# Compensating Hyperarticulation for Automatic Speech Recognition

Zur Erlangung des akademischen Grades eines

**Doktors der Ingenieurwissenschaften**

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

**Dissertation**

von
**Hagen Soltau**
aus Nordhausen am Harz

| | |
|---|---|
| Tag der mündlichen Prüfung: | 14.2.2005 |
| Erster Gutachter: | Prof. Dr. A. Waibel |
| Zweiter Gutacher: | Prof. Dr. J. Hirschberg |
| Dritter Gutacher: | Prof. Dr. R. Dillmann |

# Abstract

This thesis details the effects of hyperarticulation in the context of automatic speech recognition used for human-to-machine interaction. Hyperarticulation can be characterised as a speaking mode exhibiting an exaggerated articulation and occurs as a natural reaction in an effort to resolve recognition errors. Despite the user's attempt to disambiguate word confusions, hyperarticulation causes a significant increase in recognition errors. Current state-of-the-art technologies in automatic speech recognition fail to deal with hyperarticulated speech.

The effect of hyperarticulated speech on the recognition performance was investigated. Changes in pitch, formant frequencies, or phone duration lead to a mismatch between the train and test environment. The effects occur on a sub-phonetic, articulatory domain. The estimation of model parameters with hyperarticulated training data reduces the speaking style mismatch, but even then hyperarticulation still degrades the recognition performance drastically. This result can be attributed to wrong model assumptions in the framework of phoneme based Hidden Markov Models.

The contribution of this thesis is to show how articulatory properties can be used for recognition of hyperarticulated speech. The articulatory vector space is an algebraic representation of speech events. It provides a fine granularity for modelling of articulatory variations due to different speaking modes. This algebraic representation of speech events allows to describe hyperarticulated effects on a sub-phonetic, articulatory domain. Hyperarticulated variations can be explained using the concept of contrastive attributes. Contrastive attributes are attributes to disambiguate word confusions. Effects of hyperarticulation can be described as the activation or deactivation of contrastive attributes. The mathematical framework, developed in this thesis, provides a set of operations and basis elements to work with contrastive attributes. Hyperarticulation can be seen as warping of trajectories in an articulatory vector space. The vector model consists of probability density functions for each dimension. An exponential combination of the underlying function leads to a score function for the speech events.

The effects of hyperarticulation were studied on two languages: English and German. On both languages, similar performance degradations were observed in a hyperarticulated speaking mode. The influence of hyperarticulation on pitch, formants, and phone duration leads to similar changes both in English and German. Recognition experiments show drastic im-

provements with the vector models over pure phoneme based models. This confirms that hyperarticulation occurs on a sub-phonetic level in an articulatory domain, where standard phoneme based models are not able to capture these variations. Furthermore, a combination of normal with corresponding hyperarticulated utterances achieves a significant improvement over the recognition performance of normal speech. Thus, hyperarticulated data can be used as additional knowledge to improve the recognition of normal speech.

A further evaluation of the generalisation capability of articulatory vector spaces was conducted on the SUSAS (speech under actual and simulated stress) corpus. Significant error reductions were obtained on this type of data. The results confirm the potential of articulatory properties for modelling of speech.

# Zusammenfassung

Diese Dissertation behandelt hyperartikulierte Effekte in Kontext automatischer Spracherkennung für Mensch-Maschine Interaktion. Hyperartikulation kann charakterisiert werden durch eine übertrieben klare Artikulation und tritt auf als eine natürliche Reaktion um Erkennungsfehler zu beheben. Wir zeigen, daß trotz der eigentlich Intention des Benutzers, Wortverwechselungen aufzulösen, dieser Sprechmodus zu einer signifikanten Fehlererhöhung führt. Derzeitige Forschungsysteme im Bereich automatischer Spracherkennung sind nicht in der Lage auf Hyperartikulation angemessen zu reagieren.

Der Effekt hyperartikulierter Sprache auf die Erkennungsleistung wurde untersucht. Veränderungen der Tonhöhe, der Formanten, und der Phonemdauer führen zu einer Diskrepanz zwischen Testdaten und trainierten Modellparametern. Die Veränderungen treten auf einer sub-phonetischen Ebene in einer artikulatorischen Domäne auf. Die Schätzung von Modellparametern mittels hyperartikulierter Trainingsdaten reduziert die Unterschiede zwischen den Modellen und der Testdaten. Gleichwohl besteht ein deutlicher Erkennungseinbruch bei hyperartikulierter Sprache selbst bei Verwendung hyperartikulierter Trainingsdaten. Diese Ergebnisse können fehlerhaften Annahmen bei phonembasierten Hidden Markov Modellen zugeschrieben werden.

Der Beitrag der Dissertation ist es zu zeigen, wie artikulatorische Attribute zur Verbesserung bei der Erkennung hyperartikulierter Sprache eingesetzt werden können. Artikulatorische Vektorräume können hierbei als algebraische Repräsentation von Sprachereignissen verwendet werden. Dies erlaubt eine feinere Auflösung der akustischen Eigenschaften im Vergleich zu Phonemen. Die algebraische Repräsentation von Sprachereignissen erlaubt es, hyperartikulierte Effekte auf einer sub-phonetischen Ebene in einer artikulatorischen Domäne zu beschreiben. Hyperartikulierte Veränderungen können mittels kontrastiver Attribute erklärt werden. Kontrastive Attribute sind Attribute zur Disambiguierung von Wortverwechselungen. Effekte von Hyperartikulation können dabei als Aktivierung und Deaktivierung kontrastiver Attribute beschrieben werden. Das, in dieser Dissertation entwickelte, mathematische Grundgerüst stellt hierbei Operatoren und Basiselemente zur Manipulierung kontrastiver Attribute bereit. Dabei kann Hyperartikulation als eine Verzerrung von Trajektorien im artikulatorischen Vektorraum angesehen werden. Das Vektormodell besteht aus Wahrscheinlichkeitsdichte-Funktionen für jede Dimension. Eine exponentielle Kombination der zugrundelegenden Funktionen erlaubt es, eine Bewertungsfunktion für Sprachereignisse zu de-

finieren.

Der Einfluß von Hyperartikulation ist in zwei Sprachen untersucht worden: Deutsch und Englisch. Dabei ergaben sich für beide Sprachen ähnliche Analyseergebnisse hinsichtlich Tonhöhe, Formanten, sowie Phonemdauer. Erkennungsexperimente zeigen signifikante Verbesserungen mit Vektormodellen gegenüber herkömmlichen phonembasierten Ansätzen. Dies bestätigt, daß Hyperartikulation auf einer sub-phonetischen Ebene in einer artikulatorischen Domäne auftritt, in der traditionelle phonembasierte Modelle nicht in der Lage sind, solchen Variationen gerecht zu werden. Weiterhin konnte gezeigt werden, daß eine Kombination von normalen mit korrespondierenden hyperartikulierten Äußerungen zu einer deutlichen Erkennungsverbesserung bei normaler Sprache führt. Dies bedeutet, daß hyperartikulierte Äußerungen eine zusätzliche Informationsquelle für die Erkennung normaler Sprache darstellen.

Desweiteren ist eine Evaluation des artikulatorischen Vektorraums auf einem Korpus mit unterschiedlichen Sprechweisen durchgeführt worden. Experimente auf dem SUSAS (Sprache bei realem und simuliertem Streß) Korpus belegten, daß signifikante Verbesserungen durch artikulatorische Attribute möglich sind. Die Ergebnisse bestätigen das Potential artikulatorischer Vektorräume.

# Acknowledgement

I want to thank Prof. Waibel for advising my dissertation. I appreciate his efforts for providing funding for our research group. Also, I would like to express my thanks to Prof. Julia Hirschberg and Prof. Dillmann for kindly taking over the task of being the co-advisor of my thesis.

A really great experience was writing the IBIS decoder, a very fast and memory efficient one-pass search engine. At this point of time, I would like to express my thanks to Florian Metze and Christian Fügen for this team work.

As I joined the Interactive Systems Labs in 1997, my first task was building the speech recogniser for the Verbmobil-II evaluations. I would like to thank Thomas Schaaf for the teamwork during a series of evaluations.

As I thought I come closer to my PhD defense, the SWB evaluation did arise in 2003. It was a great experience working with such a great team. Again, Florian and Christian were involved in Karlsruhe. On the other side of the Atlantic Ocean, Hua Yu and Qin Jin were part of our team.

After the SWB evaluation, I got finally the chance to write my dissertation. I would like to thank Vicky, Tanja, and Florian for proofreading my thesis and their valuable comments.

Special thanks go to Silke Dannenmaier and Annette Römer for their administrative support.

I would like to thank my parents, Annerose and Dieter, for all their support over the years, despite my rare visits in Nordhausen. Finally, I would also like to thank my sister Elke and my nephew Philip for their understanding that I had too little time for them in the last years.

# Contents


**1 Introduction** **1**
- 1.1 Motivation . . . 1
- 1.2 Related Work . . . 4
- 1.3 Thesis Goals . . . 8
- 1.4 Outline . . . 9

**2 Statistical Methods** **11**
- 2.1 Speech recognition as a Classification Problem . . . 12
- 2.2 Extraction of Relevant Features . . . 12
- 2.3 Models and Parameter Estimation . . . 13
- 2.4 Significance Tests . . . 20

**3 Hyperarticulation in the Context of ASR** **23**
- 3.1 Definition of Hyperarticulation . . . 23
- 3.2 Corpus Collection . . . 25
- 3.3 Recognition Experiments . . . 26
- 3.4 Error Analysis . . . 29
- 3.5 Use of Hyperarticulated Training Data . . . 40
- 3.6 Summary . . . 44

**4 Compensation Techniques** **45**
- 4.1 Duration Modeling . . . 47
- 4.2 Pronunciation Modeling . . . 49
- 4.3 Separate Acoustic Model Sets . . . 58
- 4.4 Hyperarticulation in Context Decision Trees . . . 62
- 4.5 Summary . . . 67

**5 The Articulatory Vector Space** **69**
5.1 Articulatory Phonetics . . . . . 69
5.2 Articulatory Modelling for ASR - a Review . . . . . 75
5.3 Hyperarticulation - Warping in an Articulatory Domain . . . 76
5.4 Statistical Modeling of Acoustic Events . . . . . 84
5.5 Detection of Articulatory Properties . . . . . 89
5.6 Speech Recognition with Vector Models . . . . . 94
5.7 Analysis of Contrastive Attributes . . . . . 96
5.8 Vector Models and Model Selection . . . . . 99
5.9 Utterance Combination . . . . . 100
5.10 Summary . . . . . 101

**6 A perception study** **103**
6.1 Experimental Setup . . . . . 104
6.2 Results . . . . . 105
6.3 Validation . . . . . 106

**7 Investigations on Portability** **109**
7.1 Transfer to Other Languages . . . . . 109
7.2 Transfer to Other Speaking Modes . . . . . 114

**8 Conclusions** **117**

**Appendices** **119**

**A Phoneset** **121**

**B Training Data** **125**

**C Test Data** **143**

**Bibliography** **151**

# List of Tables


3.1 Database for normal and hyperarticulated speech. . . . . 26
3.2 Error Rates on normal and hyperarticulated speech. . . . . 29
3.3 Statistical Test comparing likelihoods on normal and hyperarticulated data. . . . . 30
3.4 Average phone duration. . . . . 32
3.5 Average phone duration according to manner of articulation. . 32
3.6 Average phone duration according to place of articulation. . . 32
3.7 Phone duration on normal and hyperarticulated data, t-test with $\alpha = 0.05$. . . . . 33
3.8 Speaking rate (mrate) on normal and hyperarticulated data, t-test with $\alpha = 0.05$. . . . . 35
3.9 Fundamental frequency in Hz on normal and hyperarticulated data, t-test with $\alpha = 0.05$. . . . . 36
3.10 Recognition performance with respect to $F0$. . . . . 37
3.11 Significant differences at formant differences under hyperarticulation, t-test with $\alpha = 0.05$. . . . . 40
3.12 Supervised adaptation on hyperarticulated speech. . . . . 42
3.13 Supervised MLLR on different training sets. . . . . 43
3.14 Error rate versus amount of hyperarticulated adaptation data. 43

4.1 HMM topologies for hyperarticulated speech . . . . . 47
4.2 Effect of estimating transition probabilities . . . . . 48
4.3 Speaker dependent transition probabilities . . . . . 49
4.4 Ranking of top 4 vowel recognition candidates, normal speech. 52
4.5 Ranking of top 4 vowel recognition candidates, hyperarticulated speech. . . . . 52
4.6 Patterns of phone variation. . . . . 54
4.7 Dictionary size. . . . . 57

4.8 Pronunciation Modelling. . . . . 57
4.9 Comparison of Meeting with SWB models and supervised adaptation (results in word error rate). . . . . 59
4.10 Model specialisation for normal and hyperarticulated speech. . 59
4.11 Model selection using an oracle. . . . . 60
4.12 Specialised models: likelihood selection. . . . . 61
4.13 Specialised models: speaking rate and pitch based selection. . 62
4.14 Comparison of adapted meeting models with HSC models. . . 65
4.15 Tree generation with hyperarticulated questions. . . . . 65
4.16 Splits relating to manner of articulation. . . . . 66
4.17 Splits relating to place of articulation. . . . . 66

5.1 Consonantal place of articulation. . . . . 72
5.2 Contrastive Attributes : *doubts* vs. *doubt*. . . . . 81
5.3 Detection accuracy for manner of articulation attributes. . . . 92
5.4 Detection accuracy for place of articulation attributes. . . . . 92
5.5 Detection accuracy for global attributes. . . . . 92
5.6 Basis Elements. . . . . 95
5.7 Recognition experiments with vector models (results in word error rates). . . . . 95
5.8 Predictions of contrastive attributes. . . . . 97
5.9 Predictions of contrastive manner of articulation attributes. . 98
5.10 Predictions of contrastive place of articulation attributes. . . . 98
5.11 Enforcing contrastive attributes (results in word error rate). . 99
5.12 Vector models and model selection (results in word error rates).100
5.13 Utterance combination (results in word error rates on normal speech). . . . . 101

6.1 Counts for each class and labeler . . . . . 105

7.1 German Corpus for normal and hyperarticulated speech. . . . 110
7.2 Performance degradation at hyperarticulated speech for German and English. . . . . 110
7.3 Sub test groups partitioned according to error rate changes . . 111
7.4 Phone durations versus error rate . . . . . 111
7.5 Word error rate as a function of $F0$ changes . . . . . 111
7.6 Model Selection : Comparison of German and English . . . . . 112
7.7 Articulatory Vector Space : Comparison of German and English112

7.8 Selection of vector models : Comparison of German and English 113
7.9 Utterance combination for German and English. . . . . 113
7.10 Comparison of phone and vector models on the SUSAS corpus (error rates). . . . . 116

# List of Figures


1.1 Pitch contour for the word *Leonard*, spoken normally (left) and hyperarticulated (right). . . . . 3

1.2 Class probabilities for the attributes *Fricative* (left) and *Plosive* (right) while pronouncing *doubts* normally and hyperarticulated. . . . . 4

1.3 Computer-elicited Hyperarticulate Adaptation Model (CHAM), with written permission from Sharon Oviatt [Oviatt '98] . . . . . 6

3.1 Phone alignment for *endorsement*, normally spoken (bottom) and hyperarticulated (top). . . . . 31

3.2 Phone duration vs error rate. . . . . 34

3.3 F1/F2 formant drift for speaker *spk2*. . . . . 38

3.4 F1/F2 formant drift for speaker *spk9*. . . . . 38

3.5 F1/F2 formant drift for speaker *spk4*. . . . . 39

4.1 HMM composition. . . . . 46

4.2 Pronunciation graph for *ABIDING*. . . . . 54

4.3 Pronunciation decision tree for /IX/. . . . . 56

4.4 Pitch contour for the word *Leonard*, spoken normally (left) and hyperarticulated (right). . . . . 62

4.5 Excerpt from the decision tree for /Z/. . . . . 65

5.1 Organs of the human speech production : (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, from [Lemmetty '99]. . . . . 70
5.2 Vocal tract as a system of cavities. . . . . 71
5.3 pulmonic consonants, [International Phonetic Association '99]. 73
5.4 vowels, [International Phonetic Association '99]. . . . . 74
5.5 $\Delta(o_t, a)$ for attribute *Fricative* while pronouncing *doubts*, normally and hyperarticulated. . . . . 83
5.6 $\Delta(o_t, a)$ for attribute *Plosive* while pronouncing *doubts*, normally and hyperarticulated. . . . . 84
5.7 Acoustic models for vector elements (example *doubts*). . . . . 88

6.1 Perception Study: User Interface . . . . . 104
6.2 Baseline system: error rates with respect to human perception 106
6.3 Training with hyperarticulated data: error rates with respect to human perception . . . . . 107
6.4 Articulatory vector models: error rates with respect to human perception . . . . . 108

# Chapter 1

# Introduction

The performance of today's automatic speech recognition (ASR) systems still depends on many factors limiting the usefulness of such systems. So-called speaker independent systems are state-of-the-art in ASR. They do not require an enrollment phase in order to achieve low error rates in controlled environments. However, the performance of such systems is, in fact, not speaker or speaking-mode independent. Moreover, it is often observed that an extraordinary speaking mode results in a drastic performance degradation. An important problem arises if users change their speaking mode in order to correct recognition errors. For humans, this is a natural reaction intended to disambiguate word confusions, but it causes even more recognition errors. The goal of this work is, therefore, to achieve a better understanding of the influence of speaking styles on ASR systems and, with this understanding, to develop algorithms to compensate for such variations.

## 1.1 Motivation

### Verbal Human-to-Machine Interaction

Besides the recognition of pre-recorded audio data, for example the transcription of broadcast news, automatic speech recognition plays an important role in creating user friendly computer interfaces. Dialogue systems are a key technology in supporting communication between humans and machines. Speaking style, dialect, speaking rate, accent, and even emotion can vary, depending upon the user or the system behaviour.

Examples of speaking style changes can be found in error recovery situations. Humans interacting with an automatic dialogue system change their speaking mode in order to react to recognition errors. Several studies have observed, e.g. in [Soltau & Waibel '98], that a user will expend more effort toward achieving better pronunciation in order to resolve recognition errors. From a functional point of view, the articulation efforts of the speaker depend on the listener's capability to recognise the utterance. As long as the voice interface works perfectly, sloppy speech will require only minimal articulation. If recognition errors occur and the user needs to repeat the utterance several times, the pronunciation will change to a *hyper-articulated* speaking mode. This reaction is quite similar to human conversations with hearing impaired people [Picheny et al. '86].

Contrary to the user's expectation, current state-of-the-art speech recognition systems fail to handle hyperarticulated speech. The recognition performance degrades significantly in such a speaking mode. In other words, humans make an effort to improve the recognition performance, but current systems react diametrically opposed to the speaker's efforts. This system behaviour is contrary to the way human-computer interfaces should work. A human's expectation that his or her attempt at clearer articulation will lead to better system performance will not be realized. One question that needs to be addressed is, therefore, why hyperarticulated speech has a negative impact on the performance of automatic speech recognisers.

