Waibel and Kai-Fu Lee, editors. *Readings in speech recognition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[WP02] Phil Woodland and Dan Povey. Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. *Computer Speech and Language*, Vol. 6, 2002.

[WR00] Alan Wrench and Korin Richmond. Continuous speech recognition using articulatory data. In *Proc. ICSLP 2000*, Beijing; China, October 2000. ISCA.

[Wre00] Alan Wrench. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. In *Phonus Research Report*, volume 4, 2000. ftp://bell.qmuc.ac.uk/mocha/.

[WSS+00] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze. Multilingual Speech Recognition. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Heidelberg; Germany, 2000. Springer-Verlag.

[WTHSS96] Mitch Weintraub, Kelsey Taussig, Kate Hunicke-Smith, and Amy Snodgrass. Effect of speaking style on LVCSR performance. In *Proc. ICSLP 1996*, Philadelphia, PA; USA, October 1996. ISCA.

[Yu04] Hua Yu. *Recognizing Sloppy Speech.* PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA; USA, 2004.

[ZL86] Victor Zue and Lori Lamel. An expert spectrogram reader: A knowledge-based approach to speech recognition. In *Proc. ICASSP 1986*, Tokyo; Japan, April 1986. IEEE.

[ZLR+95] Frederick Zussa, Qiguang Lin, Gael Richard, Daniel Sinder, and James L. Flanagan. Open-loop acoustic-to-articulatory mapping. *JASA*, 98(5), November 1995.

[ZR98] Geoffrey Zweig and Stuart Russell. Speech recognition with dynamic bayesian networks. In *Proc. Fifteenth National Conference on Artificial Intelligence (AAAI-98).* AAAI, July 1998.

[ZSN05] András Zolnay, Ralf Schlüter, and Hermann Ney. Acoustic feature combination for robust speech recognition. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, March 2005. IEEE.

[ZW97] Puming Zhan and Martin Westphal. Speaker normalization based on frequency warping. In *Proc. ICASSP 1997*, München; Bavaria, April 1997. IEEE.