The inability of current ASR systems to deal with hyperarticulated speech has consequences for dialogue systems. For example, the user might enter an endless loop of interaction. One possible scenario is the following: The speech recogniser will fail to recognise some words. The user will therefore support the system by switching to a clearer articulation. The recogniser will then create even more errors. The user will extend his efforts towards a hyper-clear speaking mode. This will lead to even more recognition errors. Finally, the speech interface becomes completely useless and the user will seek other modalities instead [Suhm '98].

## Disambiguation of Words

Interactive speech-based interfaces have a great potential to simplify access to modern information systems. As a matter of fact, however, the statistical nature of speech and the limitations of current ASR systems cause recognition errors. If a perfect recognition cannot be guaranteed, the interface must be

able to deal with recognition errors. If several repetitions are necessary to correct a recognition error, the user may switch to a *hyperarticulated* speaking mode. This speaking style is characterised by a very precise and accentuated pronunciation and a reduced rate of speech. Additionally, the position of the recognition error is often acoustically labeled by using several features, such as loudness, pitch, and duration. This effect is illustrated by the following example:

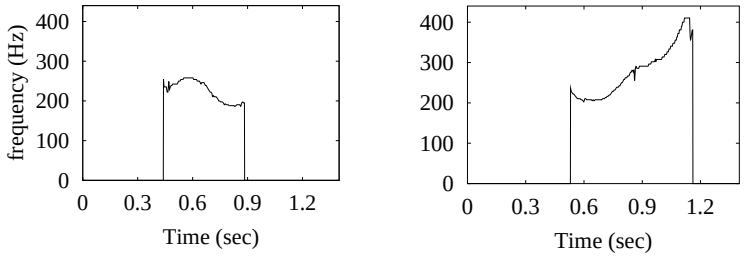


Figure 1.1: Pitch contour for the word *Leonard*, spoken normally (left) and hyperarticulated (right).

The word *Leonard* was confused with the word *Leopard*. In reaction to the error, the word was spoken again and pronounced very clearly in order to correct the mistake. Figure 1.1 exhibits changes in the pitch contour during a hyperarticulated speaking mode. The variation in the pitch contour in that particular case is used to encode the information of the previous recognition error. However, current preprocessing methods in automatic speech recognition attempt to filter pitch information, since they are normally considered to be irrelevant[1] As a consequence, later processing stages are not able to extract the information of the previous recognition error.

Further indicators used to disambiguate words can be extracted from articulatory attributes, e.g. place and manner of articulation. In an articulatory vector space, speech sounds will be treated as a composition of several articulatory attributes. Compared to a phoneme based approach, this representation allows a finer granularity and offers more insights into the kind of hyperarticulated effects that occur.

The data of the figure in 1.2 originates from the confusion of the word *doubts* with *doubt*. A different realisation of the articulatory attributes under

[1]There are a few exceptions, e.g. the fundamental frequency is sometimes used for recognition of tonal languages (Chinese).

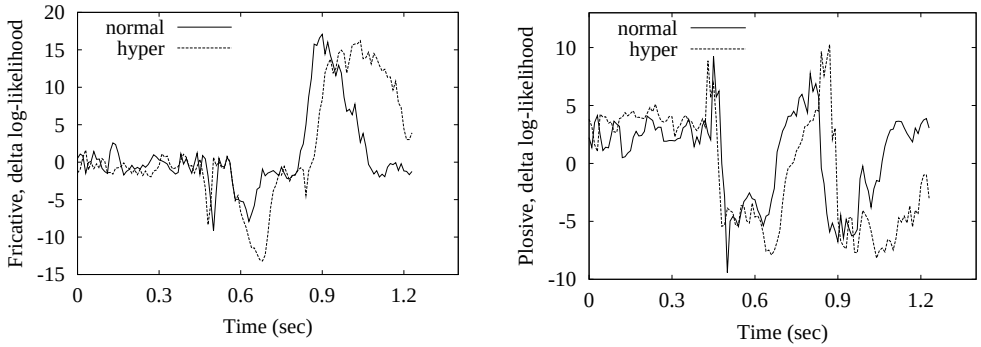


Figure 1.2: Class probabilities for the attributes *Fricative* (left) and *Plosive* (right) while pronouncing *doubts* normally and hyperarticulated.

hyperarticulation can be observed at the transition from the sound /t/ to the final /s/. The feature *Fricative* (corresponding to the /s/ sound) becomes activated at the end of the hyperarticulated audio signal compared to the corresponding realisation under normal conditions. On the other hand, the feature *Plosive* of /t/ gets suppressed at the end of the hyperarticulated utterance. This signaled that a plosive sound is not the final sound. Both the accentuation of the *Fricative* feature and the suppression of the *Plosive* feature place an emphasis of the missing /s/ sound at the mis-recognition of *doubt* instead of *doubts*.

## 1.2 Related Work

Hyperarticulated and related speaking modes have been investigated in several studies in recent years. This section is devoted in the review of those studies.

1. Picheny investigated in [Picheny '81] the acoustic characteristics of clear and conversational speech when talking to hearing impaired people. The study used a set of nonsense sentences which were spoken by three speakers. The intelligibility was tested by five listeners with sensorineural hearing losses. It was found that the human recognition accuracy is significantly higher for clear speech. Their results indicate that formant frequencies of vowels change to their "target values" in

clear speech. Furthermore, the speaking rate is greatly reduced in clear speech. Changes in the long-term spectrum were not significant.

2. Shriberg, Wade and Price presented in [Shriberg et al. '92] an analysis of factors affecting performance of spoken language systems. The authors studied how users adapt to a spoken language system for air travel information (Darpa ATIS task). The study revealed a relationship between hyperarticulation and recognition errors. Users (mal-)adapted to the system by speaking more clearly and overenunciated words which resulted in a significant higher error rate.

   In a further experiment, users were given instructions to avoid hyperarticulated effects but rather to speak naturally. The instructions resulted indeed into a smaller degree of hyperarticulation and to improved recognition performance. However, the difference in error rate was not reliable due to data sparseness.

3. Lindblom [Lindblom et al. '92] published a comparative study on acoustic-phonetic data for different speaking styles. He examined conversational speech, clear speech, and "baby talk". The observed pronunciation patterns varied significantly across the speaking style. To explain these observations, he proposes viewing the pronunciation variations as products of *adaptation*. Phonetic gestures and signals are modulated and tuned adaptively with respect to the communication demands. In other words, Lindblom's theory of speech adaptation in human-human conversations suggests that acoustic variability occurs as a functional adaptation of the speaker to the listener. A speaking mode can be explained as a point on the "articulation-axis", whereby the ends are *hypo-clear* and *hyper-clear* speech. Hypo-clear speech can be characterised by a *minimum* effort of articulation and requires that a listener is able to fill in missing phonetic information. On the other hand, *hyper-clear* speech needs more effort by the speaker. In the context of adaptation to a listener, the speaker attempts to produce an "ideal" acoustic realisation of speech units.

4. Oviatt proposed in [Oviatt '98] an adaptation model to describe changes in human speech when talking to a computer. Her model is based on Lindblom's theory in the context of human-computer error resolution. She proposed a two-stage model of a speaker's adaptation,

the Computer-elicited Hyperarticulate Adaptation Model (CHAM). The first stage of human adaptation consists of duration changes only and occurs in situations with low error rates. If the error rate increases, the second stage of a human's adaptation arises. According to the model, changes in pitch, amplitude, and articulation will be observed (see figure 1.3).

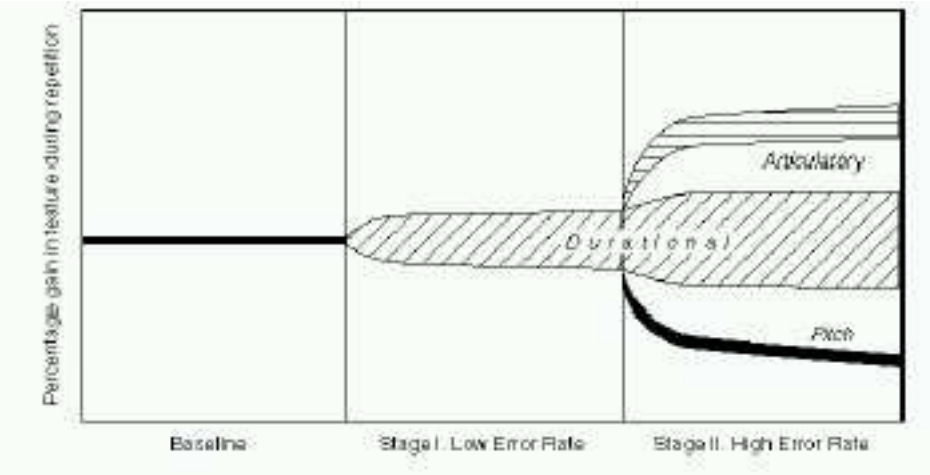


Figure 1.3: Computer-elicited Hyperarticulate Adaptation Model (CHAM), with written permission from Sharon Oviatt [Oviatt '98]

5. Levow presented in [Levow '98] a study of spoken corrections in Human-Computer Dialogues. She analysed 300 pairs of original and repeated correction utterances. The data were collected in a field trial using a voice-only interface to common desktop applications, such as e-mail, calendar, and stock quotes. She distinguished between corrections of mis-recognition errors (CME) and corrections of rejection errors (CRE). A general shift from conversational to clear speech was observed for corrections of rejection errors. In contrast, corrections of misrecognition errors exhibit additional pitch accent features. Duration and pause features exhibited significant differences. The minimum pitch decreased in male speakers. Furthermore, significant increases in the steepest rise of pitch were measured. She concludes: "These contrasts will be shown to ease the identification of these utterances as corrections and to highlight their contrastive intent". Based on these facts, she trained

a decision tree to distinguish between original utterances and repeated corrections. The decision tree uses a set of 38 attributes based on duration, pitch, and amplitude. Her system achieved a classification accuracy of 75%.

6. Studies to detect recognition errors are presented in [Hirschberg et al. '99] and [Hirschberg et al. 2000]. The experiments are based on a spoken dialogue system to retrieve information about train schedules. The dialogue manager is based on a finite state machine allowing mixed initiative approaches. Several prosodic features exhibit significant differences between correct and misrecognised utterances. Using these features and output from the ASR engine, including confidences, a decision tree is trained to distinguish correct from misrecognised utterances. The system using prosodic features and ASR output (hypotheses and confidences) has a prediction error of 10.79%. An error rate of 13.39% is achieved if only ASR output is used. Thus, prosodic features provide additional information for detecting misrecognitions.

7. In [Hirschberg et al. 2001], the question of how to detect utterances intended to correct previous recognition errors is addressed. The motivation for this work is that knowledge about recognition errors and user's correction might be useful for dapting the dialogue strategy. For example, the dialogue manager could switch to an "error repair mode", where the dialogue strategy is focused on repairing previous system errors. The dialogue strategy could also become more restricted in such situations. Instead of allowing mixed initiatives, the dialogue manager could switch to a confirmation strategy if problematic turns are detected. If high reliability is required, a human operator might be involved when correction utterances are detected. In order to detect such problematic turns, Hirschberg investigated features such as prosody, ASR output, and dialogue state to distinguish between corrections and non-corrections. A decision tree was trained based on these features and achieved a classification error of 15.72%.

8. Kienast observed in [Kienast et al. '99] articulatory changes in emotional speech. Her experimental results are based on recordings from three actors. A high degree of articulatory simplification was observed

in utterances expressing sadness and fear. "Happy" speech seemed to exhibit articular movements similar to neutral speech.

9. Holtzapfel investigated in [Holtzapfel et al. 2002] and [Holtzapfel 2003] the use of emotions for dialogue strategies. The idea behind this work is that the detection of emotions can guide the dialogue manager to choose an appropriate dialogue strategy according to the emotional state of a user. Holtzapfel integrated emotional parameters in task-oriented dialogue systems [Denecke 2002]. He introduced variables for the emotional state, both for the user and for the system. Emotional information is thereby encoded by corresponding facets of the feature structure. Two facets are used to store information about emotion: the type of emotion (neutral, stressed, happy, etc.) and the accumulated emotion score. The dialogue manager can use these facets to adapt the dialogue goals and the strategy for reaching the goals. Furthermore, emotions were used for the disambiguating of commands in a humanoid robot scenario.

It should also be mentioned that different strategies for dealing with recognition errors are possible. Suhm investigated in [Suhm '98] multiple modalities for interactive error recovery. In his study, he investigated interfaces for speech, handwriting, and text (typing). The speech interface could handle spontaneous speech as well as spelling sequences. Mouse gestures for selecting hypotheses from n-best lists were also used. He observed that users switch from one modality to another in order to resolve recognition errors. Although users initially prefer speech for correction, they learn with experience to prefer the most accurate modality. Thus, a system using multiple modalities can exploit different capabilities for reacting to recognition errors. But it was also shown that speech interfaces are the fastest way for input, if the underlying speech recogniser works well. Therefore, if speech recognisers learn to deal with hyperarticulated speech, the user might prefer to use the speech interface instead of switching to other modalities.

## 1.3 Thesis Goals

Summarising the above-mentioned investigations, acoustic and prosodic analyses of hyperarticulated speech were reported in these studies. In two studies, classifiers were presented in order to distinguish between normal and repeated

or misrecognised utterances. These studies all lack an analysis of the recognition performance itself in the context of hyperarticulation. As we will show in this thesis, hyperarticulation causes a significant increase in recognition errors.

- The first goal of this thesis is, therefore, to answer the question: What are the differences between normal and hyperarticulated speech that cause the drastic increase in recognition errors. In short, a problem analysis is needed.
- If the reasons for the performance degradation are understood, we can address the next goal: The second goal of this thesis is to achieve a comparable recognition performance for both normal and hyperarticulated speech.
- To validate the thesis' goals, a corpus of normal and hyperarticulated speech is needed. Based on this corpus, the recognition performance will be evaluated using the algorithms and models developed in this work.

The scenario of verbal human-to-machine interaction and disambiguating word confusions indicates the relevance of hyperarticulated speech. But there is also a second argument why hyperarticulated speech should be investigated. Extraordinary speaking modes might be helpful for detecting invalid or weak assumptions in the current dominant approach of phoneme based Hidden Markov Models. Matched train/test conditions can cover wrong prerequisites. Switching to a hyper-clear speaking mode might, therefore, be used to discover invalid model assumptions.

## 1.4 Outline

The structure of this dissertation is based on an "analysis-to-synthesis" approach. An exact analysis of the problem is needed to find the necessary parts for assembling a solution for the problem.

This dissertation consists, therefore, of two major parts. In the first part, an analysis of the problem will be presented. Characteristic properties of hyperarticulated speech will be investigated and their relevance for compensating hyperarticulated effects will be discussed. Based on the results, novel

techniques for modeling acoustic properties and algorithms will be presented and evaluated in the second part of this work. Extensions of this work to other languages and speaking modes will be investigated in the final part. Starting with a brief introduction of statistical methods for automatic speech recognition, the dissertation is structured as follows:

1. Background (chapter 2)
2. Problem Analysis (chapter 3)
    - Properties of Hyperarticulated Speech
    - Influence of Speaking Style on the Recognition Performance
3. Algorithms and Models to Compensate for Hyperarticulation (chapter 4,5)
    - Articulatory Feature Structures
    - Adaptation Techniques
    - Hyperarticulation in Context Decision Trees
4. Extensions to other Speaking Styles and Languages (chapter 7)

# Chapter 2

# Statistical Methods

In this chapter, we present some background information for those readers who are not familiar with the technologies and concepts of automatic speech recognition (ASR). After introducing ASR as a statistical classification problem, typical feature extraction and parameter estimation techniques will be described. A short introduction to statistical significance tests will complete this chapter.

Current state-of-the-art speech recognition systems are based on the concept of Hidden Markov Models (HMM) to represent acoustic units. An HMM is a flexible finite state automat, together with a mechanism to propagate probabilities. Belief networks [Pearl '88] build a more general framework of this kind of automat. The syntax and semantic of a language are captured mostly by statistical n-gram language models (LM). The acoustic models (AM), together with the language models, form the backbone of a speech recogniser. From an algorithmic point of view, there are two basic problems. On the one hand, techniques for estimating the model parameters robustly are required. Therefore, large training samples are needed. On the other hand, the complexity of the acoustic and language models requires efficient search techniques in order to find the state sequence with the highest probability.

## 2.1 Speech recognition as a Classification Problem

Bayes' decision theory establishes the foundation for the formulation of the classification problem in speech recognition. The recognition of a word or state sequence can be expressed as maximising the a-posteriori probability over all elements in the search space, given the acoustic observations as a sequence of feature vectors $\vec{o}$. Having an utterance consisting of a sequence of $T$ feature vectors $O = (\vec{o_1}, .., \vec{o_T})$, the classification problem can be expressed as:

$$W^* = \underset{W}{\operatorname{argmax}} P(W|O) \tag{2.1}$$

$$= \underset{W}{\operatorname{argmax}} \frac{P(O|W) \cdot P(W)}{P(O)} \tag{2.2}$$

$$= \underset{W}{\operatorname{argmax}} P(O|W) \cdot P(W) \tag{2.3}$$

The maximisation process of the a-posteriori probabilities allows 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 $P(O)$ itself and can therefore be ignored. The a-priori probabilities $P(W)$ will be computed via the language model $P(W|\tau)$, where $\tau$ are the parameters of the language model. On the other hand, the acoustic model contains the class probabilities $P(O|W, \lambda)$ with parameters $\lambda$. Given this framework, research in speech recognition focuses on the estimation of the parameter of the language model $\tau$ and of the acoustic model $\lambda$ based on large training corpora.

## 2.2 Extraction of Relevant Features

The goal of the preprocessing step is to remove problem invariant features from the digital acoustic signal and to arrange the feature space to be appropriate for the acoustic models. In the first step, a short-time spectral analysis will be performed to extract features in the spectral domain. This step is valid, since it can be assumed that the speech signal is at least short-time stationary. The next assumption is that the phase spectrum does not contain meaningful information for speech recognition. Consequently, only the

power spectrum will be 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 [Fant '60], 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 normalisation is performed by cepstral mean subtraction (CMS). Additionally, the feature values can be divided by their variances (cepstral variance normalisation, CVN), but this requires reliable variance estimates. The next step induces some context information: cepstral features from adjacent windows are concatenated to 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 maximise the inter-class variances while minimising the intra-class variances. At the end of the feature processing, a sequence of $T$ feature vectors $O = (\vec{o_1}, .., \vec{o_T})$ is available. This sequence of feature processing steps is fairly standard in the ASR community, although there are several variations possible.

## 2.3 Models and Parameter Estimation

### Acoustic Models

Acoustic modeling deals basically with probabilities $P(O|W)$, where $W$ denotes a word or, more generally, an acoustic class and $O$ is a sequence of feature vectors. Since speech signals exhibit differences in a temporal and spectral domain, an appropriate model must deal with both dimensions in a statistically consistent way. The temporal changes can be modeled as a finite state automat with associated transition probabilities between the states. Attaching observation probabilities to each state will extend the automat to a Hidden Markov Model (HMM). This model is also called "first order Markov process" since the state probability depends only on the predecessor. Defining $S = \{s_1..s_n\}$ as a set of $n$ HMM states and $\mathcal{P} = S^T$ as the set of all state sequences of length $T$, the probability $P(O|W)$, given the model $\lambda$, can be computed as:

$$P(O|W,\lambda) = \sum_{q \in \mathcal{P}} \prod_{t} a_{s_{q_t} s_{q_{t+1}}} p(o_t|q_t) \tag{2.4}$$

The element $q \in \mathcal{P}$ represents one path through the state automat, and, furthermore, $q_t$ denotes the state index at time $t$. The variable $a_{s_i s_j}$ represents the probability for the transition from state $s_i$ to $s_j$. A set of start and end of states completes the HMM definition. The Forward/Backward algorithm computes these probabilities via dynamic programming with a complexity of $O(n^2 * T)$. The forward and backward probabilities are defined as:

$$\alpha_t(j) = P(o_1..o_t, q_t = s_j|\lambda) \tag{2.5}$$
$$\beta_t(j) = P(o_{t+1}..o_T|q_t = s_i, \lambda) \tag{2.6}$$

The conditional probability $P(O|W,\lambda)$ can be expressed as a sum over the $\alpha$'s and $\beta$'s:

$$P(O|W,\lambda) = \sum_{i} \alpha_T(i)\beta_T(j) \tag{2.7}$$

Furthermore, a recursive scheme to compute $\alpha$ and $\beta$ is available:

$$\alpha_t(j) = \sum_{i} \alpha_{t-1}(i) a_{ij} p(o_t|q_t = s_j) \tag{2.8}$$
$$\beta_t(j) = \sum_{i} \beta_{t+1}(i) a_{ji} p(o_{t+1}|q_{t+1} = s_i) \tag{2.9}$$

The Viterbi algorithm is similar to the Forward/Backward algorithm, But it requires only one pass. If the $\sum$ operator is replaced by the max operator, the best state sequence can be obtained:

$$q^* = \underset{q \in \mathcal{P}}{\operatorname{argmax}} \prod_{t} a_{s_{q_t} s_{q_{t+1}}} p(o_t|q_t) \tag{2.10}$$

The decoding engine searches for the best state sequence, whereby the language model probabilities will be included. Complex acoustic and language models require an efficient search space organisation, as described for example in [Soltau et al. 2001a].

Despite the availability of efficient algorithms to work with HMM's, there are several drawbacks. One important point is that the emission probabilities depend only on the current state. Thus, certain dependency or independency relations 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 factorisation of these random variables would allow 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. Belief networks [Pearl '88] allow the factorisation of such dependencies. However, parameter estimation and decoding in the framework of belief networks gives rise to a couple of problems.

## Kullback-Leibler Statistics

Parameter estimation for ASR focuses often on the emission probabilities, which usually are modeled by mixtures of Gaussians. Furthermore, practical considerations lead to diagonal covariance restrictions. The probability density functions (pdf) for emission probabilities are as follows:

$$P(o|s,\lambda) = \sum_i w_i N(o|\mu_i,\Sigma_i) \tag{2.11}$$

$$N(o|\mu,\Sigma) = \frac{e^{-\frac{1}{2}(o-\mu)^{-1}diag(\Sigma)^{-1}(o-\mu)}}{\sqrt{(2\pi)^n det(diag(\Sigma))}} \tag{2.12}$$

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

The parameter estimation is often based on the maximum likelihood principle[1]. A direct application of the maximum likelihood principle on HMMs is, however, not possible. Instead, Kullback-Leibler statistics are used to establish an iterative algorithm, known as the Baum-Welch re-estimation procedure. Inducing a variable $q$ for the (hidden) state sequence and initial

[1]Recently, the maximum mutual information criterion has been revived and is used in a lattice framework [Woodland & Povey 2002].

parameter $\lambda^0$, the log-likelihood of parameter $\lambda$ for an HMM can be expanded as:

$$\begin{aligned}
\mathcal{L}(\lambda) &= \log p(O|\lambda) \sum_{q \in \mathcal{P}} P(q|O, \lambda^0) && (2.13)\\
&= (\log P(O, q|\lambda) - \log P(q|O, \lambda)) \sum_{q \in \mathcal{P}} P(q|O, \lambda^0) && (2.14)\\
&= \sum_{q \in \mathcal{P}} \log P(O, q|\lambda) P(q|O, \lambda^0) - && (2.15)\\
&\quad \sum_{q \in \mathcal{P}} \log P(q|O, \lambda) P(q|O, \lambda^0) && (2.16)
\end{aligned}$$

The likelihood can be expressed as the Kullback-Leibler statistics $Q(\lambda, \lambda^0) = \sum_{q \in \mathcal{P}} \log P(O, q|\lambda) P(q|O, \lambda^0)$ and a rest term. Furthermore, the concavity of the log function leads to the following (Jensen-) inequality:

$$\begin{aligned}
\sum_{q \in \mathcal{P}} P(q|O, \lambda^0) \log \frac{P(q|O, \lambda)}{P(q|O, \lambda^0)} &\leq log \sum_{q \in \mathcal{P}} P(q|O, \lambda^0) \frac{P(q|O, \lambda)}{P(q|O, \lambda^0)} && (2.17)\\
&= 1 && (2.18)
\end{aligned}$$

Maximising the parameter $\lambda$ with respect to the Kullback-Leibler statistics, $Q(\lambda, \lambda^0) \geq Q(\lambda^0, \lambda^0)$ will increase the likelihood $\mathcal{L}(\lambda) \geq \mathcal{L}(\lambda^0)$. In the HMM framework, the term $P(q|O, \lambda^0)$ in $Q(\lambda, \lambda^0)$ denotes the state occupancies obtained using initial model parameters. The Baum-Welch algorithm will increase the likelihood in each training iteration. However, the final model parameters depend on the initial settings $\lambda^0$. Kullback-Leibler statistics are not only used for HMM parameter estimation, but also for mixtures of Gaussians.

## Vocal Tract Length Normalisation

Vocal Tract Length Normalisation (VTLN) is a feature transform which attempts to normalise the frequency changes due to different vocal tract lengths [Andreou et al. '94]. 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 by warping the frequency axis. To that end, a piece-wise linear function $f(\omega)$ can be employed:

$$f(\omega) = \begin{cases} \alpha\omega & : \ \omega < \omega_0 \\ \beta\omega + \gamma & : \ \omega \geq \omega_0 \end{cases} \tag{2.19}$$

whereby $\beta$ and $\omega$ can be obtained via the constraints $f(\omega_0)$ and $f(\omega_N)$. The warping factor $\alpha$ can be estimated by a maximum likelihood criterion:

$$\mathcal{L}(\alpha) = \sum_t \log(J(\alpha)P(f(o_t, \alpha)|\lambda)) \tag{2.20}$$

A Brent search is often used, since no closed-form solution is available. Furthermore, the derivate $J(\alpha)$ is ignored and the resulting function no longer satisfies the requirements for being a pdf.

## Model Adaptation

The maximum likelihood criterion can also be used for estimating a linear transform of the model parameters [Leggetter '95]. In the context of mixtures of Gaussians, a mean adaptation can be represented by such pdf's:

$$P(o|s, \lambda) = \sum_i w_i N(o; A\mu_i, \Sigma_i) \tag{2.21}$$

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

$$Q(A, A^0) = c - \sum_{i,t} \gamma_i(t)(c_i + (o_t - A\mu_i)^T \Sigma_i^{-1}(o_t - A\mu_i)) \tag{2.22}$$

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

$$\frac{dQ(A, A^0)}{dA} = 0 \tag{2.23}$$

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

$$\sum_{i,t} \gamma_i(t) \Sigma_i^{-1} o_t \mu_i = \sum_{i,t} \gamma_i(t) \Sigma_i^{-1} A \mu_i \mu_i \tag{2.24}$$

## Feature Adaptation

Linear transforms can also be applied in the feature space. This technique has some advantages over model adaptation since combinations with adaptive training schemes and Gaussian selection algorithms are easy to realise. If a pdf $p(x)$ and a feature transform $f(x)$ are given, an appropriate pdf with respect to $f$ would be $\hat{p}(x) = p(f(x))\frac{df(x)}{dx}$. This ensures that the probability mass is conserved:

$$\int p(x)dx = \int p(y)dy = \int p(y)\frac{dy}{dx}dx = \int p(f(x))\frac{df(x)}{dx}dx = \int \hat{p}(x)dx \tag{2.25}$$

If $f : \vec{x} \rightarrow \vec{y}$ 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 $f(x) = Ax$ is, therefore:

$$Q(A, A^0) = c + \sum_{i,t} \gamma_i(t)(\log|A| - c_i - \frac{1}{2}(Ao_t - \mu_i)^T \Sigma_i^{-1}(Ao_t - \mu_i)) \tag{2.26}$$

The Jacobian $|A|$ term complicates the optimisation process. However, the Laplace development for a row $j$ allows the representation of the Jacobian as:

$$|A| = \sum_{jk} a_{jk}\tilde{a}_{jk} \tag{2.27}$$

$$\tilde{a}_{jk} = (-1)^{j+k}|A_{jk}| \tag{2.28}$$

whereby $\tilde{a}_{jk}$ denotes the adjunct of $A$, given $j$ and $k$. This allows the implementation of an iterative optimisation scheme, working row by row. The adjunct's $\tilde{a}_{jk}$ will be kept fixed while optimising the row $j$.

## Semi-tied full Covariances

Semi-tied full Covariances (STC) [Gales '99] or Maximum Likelihood Linear Transform (MLLT) [Gopinath '98] introduce linear transforms for covariance modeling. The motivation for this approach is that diagonal covariances are used for practical reasons, but the observation space does not allow this since the features are correlated. A better parameter sharing scheme may be achieved by sharing the full transform matrices. The pdf is structured as follows:

$$P(o|s,\lambda) = \sum_i w_i N(o;\mu_i, A^T\Sigma_i A) \tag{2.29}$$

whereby $\Sigma_i$ 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:

$$P(o|s,\lambda) = |B| \sum_i w_i N(Bo;B\mu_i, \Sigma_i) \tag{2.30}$$

The resulting Kullback-Leibler statistics has the same form as for the feature adaptation with the exception that the same matrix $B$ is used to transform $\mu$ additionally:

$$Q(B,B^0) = c + \sum_{i,t} \gamma_i(t)(\log|B| - c_i - \frac{1}{2}(Bo_t - B\mu_i)^T\Sigma_i^{-1}(Bo_t - B\mu_i)) \tag{2.31}$$

## Language Models

The language model (LM) deals with the probabilities $P(W)$, where $W = w_1..w_n$ 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 no algorithm to learn the structure from data is available so far. Human labour is, therefore, required to write grammars. For tasks covering large domains, statistical n-gram models are popular. The word history is constrained to the last $n$ words. Lack of training data and disc space limits the word history to three, resulting in tri-gram models.

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.

$$P(W) = \prod_i P(w_i|w_{i-1}, w_{i-2})) \tag{2.32}$$

## 2.4 Significance Tests

The non deterministic nature of speech makes it desirable to validate experimental results with significance tests. The basic idea is to establish a null-hypothesis $H_o$ *before* starting the experiment and asking if the experimental results confirm the hypothesis. Details can be found in Brandt's textbook [Brandt '75]. A statistical test $Test(T, A)$ is given by a verification function $T : O \to \mathbb{R}$ and a set $\mathcal{A} \subset \mathbb{R}$, typically a confidence interval. The Null-hypothesis $H_o$ can be rejected if $T(x) \in \mathcal{A}$. An example can illustrate this: an experiment is planned to investigate whether the formant frequencies differ between normal speech and hyperarticulated speech. An appropriate null-hypothesis is that the differences are randomly distributed. It is further assumed that the formants are in a normal distribution. An appropriate test is, therefore, the student-test. The confidence interval is the $\alpha$-quantile of the $t-$function. If the verification function is higher than a certain significance level $\alpha$, it can be concluded that the differences are not randomly distributed. Further conclusions cannot be drawn.

For our purposes, the student-test (T-Test) and the F-Test for variance homogeneity are relevant. The student-test requires that the samples are distributed by a normal density with homogenous variances. Given two sample sets $S1$ and $S2$ with homogenous variances, the verification function for the student test is given by:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}(\frac{1}{N_1} + \frac{1}{N_2})}} \tag{2.33}$$

The sample means are denoted as $\bar{x}_1, \bar{x}_2$, the variances as $s_1^2, s_2^2$, and the sample sizes by $N_1, N_2$. The critical t-score is given by the $\alpha$-quantile of the student-function:

$$s(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}} \tag{2.34}$$

The Gamma distribution is hereby defined as $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx$.

The variance homogeneity can be examined via the F-test. The verification function is given by:

$$f = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} \tag{2.35}$$

The underlying Fisher distribution $F(m, n)$ is defined as:

$$f(x) = \frac{(\frac{m}{2})^{\frac{m}{2}}\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}x^{\frac{m}{2}-1}(1 + \frac{m}{n}x)^{-\frac{m+n}{2}} \tag{2.36}$$

The $\alpha$-quantile from the Fisher distribution can be obtained analogously to the student distribution.

# Chapter 3

# Hyperarticulation in the Context of ASR

In order to understand the influence of hyperarticulation in automatic speech recognition, a comparative study of different kinds of articulation is needed. For this thesis, we collected a database of audio examples with different speaking modes. The corpus collection, which is also essential for training acoustic models later on, is described in the first section. Initial experiments investigating the influence of hyperarticulation on the recognition performance will be reported in section 3.3 followed by an error analysis.

## 3.1 Definition of Hyperarticulation

The effects of hyperarticulation have already been described in the literature [Shriberg et al. '92, Oviatt '98, Levow '98, Hirschberg et al. '99, Kienast et al. '99] although there is no well-established and precise definition of the term itself. There are two ways to characterise hyperarticulated speech: the first method relies on the concepts *hyper* and *articulation* while the second characterisation is based on a pragmatic, application-oriented approach.

1. The term *articulation* refers to the act or manner of producing a speech sound. It is the aspect of pronunciation that involves bringing articulatory organs together so as to shape the sounds of speech [Wordnet 2003]. The word *hyper* originates from the Greek language

and means "excessive, extreme, exaggerative". Thus, hyperarticulation describes an extreme speaking style producing speech sounds in an exaggerative way. Hyperarticulated sounds result from a very exact movement of the articulatory organs. In order to create hyperarticulated sounds, humans will make an effort to reach ideal positions for the articulators. For example, Picheny [Picheny et al. '86] observed that the formant frequencies for vowels move to their "target" values in clear speech.

2. Lindblom's theory of functional adaptation [Lindblom et al. '92] motivates the second approach. It is based on the assumption that humans want to achieve a particular communication goal, if they attempt to produce very precise speech sounds. It is clear that hyperarticulation needs much more effort by the speaker - and there must be a reason for this behaviour. In human-human communication, hyperarticulation occurs to improve the intelligibility. It is shown in [Soltau & Waibel '98] and [Oviatt '98] that hyperarticulated speech occurs also in human-computer interactions in order to react to recognition errors.

   Instead of defining the hyperarticulated speaking mode by the acoustic properties[1], a problem- driven definition can be used. A hyperarticulated speaking mode arises in order to react to recognition errors. The intention of hyperarticulation is to disambiguate the true from the misrecognised word. The idea is to collect data in a scenario to correct recognition errors. Regardless of the absence of an exact definition, an analysis of these data will reveal the properties of speech used in an error recovery mode.

The approach chosen in this thesis for clarifying the concept of hyperarticulation is the pragmatic one. The strategy is to collect data using a scenario to correct recognition errors. This avoids an artifical definition of hyperarticulation. The advantage of this approach is that the data used for this work are "real world" data.

Another question arises, which is related to the problematic definition of the term hyperarticulation: how do humans judge the degree of hyperarticulation for a given utterance? This is an important question since we need to ensure that our pragmatic approach to use error-repair data for our

[1] which would be somewhat artifical anyway.

study is valid with respect to human perception. Therefore, we conducted a perception study to label the degree of hyperarticulation from a human perspective. The results of this perception study are presented in chapter 6.

## 3.2 Corpus Collection

The goal of this data collection is to compare different speaking styles. Since the performance of so called speaker-independent speech recognisers is speaker-, channel-, and domain-dependent, the corpus collection needs a careful design to allow analyses across speaking styles. On the other hand, the database should contain *realistic* audio recordings from real users. In short, the database has two requirements:

1. Realistic recordings of hyperarticulated speech

2. Prevention of performance dependent conditions across speaking styles

Taking the first point into account, the recordings were collected with a simulated dialogue system. The subjects who sat in front of a computer were asked to correct previous recognition errors. The subjects were not told that the system was a simulation. In order to induce hyperarticulated speech in a realistic way, we analysed typical errors of our LVCSR system and generated a list of frequent word confusions. In most cases, inflections and phonetically similar words cause recognition errors.

The recording scenario consisted of two sessions. In the first session, the subjects used the dialogue system as usual. After that, a list of recognition errors from the first session was presented to the subjects. The users were then asked to correct the word confusions. The recognition errors were presented as phrases, e.g. "The word *recounting* was confused with *recounted*. Please repeat *recounting*.". There were up to three attempts allowed 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. Furthermore, subjects were naive users of speech technology, i.e. none were speech experts knowing how to work with ASR prototypes.

```
Session 1 : dialoge system under normal conditions
            record 50 turns per speaker
Session 2 : dialoge system in ''correction mode''
            present word confusions up to 3 times
            reverse order of words
```

The term "normal condition" refers here to a normal operation mode, where speech is produced without any attempts to diverge from a canonical pronunciation.

The advantage of this approach is that the database contains recordings comparable in domain, vocabulary, microphone, and environmental noise for each speaker across different speaking styles.

| | speaker | utterances | | speech | |
|---|---|---|---|---|---|
| | | normal | hyper | normal | hyper |
| train | 34 | 3506 | 3923 | 124 min | 158 min |
| test | 11 | 1171 | 1444 | 34 min | 57 min |
| all | 45 | 4677 | 5367 | 158 min | 215 min |

Table 3.1: Database for normal and hyperarticulated speech.

In total, the database consists of 4677 normal and 5367 hyperarticulated recordings from 45 subjects. The corpus was divided into a training set of 34 speakers and a test set of 11 speakers. The test set is approximately 91 min. The set of training speakers is rather small. The purpose of these data is, however, to allow supervised adaptation experiments using acoustic models trained on large corpora, e.g. the Switchboard and Broadcast News databases. In the following experiments, the described corpus will be referred to by the name HSC (hyperarticulated speech corpus). HSC-normal is used to denote the normal portion and HSC-hyper for the hyperarticulated part.

## 3.3 Recognition Experiments

As discussed in the previous section, the data we are interested in were collected in an error recovery mode of a dialogue system. From the user's intention point of view, the user speaks more clearly and accentuated in order to facilitate the recognition process. The question that arises is whether this

change of speaking style results in a reduction of recognition errors or not. To answer this question, we conducted a series of recognition experiments with a state-of-the-art speech recogniser.

## Experimental Setup

The system we used was trained on a large corpus (Switchboard, SWB) of around 300 hours of conversational telephony speech [Godfreq et al. '92]. The JANUS recognition toolkit [Finke et al. '97, Soltau et al. 2001b] developed at the Interactive System Laboratories provides a library and framework for building speech recognisers. The context decision tree is based on septa-phone models allowing a maximal context of 6, and it was created using a divisive clustering procedure based on an entropy criterion. The probability density functions are a mixture of Gaussians estimated with an algorithm entailing an incremental growth of Gaussians. Several normalisation and adaptation techniques are used, such as cepstral mean and variance normalization, or vocal tract length normalisation. The front-end uses linear discriminant analysis and semi-tied full covariances.

- acoustic models trained on SWB corpus
- entropy clustered poly-phones with a context of +/- 3
- 10,000 context dependent HMM states with a variable number of Gaussians
- training by incremental growing of Gaussians, 288,000 in total
- semi-tied full covariances
- cepstral mean removal, variance normalisation, linear discriminant analysis

A zero-gram language model was used together with a search vocabulary of around 8,000 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.

The acoustic models used for these experiments were developed for the RT-03 CTS (Rich Transcription 2003, conversational telephony speech) evaluation. A detailed description can be found in [Soltau et al. 2002b, Soltau et al. 2003]. In short, the training consists of these steps:

1. Train fully-continuous models (10k codebooks)
   (a) simultaneous diagonalisation to compute LDA on warped MFCC features
   (b) re-organise data according to context dependent HMM states
   (c) grow mixture components : (30 iterations)
   iterative merging and splitting of means and covariances
   (d) estimate semi-tied full covariances (4 iterations)
2. Train semi-continuous models (50k distributions)
   (a) FSA-SAT viterbi training (4 iterations)
   (b) MMIE training (1 iteration)

## Results

The recognition performance, as shown in table 3.2, indicates significant differences between normal and hyperarticulated speech. While an acceptable error rate of 25.6% is obtained under normal conditions, there is a relative error increase of more than 60% under hyperarticulation on average over all test speakers. An important aspect is the speaker dependency of the error increase. The error rate of some speakers exhibits drastic performance degradations, e.g. for *spk1*, *spk4*, or *spk5*. On the other hand, there is only a 4% increase in recognition errors for *spk10*.

These results suggest 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 hyperarticulated speech well. In summary, we showed with this experiment that:

- There are significantly more recognition errors at hyperarticulation.
- The reaction on word confusions is a speaker dependent effect in terms of an increase in recognition errors.

Additionally, the outcome of this study is confirmed by earlier experiments [Soltau & Waibel '98] using a different setup and for the German language.

| speaker | error rate | | relative increase |
|---|---|---|---|
| | normal | hyper | in error rate |
| spk1 | 16.8 | 35.4 | 110.7% |
| spk2 | 28.2 | 46.0 | 63.1% |
| spk3 | 19.4 | 23.4 | 20.6% |
| spk4 | 25.3 | 47.5 | 87.7% |
| spk5 | 12.4 | 44.7 | 260.5% |
| spk6 | 38.4 | 61.3 | 59.6% |
| spk7 | 18.2 | 21.0 | 15.4% |
| spk8 | 25.7 | 32.8 | 27.6% |
| spk9 | 38.4 | 64.5 | 41.9% |
| spk10 | 27.3 | 28.5 | 4.4% |
| spk11 | 33.3 | 53.4 | 60.4% |
| all | 25.6 | 41.6 | 62.5% |

Table 3.2: Error Rates on normal and hyperarticulated speech.

## 3.4 Error Analysis

### Acoustic Models: Observation probabilities

The question that we address now is *why* does hyperarticulation cause such a drastic increase in recognition errors. The likelihoods of the acoustic models given the observable data can be used to examine how the models fit with the different speaking modes. The likelihoods of the Hidden Markov Models can be computed via the viterbi algorithm. In order to discover systematical variations across speaking styles, we performed statistical tests to examine if the likelihoods differ between normal and hyperarticulated speech. A so-called *T-test*, or *student-test*, with an $\alpha$-quantile of 0.05 was used.

We can then interpret the results of the T-test in that the likelihoods exhibit significant differences across the speaking styles for 8 out of 11 speakers. In other words, the acoustic models do not match with hyperarticulated data using a significance level of 0.95.

| speaker | significance test |
|---|---|
| spk1 | $\surd$ |
| spk2 | $\surd$ |
| spk3 | $-$ |
| spk4 | $\surd$ |
| spk5 | $-$ |
| spk6 | $\surd$ |
| spk7 | $\surd$ |
| spk8 | $-$ |
| spk9 | $\surd$ |
| spk10 | $\surd$ |
| spk11 | $\surd$ |

Table 3.3: Statistical Test comparing likelihoods on normal and hyperarticulated data.

## Phone Duration

The likelihood differences indicated a mismatch between acoustic models and observations. There are several factors which may have contributed to the mismatch. One of these is the speaking rate. An analysis of the average phone duration gives us more detailed information of segmentation issues. We will examine the effect of hyperarticulation on speaking rate with respect to phone classes, speaker identity, and error rate.

The phone durations were estimated based on the state alignment computed with the viterbi algorithm. The procedure used true transcripts and standard three state HMM topologies.

To illustrate the duration changes between speaking styles, phone alignments for the word *endorsement* are depicted in figure 3.1. In general, the phone segments become longer in the hyperarticulated case. This stretching effect is, however, not evenly distributed for all phones. For example, the segments for /D/, /AO/, /R/ do not exhibit larger changes, while the segment for the final /T/ sound increased from 10ms to 27ms, a factor of 2.7. Another quite interesting aspect can be seen if we compare the segments for the /N/ sound. The duration of the first occurrence of /N/ is mainly twice the duration of normal articulation, but the second occurrence of /N/ is not affected by hyperarticulation. This example indicates that the increase in

duration is not only a global effect but phone and position dependent.

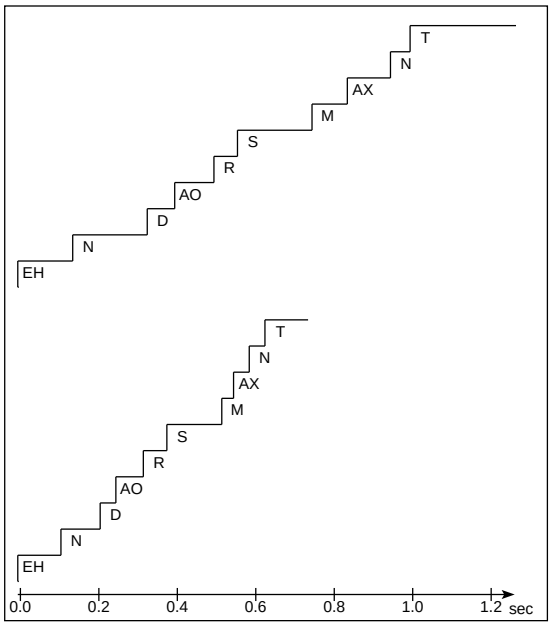


Figure 3.1: Phone alignment for *endorsement*, normally spoken (bottom) and hyperarticulated (top).

Besides this example, the results of a statistical analysis of the phone durations over the full test set are summarised in the following tables. The average segment length, reported in milli-seconds, increases by 28%. As shown in table 3.4 consonants are influenced considerably more by hyper-articulation than vowels. The effects for voiced and unvoiced sounds are comparable.

Another question is whether the *place* or the *manner* of articulation plays a role at the phone duration. To test this, we extracted the average segment length for each phone belonging to a certain place or manner of articulation. The tables 3.5 and 3.6 show the results of that examination. Plosive sounds like /p/ and /t/ exhibit an increase of around 44% on average, while the segments of fricative sounds are only 26% longer. A similar picture can be found for the place of articulation. Bilabial sounds, such as /p/, /b/, or /m/ show here significant differences with 44% longer segments, while glottal sounds are less affected.

| phone class | normal[msec] | hyper[msec] | relative increase |
|---|---|---|---|
| All | 99 | 127 | 28% |
| Vowels | 101 | 117 | 16% |
| Consonants | 100 | 132 | 32% |
| Voiced | 97 | 122 | 26% |
| Unvoiced | 106 | 137 | 29% |

Table 3.4: Average phone duration.

| phone class | normal[msec] | hyper[msec] | relative increase |
|---|---|---|---|
| Plosive | 79 | 114 | 44% |
| Nasal | 95 | 127 | 33% |
| Flap | 45 | 53 | 18% |
| Fricative | 124 | 156 | 26% |
| Approximant | 79 | 104 | 32% |
| Lateral | 92 | 119 | 29% |

Table 3.5: Average phone duration according to manner of articulation.

| phone class | normal[msec] | hyper[msec] | relative increase |
|---|---|---|---|
| Bilabial | 80 | 115 | 44% |
| Labiodental | 113 | 133 | 17% |
| Alveolar | 103 | 135 | 31% |
| Palatal | 57 | 79 | 38% |
| Velar | 86 | 118 | 37% |
| Glottal | 148 | 181 | 22% |

Table 3.6: Average phone duration according to place of articulation.

Recapitulating these results, the phone duration increases significantly if hyperarticulation occurs. The effect is phone and position dependent. Vowels are less affected. There are differences according to place and manner of articulation. Furthermore, the duration changes depend on the speaker identity as shown in table 3.7. The duration per speaker is computed as the average over all phones. Despite the speaker with a higher speaker rate (spk8), all changes are statistically significant using a level of $\alpha = 0.05$. The relative duration change varies from $-3\%$ to 63%.

| speaker | normal[msec] | hyper[msec] | relative increase | t-test |
|---|---|---|---|---|
| spk1 | 39 | 52 | 32% | $\surd$ |
| spk2 | 46 | 70 | 52% | $\surd$ |
| spk3 | 44 | 54 | 21% | $\surd$ |
| spk4 | 35 | 57 | 63% | $\surd$ |
| spk5 | 40 | 48 | 20% | $\surd$ |
| spk6 | 58 | 84 | 44% | $\surd$ |
| spk7 | 36 | 40 | 12% | $\surd$ |
| spk8 | 48 | 47 | -3% | $-$ |
| spk9 | 47 | 56 | 18% | $\surd$ |
| spk10 | 37 | 55 | 46% | $\surd$ |
| spk11 | 56 | 63 | 14% | $\surd$ |

Table 3.7: Phone duration on normal and hyperarticulated data, t-test with $\alpha = 0.05$.

Figure 3.2 shows the correlation of phone duration with error rate. It can be observed that speakers with a higher phone duration have a higher error rate. Moreover, this is valid both for normal and hyperarticulated speech. Considering the data points for normal speech only allows recognition of the correlation between phone duration and error rate. Therefore, it can be concluded that at least a part of the performance degradation at hyperarticulation can be directly attributed to higher phone durations.

## Variation in Speaking Rate

The phone duration is measured using a forced alignment procedure. This approach requires transcripts, or at least hypotheses, to estimate the speaking rate. A possible application for the speaking rate is to use this as a

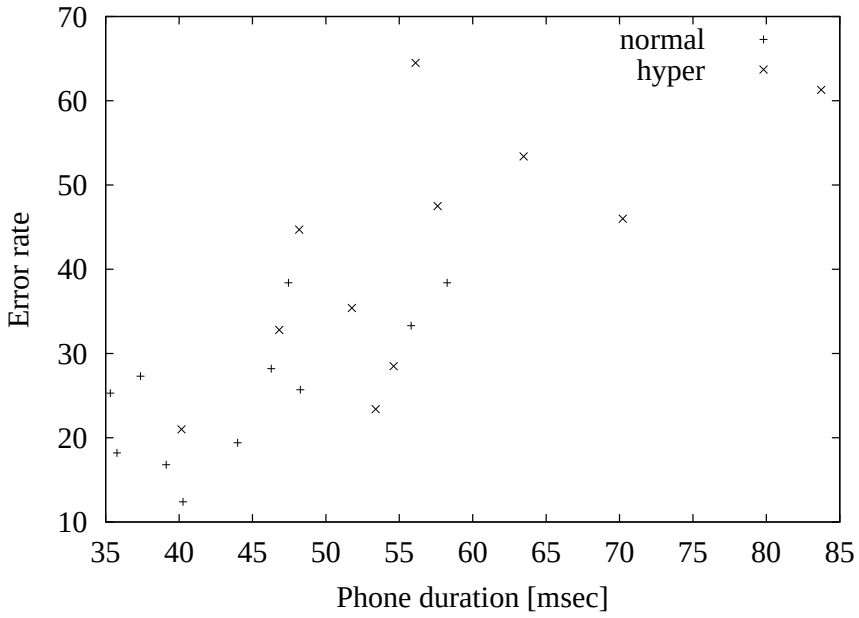


Figure 3.2: Phone duration vs error rate.

criterion for selecting an appropriate set of acoustic models. True transcripts are not available in such a scenario. Hypotheses can be used instead, but this, however, requires multiple decoding runs. A speaking rate estimator based on the audio signal only would be, therefore, advantageous for such scenarios. A combination of multiple signal based estimators is proposed in [Morgan & Fossler-Lussier '98] by Morgan and Fosler-Lussier. They used multiple measurements, like energy and peak counting, to form the *mrate* estimator. The *mrate* software is publicly available from [Morgan & Fossler-Lussier '98].

Table 3.8 shows the results obtained by this software from our data. The score *mrate* corresponds to the putative number of syllables per second. There are no significant differences (T-Test) across the speaking mode. On the other hand, the average phoneme length, measured on the true transcripts, increased by 28%.

There is a clear mismatch between the results from the previous section and the results obtained by the signal based estimator *mrate*. We need, therefore, to discuss which results are more reliable. The forced alignment

| speaker | mrate (normal) | mrate (hyper) | significant differences |
|---|---|---|---|
| spk1 | 2.54 | 2.59 | – |
| spk2 | 2.53 | 2.54 | – |
| spk3 | 2.54 | 2.47 | – |
| spk4 | 2.52 | 2.43 | – |
| spk5 | 2.60 | 2.55 | – |
| spk6 | 2.62 | 2.55 | – |
| spk7 | 2.62 | 2.59 | – |
| spk8 | 2.47 | 2.44 | – |
| spk9 | 2.62 | 2.57 | – |
| spk10 | 2.62 | 2.56 | – |
| spk11 | 2.62 | 2.56 | – |

Table 3.8: Speaking rate (mrate) on normal and hyperarticulated data, t-test with $\alpha = 0.05$.

procedure used true transcripts and complex acoustic models to compute speaking rate estimates. This information is not available for the *mrate* procedure. This procedure is solely based on the signal. Assuming that more information produces more reliable results, we can conclude that the forced alignment procedure produced more reliable speaking rate estimates. We will, therefore, assume that the results of the transcript based duration analysis are correct.

## Pitch Information

The example in the introduction, figure 1.1, compares the pitch contour between a normal and hyperarticulated speaking mode. Extracting pitch information consists of two tasks. First, the pitch detection itself needs to be performed. This step computes raw $F0$ values. The data do not necessarily provide the correct pitch values since multiples of the true $F0$ can occur. This makes it necessary to perform a smoothing step, also called pitch tracking, which takes into account the previous estimates. There are several methods to extract the raw $F0$ values which are based on auto-correlation, linear predictive coding, or cepstrum. The pitch tracker used in this work is based on the work by Medan, Yair, and Chazan [Medan et al. '91] and was

developed at Cambridge University.

| speaker | $F0$ (normal) | $F0$ (hyper) | significant differences |
|---|---|---|---|
| spk1 | 129.6 | 130.8 | – |
| spk2 | 203.5 | 199.6 | – |
| spk3 | 134.3 | 137.0 | – |
| spk4 | 126.6 | 126.8 | – |
| spk5 | 134.6 | 152.7 | $\surd$ |
| spk6 | 169.6 | 160.3 | $\surd$ |
| spk7 | 130.0 | 128.8 | – |
| spk8 | 145.5 | 138.2 | $\surd$ |
| spk9 | 186.0 | 186.1 | – |
| spk10 | 194.3 | 198.7 | $\surd$ |
| spk11 | 240.0 | 239.6 | – |

Table 3.9: Fundamental frequency in Hz on normal and hyperarticulated data, t-test with $\alpha = 0.05$.

The average fundamental frequency per speaker is shown in table 3.9. The individual $F0$ values for all utterances for a speaker were used to draw a sample for the significance test. Significant changes were observed in both directions: speakers exhibit higher, as well as lower, pitch values with hyperarticulated speech. In a second step, the effect of the fundamental frequency on the recognition rate was analysed. To that end, the test set was divided into three sub-groups: same, increasing, or decreasing $F0$. The $F0$ changes are measured as the average difference between the speaking modes. The error rate for each group is shown in table 3.10. Speakers exhibiting a decreased fundamental frequency have a relative error increase of 47.8%, while the group with increasing pitch has a 71.8% increase in recognition errors. This is an indication that the fundamental frequency has an impact on the recognition performance in a hyper-clear speaking mode. Otherwise, other factors must exist which affect the error rate, since the speakers without $F0$ changes also show significantly higher error rates.

| group | error rate | | relative increase |
|---|---|---|---|
| | normal | hyper | in error rate |
| same $F0$ | 28.1 | 49.4 | 75.8% |
| increasing $F0$ | 21.0 | 36.1 | 71.9% |
| decreasing $F0$ | 27.6 | 40.8 | 47.8% |

Table 3.10: Recognition performance with respect to $F0$.

## Vocal Tract Resonances

The duration analysis indicated that vowels do not change their characteristics in a temporal domain. However, hyperarticulation may influence vowels in a spectral domain. Fant's source-filter model [Fant '60] of the speech production process consists of three linear shift-invariant components: glottis, vocal tract, and radiation at the lips. The output of these components can be computed via a discrete convolution in the temporal domain. The discrete convolution becomes a simple addition in the log-spectrum domain. For speech recognition purposes, the point of interest is the vocal tract. However, in some tonal languages, such as Chinese, the glottis output is important as well, since the pitch contour is necessary to distinguish between phones.

The transfer function of the vocal tract can be described by its reflexion coefficients. This function can be represented as a complex polynomial. Its complex conjugate poles are called Formants. Suppose the vocal tract is simply a sequence of cylinders with different diameters. The corresponding transfer function would be:

$$V(z) = \frac{c}{1 - \sum_k^N \alpha_k z^{-k}} \tag{3.1}$$

The coefficients $\alpha_k$ of the predictor polynomial can be computed via Durbin's recursion algorithm [Rabiner '78], which minimises the predictor error using an autocorrelation approach. If we use a different representation of the predictor polynomial,

$$1 - \sum_k^N \alpha_k z^{-k} = \prod_k^{M/2} 1 - e^{-c_k T} \cos(b_k T) z^{-1} + e^{-2c_k T} z^{-2} \tag{3.2}$$

the formant frequencies $F_k = b_k / 2\pi$ can be extracted by computing the

complex poles via the Laguerre algorithm [Press et al. '88]. This all-pole model is valid for certain sounds only. Nasals and fricatives cannot be described completely by their formant frequencies since they require zeros in the transfer function to model anti-resonances. Therefore, the focus of this investigation lies on the vowels. The vowel time boundaries may be computed by a forced alignment procedure.

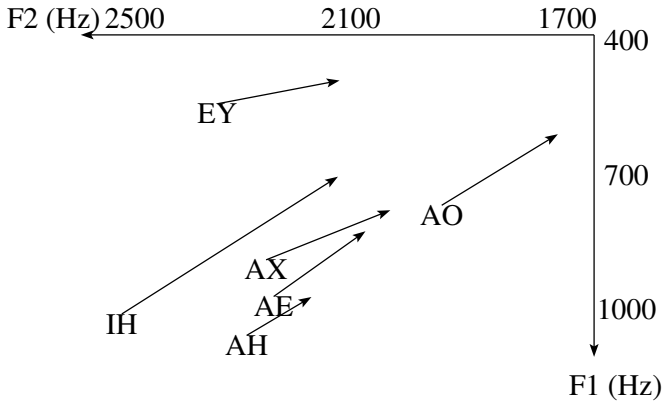


Figure 3.3: F1/F2 formant drift for speaker *spk2.*

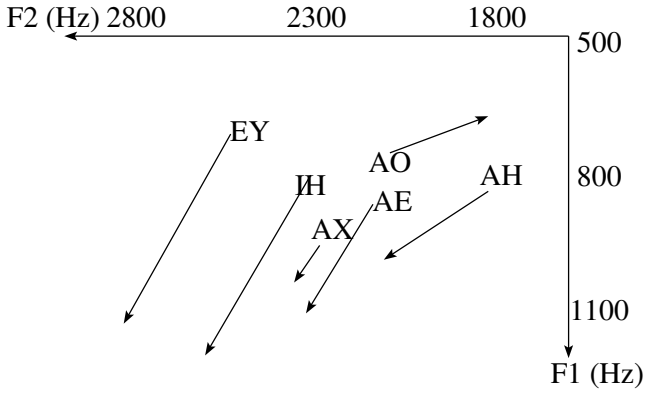


Figure 3.4: F1/F2 formant drift for speaker *spk9.*

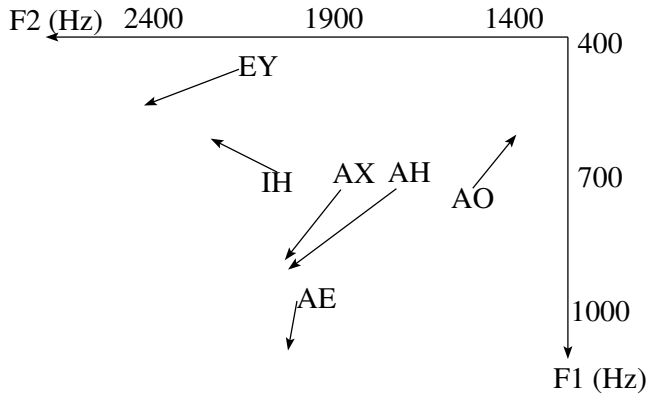


Figure 3.5: F1/F2 formant drift for speaker *spk4*.

Figures in 3.3, 3.4 and 3.5 exemplify how the formants drift under hyperarticulation. The formant frequencies F1 and F2 for the vowels /AE/, /AO/, /AH/, /AX/, /IX/, and /EY/ are computed via the LPC method as explained above. In all cases, the average formant frequencies change drastically under hyperarticulation. Independent of the absolute values of the average formant frequencies, the changes in the spectral domain are dependent on the speaker. Moreover, the phones exhibit different spectral changes even for the same speaker. In figure 3.4, the average formant frequencies F1 and F2 increase for both /EY/ and /IH/, while the formant shift for /AO/ moves in the opposite direction.

Besides these illustrations of spectral changes, a statistical test was performed to examine if there are significantly different formant frequencies in a hyper-clear speaking mode. As for the likelihood analysis, a T-test in conjunction with an F-test for the variance homogeneity was performed on a significance level of $\alpha = 0.05$.

The results in table 3.11 do not provide strong evidence that the formant frequencies change at a significant level. Only certain vowels exhibit significantly different spectral features for some speakers. The outcome of this analysis runs parallel with the observations of the phone durations. According to these results, vowels are only weakly affected by hyperarticulation both in a temporal and in a spectral domain.

| | AE | | AH | | AO | | AX | | EY | | IX | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 | F1 | F2 |
| spk1 | √ | √ | | | | | | | | √ | | |
| spk2 | | | | | √ | | | √ | | | | |
| spk3 | | | | | | | | | | | | |
| spk4 | | | | | | | √ | | | √ | | |
| spk5 | | | | | | | | | | | | |
| spk6 | √ | | | | | | | | | | | |
| spk7 | | | | | | | | | | | | |
| spk8 | | | | | | | | √ | | | | |
| spk9 | √ | √ | | | | | | | √ | √ | | |
| spk10 | | | | | | | | | | √ | | √ |
| spk11 | √ | | | | | | | | | | | |

Table 3.11: Significant differences at formant differences under hyperarticulation, t-test with $\alpha = 0.05$.

## 3.5 Use of Hyperarticulated Training Data

Assuming we are interested only in reducing error rates and do not need to understand hyperarticulated phenomena, we can simply try to collect hyperarticulated training data and estimate the model parameters using these data. It is obvious that such a solution has a limited applicability. First, it is rather difficult to collect sufficient training data for a hyperarticulated speaking mode. Secondly, the error analysis in the previous section gave us some clues that *invalid* model assumptions are at least one reason for the performance degradation. Now, invalid model assumptions cannot be "repaired" by just estimating the model parameters for this speaking style.

The intention of estimating the model parameters using hyperarticulated speech is to investigate how much error reduction is possible by using a brute-force method, while ignoring the reasons for the performance degradation. To that end, a corpus of 34 speakers with 2.6 hours of speech is available as table 3.1 shows. A pure Maximum-Likelihood estimation of the model parameters would be problematic due to the limited corpus size. For the recognition experiments reported in section 3.3, acoustic models trained on the SWB corpus were used. The hyperarticulated data can now be used to adapt these models. Two approaches were investigated:

1. Maximum-A-Posteriori Adaptation (MAP)
   MAP makes use of the knowledge about a *prior* distribution $g(\lambda)$ of the model parameters. Given a probability density function (pdf) $f(x|\lambda)$, the MAP solution for a data set $X$ is given by:

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}}\, f(X|\lambda)g(\lambda)$$

   Gauvain and Lee [Gauvain & Lee '94] have formulated the MAP solution for mixtures of Gaussians $\sum w_i N(x|\mu_i, \sigma_i)$ if the prior distribution function belongs to the conjugate family of the pdf. In that case, the prior distribution for $\mu_i$ and $\sigma_i$ is from type Normal-Wishart and accordingly the Dirichlet function for the mixture weights $w_i$. The parameter for the prior distribution can be estimated on a large training corpus while the MAP estimates are based on the in-domain adaptation data. MAP adaption can, therefore, be interpreted as an interpolation of out-of-domain and in-domain models.

2. Maximum Likelihood Linear Regression (MLLR)
   Leggetter and Woodland [Leggetter '95] used a set of linear transforms to adapt the mixture components. There are two types of transforms:

$$\tilde{\mu} = A\mu + b$$
$$\tilde{\sigma} = B^T \sigma B$$

   Maximising the corresponding Kullback-Leibler statistics leads to an estimation of the adaptation parameters. To make these transforms suitable for adaptation purposes, a regression tree is used to define a set of adaptation matrices. In these experiments, the basic regression classes rely on the individual Gaussian components. First, a binary tree is created by applying the k-means algorithm in a hierarchical way. The Gaussian components will hereby be clustered using the Euclidian distance of the means. Pruning of the regression tree depends on the amount of adaptation data available. The adaptation data associated with a node will be pushed to its parent node until a specific amount of data is collected. This ensures a reliable estimation of the adaptation matrices. As a consequence, the number of regression classes will be chosen dynamically, depending on the amount of adaptation data.

The experiments are based on the SWB models as described in section 3.3. 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. As mentioned before, the adaptation data is approximately 2.6 hours of hyperarticulated speech from 34 speakers. The results are given in table 3.12.

| acoustic models | error rate | | relative error increase |
|---|---|---|---|
| | normal | hyper | at hyperarticulation |
| baseline | 25.6% | 41.6% | 62.5% |
| MLLR | 21.9% | 35.0% | 59.8% |
| MAP | 23.4% | 37.9% | 61.9% |

Table 3.12: Supervised adaptation on hyperarticulated speech.

Supervised MLLR adaptation leads to an error reduction of 19% on hyperarticulated speech. MAP adaptation seems to be less effective. An error reduction of 10% was obtained with MAP. This can be attributed to the huge number of Gaussians of the seed acoustic models. The system has about $10,000$ context dependent states with more than $288,000$ Gaussians. Gaussian components having very small occupancy counts will more or less remain unchanged in MAP adaptation. The advantage of the MLLR regression tree is a better tying of Gaussian components. It therefore allows a better exploitation of the adaptation data.

Another interesting observation can be made by comparing the error rates of normal speech with hyperarticulated speech before and after adaptation. The results indicate that both normal and hyperarticulated speech profit from the adaptation using *hyperarticulated* training data. The relative error increase due to a hyper-clear speaking mode was reduced from 62.5% to 59.8% only. This result is surprising. The adaptation process itself works well; the problem rather is that the ratio between the errors of normal and hyperarticulated speech does not improve. One possible explanation could be that the adaptation compensates channel and domain mismatches[2] well, while only a moderate compensation of the speaking style is achieved. To examine this hypothesis, we conducted adaptation experiments using the

[2]The acoustic models were trained on conversational telephony speech.

normal speech portion of the database in table 3.1. These data come from the same 34 speakers used for the hyperarticulated adaptation experiments.

| adaptation data | error rate | | relative error increase |
|---|---|---|---|
| | normal | hyper | at hyperarticulation |
| 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 3.13: Supervised MLLR on different training sets.

The results in table 3.13 confirm the hypothesis. Independent of the speaking style of the adaptation data, significant error reductions were obtained both for normal and for hyperarticulated speech. However, the ratio between recognition errors of normal and hyperarticulated speech improved barely compared to the baseline models. A similar performance is achieved when using both speaking styles for adaptation.

| adaptation speaker | error rate |
|---|---|
| 0 | 41.6% |
| 5 | 37.6% |
| 10 | 37.0% |
| 15 | 36.3% |
| 20 | 35.9% |
| 25 | 35.6% |
| 30 | 35.3% |
| 34 | 35.0% |

Table 3.14: Error rate versus amount of hyperarticulated adaptation data.

In our final experiment, the influence of the limited training data size was investigated. To that end, a series of adaptation experiments using only fractions of the available training data were conducted. The number of regression classes was chosen automatically depending on the data size. The results are shown in table 3.14. The corresponding performance for normal speech is 21.9%. The error rate curve for hyperarticulated speech is not yet

in the saturated range, though the gap between normal and hyperarticulated speech is about 60% relative using all available data.

## 3.6 Summary

The analysis of hyperarticulated speech in context of automatic speech recognition led to the following observations:

1. Hyperarticulated speech causes a drastic increase in recognition errors.
2. Hyperarticulated changes depend on the speaker and phone identity.
3. Significant changes were observed both in a temporal and a spectral domain.
4. Vowels are less affected than consonants.
5. The changes of formant frequencies in a hyperarticulated speaking mode depends on the speaker identity (see figures 3.3, 3.4 and 3.5).
6. There is no evidence that the formants move toward their target values.
7. Adaptation of the acoustic models using hyperarticulated training data did not compensate for hyperarticulated effects (see table 3.13).

# Chapter 4

# Compensation Techniques

The intention of this chapter is to investigate compensation techniques for hyperarticulation in the context of a traditional ASR system. We will show that a limited amount of recognition errors can be reduced, but a real compensation of hyperarticulated effects cannot be achieved. We will systematically examine the ASR components regarding their behaviour on hyperarticulated speech. Despite the linguistic knowledge (which is not examined in context of hyperarticulation), ASR systems rely on the following knowledge sources:

1. Dictionary
2. Acoustic Models
   (a) Front-End
   (b) Model Topology
   (c) Transition Probabilities
   (d) Emission Probabilities

The preprocessing steps in the front-end module rely on psycho-acoustic knowledge, e.g. a logarithmic scaling of the signal energy and a frequency scaling by applying a filter bank. In this thesis, it is assumed that hyperarticulated speech does not effect such basic principles, and the front-end module does not need to be re-designed. Indeed, preliminary experiments did not show any evidence that a hyperarticulated front-end improves recognition results.

The American Heritage dictionary of the English language [Pickert 2000] defines *phonotactics* as "the set of allowed arrangements or sequences of speech sounds in a given language". Basically, the purpose of the dictionary is to map words to phone sequences. Pronunciation variants can be handled by allowing alternative entries in the dictionary or using more general pronunciation networks.

The HMM topology defines a structure on the sub-phonetic level. Thereby, a phone will be split into several temporal pieces, typically into a beginning, a middle, and an end state. The network layout, plus the corresponding transition probabilities, has an impact on the average phone duration. Modeling phone duration for ASR purposes often means working on the HMM topology. Both the dictionary and the Hidden Markov Model can be considered as (probabilistic) finite state automata. Moreover, both knowledge sources can be formed into a single Hidden Markov Model, as shown in figure 4.1 exemplarily.

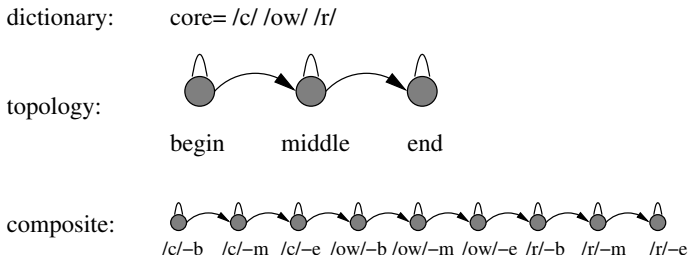


Figure 4.1: HMM composition.

Furthermore, this chapter discusses the question: Can hyperarticulated training data be efficiently used to build acoustic models? Under the premise that the model assumptions are valid in a hyper-clear speaking mode, compensation of hyperarticulation can be treated as reducing the mismatch between the model parameter and the test data. From an abstract point of view, this problem can be solved by estimating the parameters using appropriate training data. This will, however, introduce a new problem: If the models are trained on hyperarticulated speech only, a mismatch between these model parameters and normal speech will occur. An important aspect hereby is that a performance degradation for normal speech should be avoided if special hyperarticulated models are used. In the third section of this chapter, separate acoustic models for each speaking mode will be investi-

gated. Instead of generating two separate model sets, an integrated approach is evaluated in section four which relies on context decision trees.

## 4.1 Duration Modeling

Focusing on the HMM topology and the transition probabilities, which define a temporal partition or structure for acoustic units, the examination of the average phone durations in section 3.4 indicates the need of hyperarticulated duration models[1].

### HMM Topologies

The easiest way to perform a kind of duration modeling is to work on the HMM topology. The minimum phone duration is linked to the number of states since the decoding engine aligns at least *stateN* time segments to a phone model. So, varying the number of states is a very simple way to compensate for changes in the speaking rate. It is obvious that this technique has several disadvantages. First, the "design principle" of HMM topologies is mostly based on a trial-and-error method. This also makes it rather hard to work on phone and speaker dependent models. Secondly, the modeling power is rather limited.

Nevertheless, this type of duration modeling is examined for the sake of completeness. As shown in table 4.1, doubling the number of HMM states gives a small improvement on hyperarticulated data but degrades the performance on normal speech.

| topology | error rate | |
|---|---|---|
| | normal | hyper |
| 3state | 25.6% | 41.4% |
| 6state | 25.8% | 40.9% |

Table 4.1: HMM topologies for hyperarticulated speech

[1]To be more precise, duration models which fit hyperarticulated speech are needed.

## Transition Probabilities

Given the stochastic nature of speech, a more sophisticated approach for duration modeling should rely on a mathematical, statistical estimation method for transition probabilities in the HMM framework. The JANUS toolkit usually keeps the transition probabilities fixed.[2] However, the expectation-maximisation algorithm, known from the estimation problem for mixtures of Gaussians, can be applied here. Given the conditional probability for a transition from state $i$ to $j$ at time $t$:

$$\gamma_t(i,j) = P(q_t = i, q_{t+1} = j|O, \lambda) \tag{4.1}$$

which can be computed via the forward/backward method, the maximum likelihood solution for transition probabilities can be expressed as follows:

$$\pi(i,j) = \frac{\sum_t \gamma_t(i,j)}{\sum_{k=1}^{n} \sum_t \gamma_t(i,k)} \tag{4.2}$$

| transition probabilities | error rate | |
|---|---|---|
| | normal | hyper |
| fixed | 25.6% | 41.4% |
| trained | 24.3% | 38.7% |

Table 4.2: Effect of estimating transition probabilities

The 2.7% absolute improvement shown in table 4.2 demonstrates that keeping constant transition probabilities is not adequate for hyperarticulated speech. On the other hand, the normal portion of the test set also profits from trained transition probabilities. A possible cause for this is the mismatch with the original SWB training data.

## Speaker Adaptation

The term Speaker Adaptation is usually associated with acoustic modeling, particularly Maximum Likelihood Linear Regression. Speaker adaptation,

[2]It was found on various LVCSR tasks that these probabilities do not have an impact on the error rates. Therefore, these transition probabilities are often ignored and set to constant values.

however, is more than just transforming mixtures of Gaussians. Since the speaking rate is a speaker dependent factor, there is reason to believe that transition probabilities should be used in a speaker adaptive framework. This will require two decoding passes:

1. Decoding with speaker independent transition models
2. Estimation of transition models per speaker based on the hypotheses from the first pass
3. Decoding with speaker dependent transition models

| transition probabilities | error rate | |
|---|---|---|
| | normal | hyper |
| speaker independent | 24.3% | 38.7% |
| speaker dependent | 24.2% | 36.5% |

Table 4.3: Speaker dependent transition probabilities

The outcome of introducing speaker dependencies in duration modeling is shown in table 4.3. A significant error reduction of 2.2% was obtained for the hyper-clear portion of the test set. Returning to the analysis of phone durations in section 3.4, these results are in line with the observations made in this section: Speaking rate variations caused by a hyper-clear speaking mode, depends both on phone and speaker identity. Compensating hyperarticulation in a temporal domain leads to an error reduction of 4.9% absolute.

## 4.2 Pronunciation Modeling

When designing a dictionary, a few assumptions about the speaking style are made. The phonotactic knowledge encoded in the dictionary is based mainly on canonical, speaker independent pronunciations for each word. Typically, this assumption is valid for read speech only. A mismatch between the dictionary and the actual pronunciation can often lead to a significant performance degradation. For example, phones may be slurred or even omitted in spontaneous speech. Comparing hypo-clear with hyper-clear speech on an axis of "sloppiness", these speaking styles would lie on the opposite ends, while read

speech would be the "centre" on this axis. Obviously, this is not a precise model of the situation, but it can be used as an argument why it might be worthwhile investigating which phonological rules apply for hyperarticulated speech.

As we discussed in the introduction of this chapter, pronunciation modeling means finding appropriate phoneme sequences or networks[3]. What we discuss here is not building a dictionary from scratch, but investigating hyperarticulated variations from a standard pronunciation. In this sense, there are three types of phonological variations:

1. Substitutions
2. Insertions
3. Deletions

These types are sufficient enough to define new pronunciation variants if a dictionary is already given. The deletion of phones occurs quite frequently in sloppy speech, but it is rather unlikely that this happens for hyperarticulation since the speaking rate is significantly slower. But, as we have seen in the previous chapter, the slower speaking rate means higher phone durations but not necessarily inserted phones. Indeed, an informal investigation has shown that deletions and insertions of phones do not occur very frequently in a hyper-clear speaking mode.

The remaining type of variation is substitution of phones. To investigate this variation, we need to find which phones are confusable and in which context. A well known practice is to use a phone recogniser producing a set of phonetic transcripts and compare them with the references [Humphries '97]. By aligning the phone hypotheses with their counterparts from the dictionary, a set of phonetic exchange rules can be obtained as a byproduct. There is a dilemma if we want to use a phone recogniser to find pronunciations for a certain speaking style. If the phone recogniser uses a language model, we have to presume that the phonotactical information (encoded in the language model) is valid. But if we want to compensate for pronunciation rules invalid at hyperarticulated speech, we cannot expect that a phone based language model (derived from the dictionary) is correct. Therefore, the language model has to be excluded from the phone recogniser. As a consequence, we face

[3]To be exact: acyclic, directed graphs

much higher phone error rates. But phone hypotheses with error rates of more than 60% or 70% are not suitable to extract reliable phonetic variations. A different approach is, therefore, proposed to avoid such problems.

In the first step, the confusability of phones will be computed. Given a two-dimensional array of confusion costs, the phone sequences can be expanded to constrained phone networks. A decoding along these networks will then produce a set of phone hypotheses. A set of context dependent substitution rules can then be obtained via a dynamic programming technique.

## Phone Confusions

The Kullback-Leibler divergence can be used for computing the distance between phone models for normal and hyperarticulated speech. The Kullback-Leibler divergence is a criterion based on information theory to measure the additional information mass to code a distribution $f$, given the information of distribution $g$ for a random variable $Y$.

$$I(g;f) = E_Y \log \frac{f(Y)}{g(Y)} = \int_{-\infty}^{\infty} \log \frac{f(y)}{g(y)} f(y) dy$$

Now, this measure tells us something about the similarity of the models and, accordingly, their probability density functions. This allows an indirect measurement of the phone confuseability with respect to hyperarticulated effects. Given sufficient statistics, phone models can be trained and the model similarity can be measured via the Kullback-Leibler divergence. The disadvantage of this approach is that we have a rather indirect method for measuring how two phones compete with each other. If we are interested in reducing the recognition errors caused by wrong pronunciations, a more direct measurement is desired. For example, the decoding engine takes the input data and uses the conditionals $P(x|\lambda)$ to prune away unlikely models.

The approach chosen here finds competing models similar as the decoder Does, with the exception of the segmentation. A forced alignment with the correct phone transcript is used to retrieve the phone boundaries. Given this segmentation, the conditionals $P(x|\lambda)$ can be computed for each model and for each phone occurrence in the data. Finally, frequent phone confusions are extracted from the likelihood matrix, both for normal and hyperarticulated speech.

| vowel | cnt | hypotheses | | | |
|---|---|---|---|---|---|
| AH | 188 | AH 12.2% | UH 11.2% | AE 9.0% | &AH[4] 6.4% |
| AY | 247 | AY 19.8% | IY 10.9% | AE 8.1% | EY 7.7% |
| EH | 665 | IY 19.5% | AE 12.9% | EY 8.1% | UW 7.2% |
| IY | 725 | IY 36.6% | AE 12.0% | EY 11.0% | UW 9.5% |
| OW | 392 | OW 15.3% | UH 12.0% | AE 10.2% | EY 9.4% |

Table 4.4: Ranking of top 4 vowel recognition candidates, normal speech.

| vowel | cnt | hypotheses | | | |
|---|---|---|---|---|---|
| AH | 234 | AE 9.4% | AH 7.3% | IY 7.2% | UH 6.0% |
| AY | 287 | AY 15.7% | AE 12.2% | IY 9.4% | UW 7.9% |
| EH | 760 | IY 20.1% | AE 18.0% | EY 8.2% | EH 8.2% |
| IY | 836 | IY 37.3% | AE 15.1% | EY 10.9% | UW 9.9% |
| OW | 471 | OW 20.2% | EY 10.4% | AE 7.6% | &OW 5.9% |

Table 4.5: Ranking of top 4 vowel recognition candidates, hyperarticulated speech.

Tables 4.4 and 4.5 can be read as follows:

The number of occurrences per set of vowels is given in the second column. The higher occurrences for hyperarticulated speech are due to word repetitions in the error repair mode, as explained in section 3.2. The remaining columns contain how often a vowel was recognised as another phone. For example, 9.4% of hyperarticulated *AH* occurrences were recognised as *AE*. The phone hypotheses are sorted according to their frequency and only the top-4 ranks are displayed. It should be noted that we consider not only vowel confusions here, but also the conditionals are computed for all phones, e.g. vowel-consonant confusions are considered as well.

Interpreting the above tables, the first thing noted is that there are only two top-rank misclassifications for hyperarticulated vowels. On the other hand, the hyperarticulated version of */OW/* seems easier to discern from other phones. Besides this observation, there is not a clear change in vowel confuseability for hyperarticulated speech. A similar picture arises if we analyse consonantal sounds. Higher confusion rates for hyperarticulated speech can be observed for some plosives only. Indeed, a significance test shows no evidence for systematical phone variations.

The results of this analysis indicate that the pronunciation dictionary is *not* the source of our problem- that recognisers fail to recognise hyperarticualted speech. Nevertheless, we want to complete the series of experiments to anwser the question whether hyperarticulated speech requires a specialised pronunciation dictionary.

## Constrained Phoneme Networks

The next step toward a new pronunciation dictionary is to transcribe the training data on a phone level. To that end, the confusion matrices are used to convert the flat dictionary entries to phoneme networks. An example of the resulting graphs is shown in the following figure:

For sake of clarity, this example represents the context independent variant only. For the experiments conducted here, however, the networks are expanded into their context dependent counterparts. The training data is then retranscribed along these phoneme networks and an expanded list of training pronunciations, together with their frequencies, is gathered.

[4]The & symbol indicates an interjection

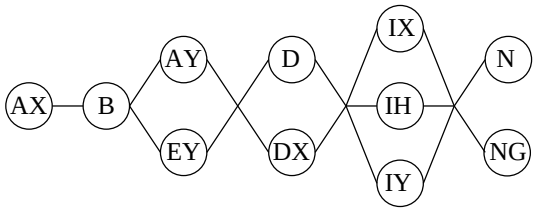


Figure 4.2: Pronunciation graph for *ABIDING*.

## Decision Trees

Before building a new dictionary for recognition experiments, we have faced the problem of predicting pronunciations for unseen test data. Let us first note that it is possible to describe the variations in the training pronunciations as a set of rewrite rules which will be applied to the base-forms. These rewrite rules consist of a substitution pair plus surrounding phonetic context. Limiting the context size will lead to more robust estimations of the phonetic variation. The following list of variation patterns should illustrate the process, whereby the format is given as:
left context / base-phone / right context → replaced phone.

| | | | | |
|---|---|---|---|---|
| EY | / T / | IX | → | DX |
| DH | / AH / | WB[5] | → | AX |
| IX | / NG / | WB | → | N |
| WB | / AE / | N | → | EH |
| N | / S / | M | → | Z |
| T | / IX / | NG | → | AE |
| D | / IX / | L | → | IY |
| WB | / IX / | R | → | AY |

Table 4.6: Patterns of phone variation.

Decision trees are an elegant technique for representing these patterns. Briefly, decision trees are binary trees augmented with questions in each node to select the branch. For the purposes here, the questions pertain to phonetic

[5]The /WB/ symbol indicates a word boundary. Depending on whether /WB/ occurs in the left or right context, a start or end of word is marked.

context. A way to induce *generalisability* is to constrain the phonetic context in the decision tree to phone clusters. These phone clusters can be obtained by a data driven method, or phonetic knowledge may be used to design the groups. The decision trees in these experiments used phone clusters grouped according to place and manner of articulation.

Given the basic components of decision trees, an algorithm for constructing the tree is needed. The approach chosen here is a divisive clustering procedure. This involves several iterations over a list of active nodes, as the following scheme illustrates:

```
root     = Node()
nodes    = NodeLst()
root.addSamples(TrainingData)
nodes.add(root)
while root.complexityCost() < threshold :
    node = nodes.findBest()
    node.split()
    nodes.add(node.leftChild)
    nodes.add(node.rightChild)
    nodes.del(node)
root.saveTree(filename)
```

As the reader may have noticed, an important issue when building a decision tree is still missing: What is the best node to split? In other words, what is the optimisation criterion? Predicting the correct pronunciation variant can be reduced to a classification problem, thus we want to minimise the classification error. Let us assume we have a list $\{(l_i, c, r_i, s_i) : i < N_x\}$ of pronunciation patterns attached to a node $n_x$ for the base-phone $c$. Furthermore, the conditional probabilities $P(s_i|c)$ can be estimated from the training data. The entropy may serve as an optimisation criterion because the negative probabilistic "uncertainty" can be interpreted as the *purity* of a node. However, optimising an entropy criterion does not necessarily translate to minimal classification errors. In [Johnson et al. 2002], the authors report that the Gini-Index has a better correlation to the classification error on a text categorisation task. Buntine and Niblett [Buntine & Niblett '92] found similar classification results for entropy and Gini-Index based tree generators. Their results are based on an experimental evaluation of several optimisation criteria performed on a suite of various artifical classification tasks.

1. Entropy : $E(n_x) = -\sum_i^{N_x} P(s_i|c) * \log(P(s_i|c))$
2. Gini-Index : $G(n_x) = 1 - \sum_i^{N_x} P(s_i|c) * P(s_i|c)$

A split of a node into two subtrees can be scored by measuring the gain of these criteria. If a node $n_x$ is partitioned into two subtrees $x_1$ and $x_2$, an Entropy based splitting score would look like:

$$H(n_x, x_1, x_2) = P(x_1) * E(x_1) + P(x_2) * E(x_2) - P(n_x) * E(n_x)$$

A similar one can be obtained for the Gini-Index. In speech recognition, the entropy criterion is often used for clustering tasks, while the Gini-Index occurs frequently in the data mining community as a criterion for the CART algorithm [Breiman et al. '84].

Pruning is an essential part of the tree construction since it ensures that the tree fit unseen test data. Two thresholds are defined to control the model complexity. The maximum depth limits the tree growing and a minimum sample count is used to avoid unreliable parameter estimations. The pruning has an effect on the leaf's uniqueness. Leafs with $G(n_x) > 0$ will occur depending on the pruning parameter. To apply the tree to pronunciation generation, there are basically two options to deal with this situation. First, dictionary entries including probabilities for all possible substitutions can be generated. This will, however, increase the word confuseability significantly. Therefore, only the pronunciation pattern with the highest probability $P(s_i|c)$ will be selected for further processing.

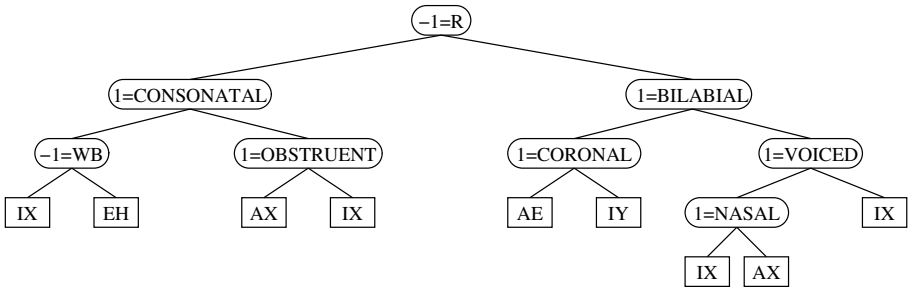


Figure 4.3: Pronunciation decision tree for /IX/.

Figure 4.3 shows the pronunciation decision tree for /IX/. The vowel /IX/ will be substituted by /EH/, /AX/, /IY/, or /AE/ depending on the context.

Questions about right phonetic context are written as "1=..", whereas "-1=..." are used for left context.

## Recognition Experiments

This section describes recognition experiments using the newly trained dictionaries. The basis dictionary already contains some pronunciation variants. The experimental setup is identical to the setup used for the transition probabilities in section 4.1.

| dictionary | variants per base-form |
|---|---|
| baseline | 1.41 |
| full expanded | 3.72 |
| Gini-Index | 1.69 |
| Entropy | 1.65 |

Table 4.7: Dictionary size.

Two optimisation criteria for the decision tree were investigated: Entropy distance and Gini-Index. The trained decision trees resulted in 12% new pronunciations. The fully expanded dictionary was also used as a contrast experiment. The number of pronunciation variants are compared in table 4.7 and the corresponding error rates are summarised in table 4.8.

| dictionary | error rate | |
|---|---|---|
| | normal | hyper |
| baseline | 24.3% | 38.7% |
| full expanded | 25.8% | 40.2% |
| Gini-Index | 24.4% | 38.6% |
| Entropy | 24.4% | 38.8% |

Table 4.8: Pronunciation Modelling.

The decision tree optimised with the Gini-Index gives modest improvements on the hyperarticulated part of the test set (38.7% → 38.6%). This change in error rate is not significant. The error rates hardly changed for

normal and hyper-clear speech. The contrast experiment with the fully expanded dictionary shows a significant increase in errors. This is actually not surprising, since 3.7 variants per base-form on average produce a much higher confuseability of the vocabulary words in the search space.

## 4.3 Separate Acoustic Model Sets

On our path to investigate which knowledge sources in an ASR system are causal for the poor recognition performance at hyperarticulation, we examine in this section the observation models. In chapter 3, we showed that the hyperarticulated speech has significant differences at the likelihood level (table 3.3). So, there is evidence that the observation models do not fit with hyperarticulated speech.

The approach chosen here for reducing the data-model mismatch consists of training separate acoustic models for each speaking mode. As a result, the decoder has to deal with two model sets, one for normal speech and one for hyperarticulated speech. The first question is how to derive mode dependent models and, secondly, how to decide which model is used when. The model selection can be done before decoding or after decoding. The later option requires more computational resources since two decoding runs are necessary, but it has the advantage that the hypotheses can be compared directly.

### Experimental Setup

The SWB corpus was used to train acoustic models for the initial experiments in chapter 3. These acoustic models make use of more than $288,000$ Gaussians defining $10,000$ context dependent HMM states. For the meeting recognition project, acoustic models trained on multiple domains were investigated [Soltau et al. 2002b]. Similar acoustic modeling techniques were used both for the meeting and the SWB system. It turned out, however, that the "meeting" models are better in conjunction with adaptation on the HSC-normal data. This data was originally collected to reduce a possible channel or domain mismatch. MLLR regression classes were estimated on two hours of speech data.

As shown in table 4.9, the SWB models have a lower error rate than the meeting models before adaptation. Nevertheless, the adaptation is more effective for the meeting models, resulting in an error rate of 18.9% for nor-

| Adapt on | SWB models | | Meeting 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 4.9: Comparison of Meeting with SWB models and supervised adaptation (results in word error rate).

mal speech and 29.9% for hyperarticulated speech. These results can be attributed to the fact that the SWB models have about 50% more model parameters to estimate. Experiments confirmed this hypothesis. As a consequence, the adapted meeting models will serve as a baseline for all further experiments.

## Generating Specialised Models

MLLR regression classes were estimated using the training data HSC-normal, HSC-hyper, or both parts. The regression tree is pruned based on the occurrence statistics. The meeting models were then transformed by these trees to the new acoustic models.

| Acoustic Models | error rate | |
|---|---|---|
| | normal | hyper |
| adapted on HSC-normal | 18.9% | 29.9% |
| adapted on HSC-hyper | 18.7% | 25.2% |
| shared models | 18.1% | 26.7% |

Table 4.10: Model specialisation for normal and hyperarticulated speech.

The recognition performance for each of this model sets is shown in table 4.10. The use of hyperarticulated training data gives an error reduction from 29.9% to 25.2%. On the other hand, if only one set of models is required to reduce the computational load, the shared models have an error rate of 26.7% on hyperarticulated speech. This is a significant improvement over the "normal" models, but 1.5% worse than the special models for hyperarticulated speech.

## Model Selection

Focusing now on the question of how to select the right models, we start with two "cheating experiments" to evaluate what would be the maximal improvement we can obtain.

| selection criterion | error rate | |
|---|---|---|
| | normal | hyper |
| database info | 18.1% | 25.2% |
| oracle | 16.8% | 23.1% |

Table 4.11: Model selection using an oracle.

In the first case, it is assumed that all word repetitions are uttered in a hyper-clear speaking mode. That means the dialogue state needs to be given. Secondly, an oracle selecting the models with respect to the error rate is used. It simply selects the output that produces a lower error rate by aligning the hypotheses with the reference. This oracle reduces the error rate to 16.8% for normal speech and 23.1% for hyperarticulated speech and is the best that can be obtained using model selection. The relative error reduction using an oracle is similar for normal speech (18.1% $\rightarrow$ 16.8% = 7.2%) and hyperarticulated speech (25.2% $\rightarrow$ 23.1% = 8.3%). This means that the database information for hyperarticulated speech is as correct as for normal speech. In other words, classifying all word repetitions as being hyperarticulated does not introduce additional errors on this corpus.

The experiments above were "cheating experiments" which used information about the test set. Real model selection cannot use such information, obviously. If real-time operation is not required, decoding runs with both model sets can be performed and the selection is based on the likelihood of the hypotheses. The setup using shared models is the baseline for comparison. As shown in table 4.12, normal speech does no profit from the model selection, while the performance on hyperarticulated speech improves from 26.7% to 24.8%. The disadvantage of this setup is that two decoding runs are required.

A model selection prior to the decoding run would reduce the computational overhead significantly. To that end, pitch and speaking rate were investigated for selecting the models prior to the decoding run.

| acoustic models | error rate | |
|---|---|---|
| | normal | hyper |
| shared models | 18.1% | 26.7% |
| likelihood selected models | 18.0% | 24.8% |

Table 4.12: Specialised models: likelihood selection.

1. Speaking Rate
   The speaking rate was estimated as described in chapter 3. Prior to the decoding run, the *mrate* score is extracted from the audio signal. If the score is lower than a predefined threshold, the utterance is treated as being hyperarticulated. The threshold was optimised on a cross-validation set.

2. Pitch Average
   The average fundamental frequency is used in an analogous manner. The pitch tracker extracts the $F0$ values before decoding. Since the absolute values are not meaningful in the context of hyperarticulated changes, the $F0$ values are compared with their counterparts obtained in the normal speaking mode. This means that the classification is based on the $F0$ differences between the normal and repeated utterances. This approach requires that the dialogue state is given.

3. Pitch Contour
   Despite the average fundamental frequency, the pitch contour can provide a clue for the speaking mode. An example for different pitch contours is shown in figure 4.4. Analogous to the $F0$ average, the dialogue state needs to be given. As we reported in [Soltau & Waibel 2000b], the pitch contour is described for this purpose as a sequence of rising and falling segments. Only the direction of the gradients is considered, but not their absolute values. A change of the speaking mode is assumed when the sequences of rising and falling segments do not match.

The results for the different selection methods are summarised in table 4.13. The average $F0$ values do not contain meaningful information for selecting the appropriate set of acoustic models. The speaking rate based selection performs best and gives an error reduction from 26.7% to 24.9%

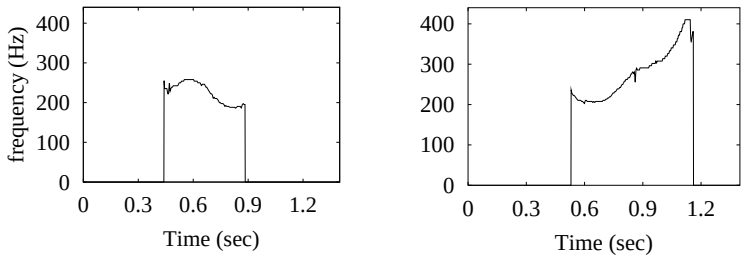


Figure 4.4: Pitch contour for the word *Leonard*, spoken normally (left) and hyperarticulated (right).

for hyperarticulated speech. On the other hand, this gain comes along with a performance degradation for normal speech. Summarised: if the computational load is not an important aspect, the likelihood criterion is the best choice for model selection. Comparing the adapted meeting models with the specialised models, an error reduction from 18.9% to 18.0% for normal speech and from 29.9% to 24.8% for hyperarticulated speech is achieved. This means that model specialisation and selection reduces the error rate at hyperarticulated speech sucessfully.

| acoustic models | error rate | |
|---|---|---|
| | normal | hyper |
| shared models | 18.1% | 26.7% |
| speaking rate selected models | 18.5% | 24.9% |
| pitch average selected models | 18.1% | 26.3% |
| pitch contour selected models | 18.1% | 25.5% |

Table 4.13: Specialised models: speaking rate and pitch based selection.

## 4.4 Hyperarticulation in Context Decision Trees

One important observation in chapter 3 has not been considered so far in the context of model specialisation. The analysis of the phone durations, as well as the formant frequencies, has shown that hyperarticulated changes

occur as phone dependent effects in a temporal and spectral domain. For example, the phone duration is increased by 44% for plosive, but only 16% for vowels. Significant changes of the formants are observed for /AE/ but not for /AH/. This point was not taken into account while generating specialised models, as discussed in the last section. The regression classes were built independently of any phone dependent effects. What is therefore wanted is a data-driven method to decide which models are affected by hyperarticulation and will therefore need special treatment. The fundamental difference is that the complete set of models does not need to be separated into normal and hyperarticulated parts. Only the models which are affected in a hyper-clear speaking mode need a mode dependent parameter. Two questions need to be addressed:

1. Splitting criterion
   Let us first define two labeled training sets $N_p$ and $H_p$ of normal and hyperarticulated speech for a phone $p$. We further denote corresponding cross-validation data as $\tilde{N}_p$ and $\tilde{H}_p$. The question is now whether it is better to share the data $N_p$ and $H_p$ and train one model $m_{nh}$ or to train two models $m_n$ and $m_h$ for each data set. The likelihood for $\tilde{N}_p$ and $\tilde{H}_p$ can serve as a criterion.

$$\begin{aligned} \mathcal{L}_1 &= \log P(\tilde{N}_p|m_{nh}) + \log P(\tilde{H}_p|m_{nh}) \\ \mathcal{L}_2 &= \log P(\tilde{N}_p|m_n) + \log P(\tilde{H}_p|m_h) \end{aligned}$$

   It should be noted that an increase of model parameters will not necessarily increase $\mathcal{L}_1$ or $\mathcal{L}_2$ since the likelihoods are measured on a cross-validation set [Rogina '97]. We can, therefore, split a model into a normal and a hyperarticulated part if $\mathcal{L}_2 > \mathcal{L}_1$.

2. Training procedure
   The splitting into normal and hyperarticulated models can be embedded into the clustering procedure for the polyphone models [Soltau & Waibel 2000a] and [Fügen & Rogina 2000]. The training procedure consists of two steps. In the first step, probability density functions are trained for each phonetic context and speaking mode. In the second step, a set of questions will be evaluated finding the best

split with respect to the likelihood according to the phonetic context or speaking mode. Starting with context and hyper-clear independent root nodes, all possible splits will be scored and children nodes will be generated. Thus, questions about the phonetic context will compete with questions about the hyperarticulated speaking mode. If a phone in a certain context is not affected by hyperarticulation, phonetic questions will probably obtain better scores and an undesired data split into the normal and hyperarticulated parts will not occur. This ensures that exactly these phones will use separate models for normal and hyperarticulated speech, which indeed exhibit differences across the speaking mode.

## Experimental Setup

This approach obviously makes it necessary to train a set of models completely from scratch since the decision trees will change. Furthermore, the training data for normal and hyperarticulated speech need to be balanced. The clustering procedure would otherwise tend to bias phonetic questions. This is one of the drawbacks of this approach. The meeting data can, therefore, not be used in this setup. Instead, a new system was conventionally trained on the HSC database only. Besides the different training data, the same acoustic modeling techniques were applied in both setups. The HSC trained model set will serve as the baseline for the tree generation experiments. The phonetic context size is one phoneme, e.g. the resulting models are generalised tri-phones.

## Results

In a first experiment, the conventionally trained system is compared against the adapted meeting models. The error rate on hyperarticulated speech for the HSC system is as good as for the adapted meeting models. A significant performance degradation, however, occurs for normal speech. Nevertheless, the HSC models provide a good performance for hyperarticulated speech and serve as a baseline.

The experiments in table 4.14 were conducted using only phonetic context questions as usual. In the next experiment, hyperarticulated questions were integrated into the clustering procedure. An excerpt of the tree is shown in figure 4.5. Questions "-1=?" will ask about the left context, "0=?" about

| acoustic models | error rate | |
|---|---|---|
| | normal | hyper |
| adapted meeting models | 18.9% | 29.9% |
| trained on HSC | 23.0% | 29.8% |

Table 4.14: Comparison of adapted meeting models with HSC models.

the centre- phone, and "1=?" about the right context. Left/right branches correspond to no/yes answers. The root node is completely independent of context and speaking mode. All available data were used for training the root. For each question, including hyperarticulation, the node is split into two children and the likelihood on the corresponding cross-validation set is computed. The best question will be used to enlarge the tree. In this example, the first node splits the root according to the word boundary flag. All nodes after the third tree level at the right branch depend on hyperarticulation.

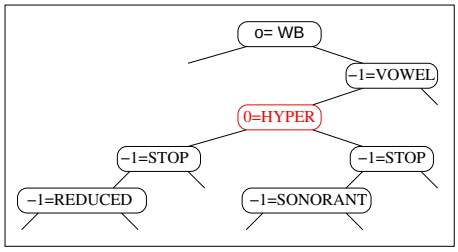


Figure 4.5: Excerpt from the decision tree for /Z/.

| questions | error rate | |
|---|---|---|
| | normal | hyper |
| phonetic context | 23.0% | 29.8% |
| + hyperarticulation | 23.3% | 27.1% |

Table 4.15: Tree generation with hyperarticulated questions.

A comparison between the conventional tree generation and the clustering, including hyperarticulated questions, is given in table 4.15. An error

reduction from 29.8% to 27.1% for hyperarticulated speech is achieved. A minor degradation is, however, observed for normal speech. In this new decision tree, 15% of all nodes do depend on the speaking mode. This confirms also that only certain speech states are affected by hyperarticulation. It should be noted that the number of parameters is the same for both trees.

| phone class | hyperarticulated questions |
|---|---|
| vowels | 3.5% |
| consonants | 20.8% |
| - nasals | 23.8% |
| - plosives | 21.6% |
| - fricatives | 24.6% |
| - approximants | 9.8% |

Table 4.16: Splits relating to manner of articulation.

In the next examination, we analysed which phones are mainly separated into a normal and a hyperarticulated part. To that end, the number of leaves was counted for each base phone and each speaking mode. It seems, that the acoustic space of vowels does not change in an error recovery mode in contrast to consonants. Only 3.5% of the vowel models depend on the hyperarticulated speaking mode, in contrast to more than 20% of the consonants. This result is not surprising if we keep the analysis of phone durations and formant frequencies from chapter 3 in mind. Additionally, the experiments on modeling vowel confusions as pronunciation variants in the previous sections led only to minor improvements (from 38.7% to 38.6%, see table 4.8).

| phone class | hyperarticulated questions |
|---|---|
| bilabial | 8.0% |
| labiodental | 0.0% |
| alveolar | 24.3% |
| retroflex | 0.0% |
| velar | 41.7% |

Table 4.17: Splits relating to place of articulation.

The distribution of hyperarticulation dependencies regarding place of ar-

ticulation is shown in table 4.17. Mainly, alveolar and velar sounds exhibit acoustic changes in a hyper-clear speaking mode. The distribution of hyperarticulated changes fits the observed duration changes: for example, bilabial sounds have a 44% long duration in a hyperarticulated speaking mode, but labiodental sounds show an increase of 17% only (see table 3.6).

## 4.5 Summary

As we have seen in this chapter, there are a couple of issues regarding hyperarticulated phonotactics. Duration modeling is one important aspect for compensating hyperarticulated effects. Significant improvements were achieved by introducing phone *and* speaker dependent transition probability functions. Starting with an error rate of 41.4%, training of phone and speaker dependent transition models led to an improved error rate of 36.5%. These improvements are in agreement with the results of the error analysis in the previous chapter.

Treating hyperarticulation as a dictionary problem did not lead to a major error reduction. The figures 3.3, 3.4, and 3.5 help to explain the situation for vowels. These figures suggest that hyper-clear speech exhibits a drift of formant frequencies. Dictionary learning can only be successful for compensating hyper-clear speech if the hyperarticulated realisation of a phone corresponds to any canonical phone model.

Otherwise, replacements in the pronunciation dictionary cannot reduce the mismatch between data and models. Now, the spectrum changes observed in the above mentioned figures do not support this point. The formant frequencies for the hyperarticulated vowels do not correspond to "standard" vowel values for normal speech. A similar picture can be drawn from the tables 4.4 and 4.5: the phone confusion matrices computed by comparing the likelihoods of the models, given the data, do not show strong evidence that humans simply substitute phones in a hyper-clear speaking mode. All these facts together explain why generating hyperarticulated pronunciation variants does not improve recognition performance.

In summary, only certain phones are affected by hyperarticulation. By extending phonetic context decision trees with dynamic questions about hyperarticulation, we achieved an error reduction of 9% relative. Despite this improvement on hyperarticulated speech, the adapted meeting models provide a better performance for normal speech. Overall, the decision tree using

hyperarticulated questions led to improvements, but did not outperform the "separate model" approach. The best model selection has an error rate of 18.0% on normal speech and 24.8% on hyperarticulated speech. Integrating hyperarticulation into the decision tree led to a performance of 23.3% for normal speech and 27.1% for hyperarticulated speech.

# Chapter 5

# The Articulatory Vector Space

This chapter shows that canonical phoneme models are inadequate for representing hyperarticulated sounds. Articulatory vector spaces provide an alternative framework to pure phone based models. Hyperarticulated effects can be described as changes of articulatory properties. The articulatory vector space allows the definition of an elegant representation of changes in a hyper-clear speaking mode. We introduce the concept of contrastive attributes, which explains hyperarticulation as an inversion of those attributes which discriminate between the spoken and the recognised word. This allows the definition of translation vectors for modeling hyperarticulated changes from a canonical pronunciation and therefore allows the prediction of hyperarticulated effects. The phenomena of hyperarticulation can then be interpreted as a warping of trajectories in an articulatory vector space. This chapter starts with a brief introduction into articulatory phonetics and explains how articulatory features can be used for a hyper-clear speaking mode. It reports on experiments conducted to detect articulatory properties as well as on recognition experiments for hyperarticulated speech. Furthermore, an analysis of translation vectors between true and recognised words confirms the concept of contrastive attributes.

## 5.1 Articulatory Phonetics

The goal of this section is to give a *functional* view of the phonation process, particularly with regard to hyperarticulated speech. In order to understand how humans change the way they produce sounds in an error recovery mode,

it is necessary to address the questions, what are the essential components of the speech production process and how do they work. As this section covers only those topics in articulatory phonetics that are relevant to understanding hyperarticulated phenomena, more detailed information about phonetics is available in [Ladefoged '75] by Peter Ladefoged.

Basically, there are three major processes involved in producing speech sounds. These processes describe the *airstream*, the *phonation*, and the movements in the vocal tract (*oro-nasal* process). Fant's source filter model [Fant '60] interprets these processes as a system of linear, time shift invariant components.

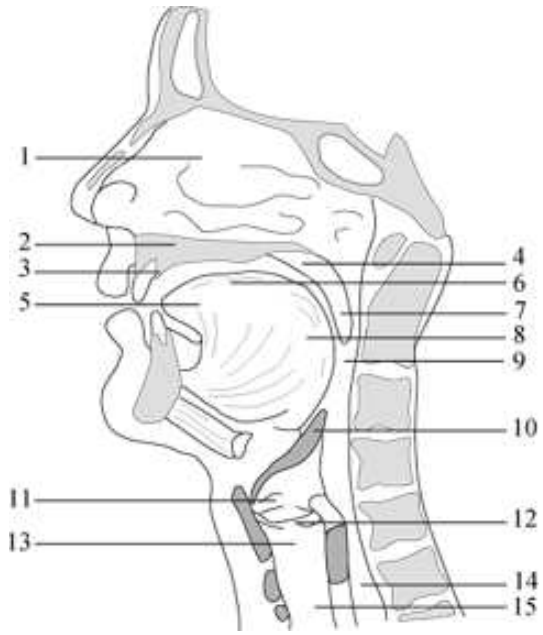


Figure 5.1: Organs of the human speech production : (1) Nasal cavity, (2) Hard palate, (3) Alveolar ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea, from [Lemmetty '99].

1. The Airstream Process
   The airstream 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. However, ingressive sounds occur rather rarely. Besides these pulmonic sounds, a closure of the glottis leads to the so-called *glottal* airstream mechanism. There are *ejective* and *implosive* glottal sounds, depending on whether the air is directly pushed outward or whether the glottis will be lowered. A special sound is the glottal stop produced by the trapping of air by the glottis.

2. The Phonation process
   The phonation process itself is based on the vocal chords. *Voiced* consonants are produced by narrowing the vocal chords. The Bernoulli effect leads to a fast cycle of opening and closing of the glottis. Depending on the length of the vocal chords, the frequency of this process can be in the range of 120-230 Hz. On the other hand, an open glottis leads to *unvoiced* consonants. In that case, air passes without obstruction through the glottis.

3. 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 5.1. The vocal tract consists of three cavities: the *oral* cavity, the *nasal* cavity, and the *pharyngeal* cavity. These components provide a mechanism for producing speech sounds by obstructing air. Several articulators can be moved in order to change the vocal tract characteristic.

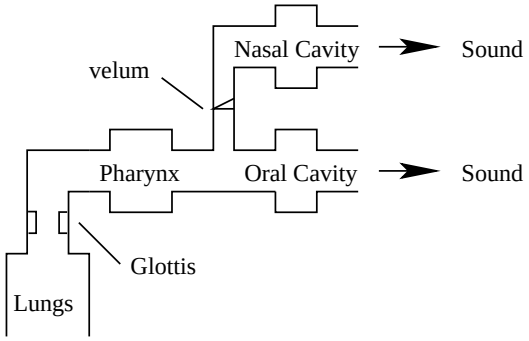


Figure 5.2: Vocal tract as a system of cavities.

The sounds depend on how the air is being modified and on the place of the modifiers. This system results in a classification scheme for consonants that groups sounds according to *place* and *manner* of articulation.

### Place of Articulation

There are several points in the vocal tract where the air stream can be modified. The *articulators* are lips, teeth, tongue, dorsum, soft and hard palate, uvula, and glottis. These articulators are depicted in 5.1. The modification of the air stream involves a pair of articulators defining the place of articulation. This results in the following sound groups:

| place | phones | articulators |
|---|---|---|
| alveolar | /t/ /d/ /n/ | tongue and alveolar ridge |
| bilabial | /p/ /b/ /m/ | lips |
| glottal | /?/ /h/ | glottis |
| labiodental | /f/ /v/ | lips and teeth |
| interdental | /T/ /D/ | teeth |
| retroflex | /r/ | tongue tip and soft palate |

Table 5.1: Consonantal place of articulation.

### Manner of Articulation

The sounds can also be distinguished according to the manner of articulation. The vocal tract allows various ways to modify or obstruct air.

1. Plosives : /p/ /b/ /t/ /d/ /k/ /g/
   Plosive sounds are produced by a complete oral closure. A re-opening of the vocal tract leads to a burst.

2. Nasals : /m/ /n/
   Nasal sounds are also produced by a closure of the vocal tract. However, the velum is in the lower position. The air stream is affected both by the oral and the nasal cavity.

3. Fricatives : /f/ /v/ /s/ /z/
   The vocal tract is constricted but there is not a complete closure. This

results in turbulent air which is then modified by the vocal tract resonators.

4. Approximants : /r/ l/ /j/ /w/
   In contrast to fricatives, the air flow is here rather smooth for approximants and the vocal tract is less constricted than for fricatives.

The degree of constriction is a major factor in describing the manner of articulation. Sounds produced by obstructing the air stream are called *obstruents*. Their counterparts are called *sonorants*.

### Voicing

The place and manner of articulation are not sufficient enough for defining speech sounds. For example, /p/ and /b/ are both bilabial plosives. But the phone /b/ is a voiced sound while /p/ is voiceless. As explained above, the phonation process determines whether a sound will be voiced. The International Phonetic Association (IPA) [International Phonetic Association '99] established an inventory of sounds and provides a classification scheme based on place, manner, and voicing. The following table contains the official IPA phone chart.

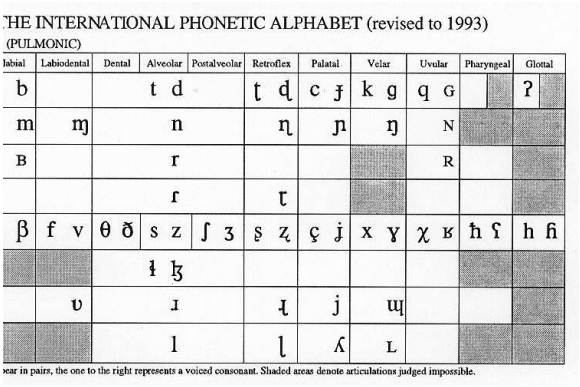


Figure 5.3: pulmonic consonants, [International Phonetic Association '99].

### Vowels

So far, we have discussed consonantal sounds only. There is a similar classification scheme for vowels. Vowels are *voiced* sounds in almost all languages. The Japanese language is one of the languages having voiceless vowels between voiceless consonants. The characteristic features of vowels are produced in the oral cavity. There are basically three parameters used to characterize vowels:

1. Vertical position of the tongue
   There are three possible values for this attribute: low, middle, and high.

2. Horizontal position of the tongue
   The tongue can be in a front, central, or back position.

3. Lips rounding
   Rounded lips are used to produced sounds like /u/ or /o/ while unrounded lips are characteristic for sounds like /a/ or /i/.

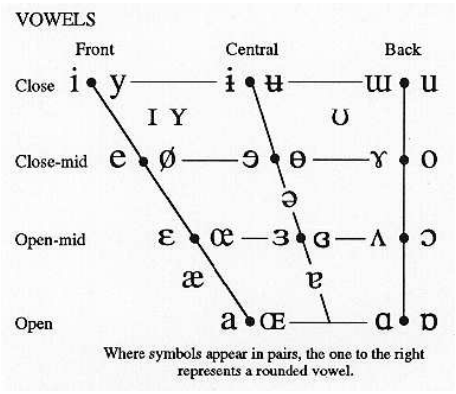


Figure 5.4: vowels, [International Phonetic Association '99].

## 5.2 Articulatory Modelling for ASR - a Review

There are several attempts for using articulatory phonetics in systems for automatic speech recognition (ASR). Ellen Eide [Eide 2001] used articulatory attributes to enhance the front-end of a speech recogniser. She trained a classifier based on Gaussian mixture models for the attributes. The output of these classifiers is then combined with the cepstral observation vector to form the front-end. The extended front-end was used to train new acoustic models. She observed an error reduction of up to 25% on car audio data. Li Deng [Deng '98] developed a framework based on neural networks and the 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. Kirchhoff developed in her thesis [Kirchhoff '99] an approach using articulatory information for robust speech recognition. She used neural networks for classifying 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 [Kirchhoff et al. 2000]. This is a similar approach to that used in the "multi-stream"-community, where different feature streams are used for computing acoustic scores. For example, streams can be used to model cepstral features and their delta's and delta-delta's separately [Rogina & Waibel '94]. The same stream technique can be used to build acoustic models with articulatory attributes. Metze proposed in [Metze & Waibel 2002] articulatory attributes with corresponding anti-attributes to form a flexible stream architecture. His approach achieved an error reduction of 15% on a dictation task, and 7% on a spontaneous scheduling task. The potential use of articulatory attributes for speaker adaptation is explored in [Metze & Waibel 2003]. The approach is based on the selection of speaker dependent attributes. The use of articulatory attributes to compensate for hyperarticulated effects is investigated in [Soltau et al. 2002a].

## 5.3 Hyperarticulation - Warping in an Articulatory Domain

As discussed in the first section, it is possible to obtain a complete description of phones by composites of attributes. More particularly, these attributes can represent multi-value structures such as place and manner of articulation or binary features such as voicing or rounding of lips. Moreover, 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 will obviously induce a correlation between the attributes[1]. On the other hand, switching to binary attributes allows the creation of a unified view of hyperarticulated effects in an articulatory domain.

### Algebraic Representation of Articulatory Attributes

Before we start to highlight the advantages of articulatory attributes in the context of hyperarticulation, it is necessary, or at least desirable, to establish a formalism for representing these units. Although a sort of phonological similarity measure is used in the form of a set of questions to generate the context decision tree, there is basically no inherent structure inside the set of phones used for ASR. Therefore, the algebraic term *set* is the adequate name for describing phones.

Coming now to the articulatory attributes, let us first introduce an abbreviation for the composites of articulatory attributes : CAA. A CAA can be seen as an element of a vector space $V$ spanned over the articulatory attributes. The neutral element is a vector representing the absence of all attributes - corresponding to silence. A natural choice of the additive operation would be the binary OR relation. This choice would, however, conflict with one of the axioms for Abelian groups : For each element $x$ there must exist an element $y$ with $x \oplus y = e$, whereby $e$ denotes the neutral element. This axiom cannot be fulfilled by a binary OR, whereas a concatenation of an addition and modulo function defines a valid associative

[1] At this point of time it should be noted that articulatory attributes will not be used to enhance the front-end. In that case, this transformation step would conflict with the diagonal assumption for covariance modeling, even for semi-tied full covariances or similar approaches.

operation for groups. More interesting is the scalar operation $\otimes$. The function $\otimes$ maps a CAA $v$ using a scalar $\alpha$ to a new CAA $v'$. This provides a framework for activating or deactivating certain attributes for a CAA. As we will see, a sequence of such scalar operations can be used for describing hyperarticulated effects. Having a more or less descriptive definition of the structure, a more formal definition follows now: An ordered set $\mathcal{A}$ of attributes

$$\mathcal{A} = \{\alpha_1, .., \alpha_n\}$$

allows to define a scalar field $\mathbb{K}$

$$\mathbb{K} = 2^{\mathcal{A}} = \{(x_1, .., x_n) | x_i \in Z_2\}$$

and finally a vector space defined by an Abelian group $(V, \oplus)$ over $\mathbb{K}$ together with a scalar operation $\otimes$ :

$$V = \mathbb{K}$$

$$\oplus : V \times V \rightarrow V$$
$$x \oplus y = (x_1 + y_1, .., x_n + y_n)$$

$$\otimes : \mathbb{K} \times V \rightarrow V$$
$$\alpha \otimes x = (\alpha_1 \wedge x_1, .., \alpha_n \wedge x_n)$$

It should be noted that we used the fact that each field $\mathbb{K}$ itself can be extended to a vector space $V$. Therefore, the scalar operation $\otimes$ works in the same domain as the vector operation $\oplus$. Remembering that $\oplus$ was defined via the $+$ operation in $Z_2$, it is easy to show that the tuple $(\mathbb{K}, V, \oplus, \otimes)$ satisfies the definition of a vector space. The reader may ask why we did not choose the vector space as $V' = 2^{\mathcal{A}}$ over a field $\mathbb{K}' = \mathcal{A}$ which would be more common. The disadvantage of such a definition is the scalar operation, or more exactly the domain $\mathbb{K}' \times V'$. This "trick" of choosing $V = \mathbb{K}$ allows both $\oplus$ and $\otimes$ to operate in the same domain. Moreover, the neutral element according to the $\oplus$ operation can be interpreted as silence.

## Basis Elements

The neutral element of V with respect to $\oplus$ is denoted by $e = (0, .., 0)$. The inverse $\bar{x}$ of an element $x$ has the property of $x \oplus \bar{x} = e$. This requires setting $\bar{x} = x$. It is straightforward to show that $x \oplus x = e$ and additionally $e \oplus x = x$. The neutral element with respect to $\otimes$ is set as $E = (1, .., 1)$, therefore $E \otimes x = x$.

A family $B = (b_i)_{i \in I}$ of vectors $b_i \in V$ forms a basis of $V$, if each vector $v \in V$ can be represented as a linear combination of $b_i$ with respect to the operations $\oplus$ and $\otimes$. Choosing $(b_i)_j = \delta_{ij}$ provides a basis enabling the creation of all elements in $V$, whereby $\delta$ denotes the Kronecker operator.

The basis elements are important vectors for describing hyperarticulated effects. These vectors can be used to represent a flipping of articulatory attributes. The operation $b_i \oplus x$ will exactly invert the attribute $a_i$ of the vector $x$ thanks to the modulo property of $\oplus$. Having a CAA $v$ representing a canonical phone and $v'$ produced in a hyperarticulated speaking mode, it is possible to describe the changes between $v$ and $v'$ as a sequence of $\oplus$ operations using the basis elements $b_i$.

## Metric

A metric of the articulatory vector space defines a similarity score. If all dimensions are treated as equivalent, an appropriate definition is given as follows:

$$|v| = \sum_i v_i$$

The distance of two vectors is, therefore, the sum of required basis elements for moving from one vector to another.

## Correspondence between Phones and Articulatory Attributes

What is still missing is the discussion of the relationship between phones and CAAs. As discussed earlier, the inherent structure of phones and CAAs is different. Thus, a function mapping CAAs to phones will not be able to conserve the structure. In mathematical language, this function can not be considered to be a homomorphism. Despite the structure information,

there exists a CAA for each phone but not every CAA has a corresponding phone. What we can define is a partial function $f : P \rightarrow V$ to map phones to CAA's. This function $f$ is injective but not surjective:

$$\forall u, v \in P : f(u) = f(v) \Rightarrow u = v$$

$$f(P) \neq V$$

The inverse function $f^{-1} : V \rightarrow P$ is not a partial function: the domain of $f^{-1}$ is constrained on $f(P)$, whereby $f(P) \subset V$. This shows already that the vector space $V$ provides a *richer* language for describing acoustic events compared to the set $P$.

After the development of this mathematical formalism to describe articulatory attributes, it is time for an illustration of the above definitions using an example. The word *doubts* would be represented in a $P$ domain as the following sequence : /D/ /AW/[2] /T/ /S/. A sufficient[3] set of articulatory attributes would cover the following elements:

$$\begin{aligned}
\mathcal{A}_{place} &= \{alveolar, bilabial, interdental\} \\
\mathcal{A}_{manner} &= \{plosive, fricative\} \\
\mathcal{A}_{vowel} &= \{front, round\} \\
\mathcal{A}_{global} &= \{consonantal, voiced\} \\
\mathcal{A} &= \mathcal{A}_{place} \cup \mathcal{A}_{manner} \cup \mathcal{A}_{vowel} \cup \mathcal{A}_{global}
\end{aligned}$$

The field $\mathbb{K}$ consists of all possible combinations of elements in $\mathcal{A}$. The word *doubts* would, therefore, be represented as a vector sequence : $v_1 v_2 v_3 v_4$, whereby the following definitions are used:

$$\begin{aligned}
v_1 &= (1, 0, 0, 1, 0, \cdot, \cdot, 1, 1) \\
v_2 &= (\cdot, \cdot, \cdot, \cdot, \cdot, 1, 1, 0, 1) \\
v_3 &= (0, 0, 1, 1, 0, \cdot, \cdot, 1, 0)
\end{aligned}$$

[2]The unit /AW/ denotes a diphthong describing a gliding vowel sound normally represented by two adjacent vowels.

[3]sufficient in the context of describing the word *doubts*.

$$v_4 = (1, 0, 0, 0, 1, \cdot, \cdot, 1, 0)$$

The dimensions correspond to the ordered list of attributes $\mathcal{A}$, e.g. the first dimension contains information about *alveolar*. For example, the vector $v_1$ corresponds to a sound with the activated attributes *alveolar*, *plosive*, *voiced*, and *consonantal*. Deactivated attributes are indicated by 0, and a dot tag is used for irrelevant attributes. This example demonstrates some of the advantages of representing acoustic units in the vector space $V$. The transition from one vector to the next vector is not an abrupt change, but some dimensions are not affected. For example, the *consonantal* attribute does not change from $v_3$ to $v_4$. In the $P$ domain, the same transition would be represented as a change from one element to a completely different element.

## Disambiguating Errors : Contrastive Attributes

As we have seen in chapter 3, hyperarticulation is not a global effect. For example, the influence of hyperarticulation depends on the phone identity. It is desirable to analyse these effects with a finer granularity in order to understand the underlying principles of hyperarticulated speech.

The formalism developed in the previous section allows us now to model these effects with a much finer granularity. This section will introduce the idea of **contrastive attributes** which are a key concept for describing changes occurring while disambiguating recognition errors. A contrastive attribute is an attribute in the context of a word error which can be used to discriminate between the true and the recognised token. In a hyperarticulated speaking mode, such a contrastive attribute can be inverted to stress the mis-recognised part of the word. The following example should illustrate this process:

Again, we have the word *doubts*, but we add a silence unit at the end: /D/ /AW/ /T/ /S/ /SIL/. Let us now suppose that the word *doubt* was recognised, e.g. the recognised phone sequence is /D/ /AW/ /T/ /SIL/. In the vector space $V$, we have the vector sequences *doubts*= $v_1v_2v_3v_4v_5$ and *doubt*= $w_1w_2w_3w_4$.

Let us now perform an alignment between both vector sequences. The first part of the sequences are identical. The interesting part of this example starts at $v_4$ and $w_3$ respectively. Keeping in mind that $v$ represents the observation and $w$ the hypothesis, the observable variables belonging to $v_3$

| | place | | | manner | | vowel | | global | |
|---|---|---|---|---|---|---|---|---|---|
| | alv | vel | int | plo | fri | fro | rnd | con | voi |
| $doubts = v_1v_2v_3v_4v_5$ | | | | | | | | | |
| $v_1$ | 1 | 0 | 0 | 1 | 0 | · | · | 1 | 1 |
| $v_2$ | · | · | · | · | · | 1 | 1 | 0 | 1 |
| $v_3$ | 0 | 0 | 1 | 1 | 0 | · | · | 1 | 0 |
| $v_4$ | 1 | 0 | 0 | 0 | 1 | · | · | 1 | 0 |
| $v_5$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $doubt = w_1w_2w_3w_4$ | | | | | | | | | |
| $w_1$ | 1 | 0 | 0 | 1 | 0 | · | · | 1 | 1 |
| $w_2$ | · | · | · | · | · | 1 | 1 | 0 | 1 |
| $w_3$ | 0 | 0 | 1 | 1 | 0 | · | · | 1 | 0 |
| $w_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5.2: Contrastive Attributes : *doubts* vs. *doubt*.

and $v_4$ have to be mapped to the CAA $w_3$. This is actually a little bit oversimplified, since segmentation issues are ignored so far. The vectors $v_3$ and $w_3$ are identical. The question that needs to be addressed is what would be a reaction to disambiguate $v_4$ and $w_3$. The formalism developed so far allows a characterisation of the changes between $v_4$ and $w_3$ as follows:

$$v_4 = w_3 \oplus b_1 \oplus b_3 \oplus b_4 \oplus b_5$$

It should be noted that the basis vectors $b_i$ do not represent the presence of articulatory attributes themselfes in context of $\oplus$ operations, but the **inversion** of attributes with respect to a certain CAA. Taking $w_3$ as the recognised token and $v_4$ as the reference token, a hyperarticulated effect can be modeled by a translation vector *hyper* as follows:

$$\begin{aligned} w_3 &= v_4 \oplus hyper \\ hyper &= b_1 \oplus b_3 \oplus b_4 \oplus b_5 \end{aligned}$$

We used here the fact that the inverse element is identical to the element itself. The vector *hyper* can only be interpreted in the context of the "starting point" $w_3$. Decoding the vector components leads to the following in the articulatory domain:

1. deactivate *alveolar*
2. activate *interdental*
3. deactivate *plosive*
4. activate *fricative*

We can now *predict* what kind of changes will occur during hyperarticulation. In order to correct the mis-recognised word *doubt*, a hyperarticulated variant of *doubts* will exhibit *activated* attributes for interdental and fricative. On the other hand, attributes for alveolar, plosive, and voiced will be deactivated. To demonstrate that these predications will also actually occur in real utterances, we will anticipate some results from section 5.5.

Let an utterance $u$ be represented as a sequence of observable feature vectors $o_1 ... o_t$, whereby $t$ denotes the length of the utterance in terms of number of frames. The probability density functions for $P(o_t|a)$ are modeled by mixtures of Gaussian densities. The pdf's 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. The conditionals are used to define a distance function:

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

Figure 5.5 shows two curves. The solid line represents the word *doubts* in a normal speaking mode. In a hyperarticulated speaking mode, the same word *doubts* results in the $\Delta(o_t, a)$-curve shown by the dashed line. Both words were uttered by the same speaker. The hyperarticulated variant arose as a reaction resolving the recognition error *doubts* vs. *doubt* in the framework of the dialogue system described in section 3.2.

The input features $o_t$ were computed with a front-end consisting of a filterbank analysis, inverse cosine transform, cepstral liftering, channel and speaker normalisation, linear discriminant analysis, and semi-tied full covariances. The models are trained via the Baum-Welch re-estimation procedure. An incremental growing of Gaussians approach was applied as well. The full SWB corpus was used to train the models. It should be noted that there is only one set of models: the GMMs are independent of the speaking mode.

Discussing now the figure 5.5, it can be seen that both curves are quite similar in the first half of the figure. The main changes occur actually in

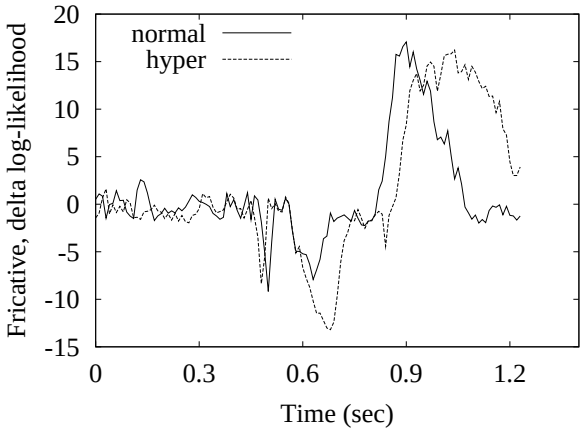


Figure 5.5: $\Delta(o_t, a)$ for attribute *Fricative* while pronouncing *doubts*, normally and hyperarticulated.

the range of 0.9 to 1.2 seconds. The $\Delta(o_t, a)$-scores are much higher for the hyperarticulated word. In other words, the likelihood for being a fricative is increased in this area. This observation agrees perfectly with the theory of contrastive attributes in the vector space $V$. The concept of contrastive attributes led to the prediction that some attributes will be activated and deactivated in order to resolve recognition errors. Representing *doubts* and *doubt* as sequences of CAA's in the vector space $V$ (see table 5.2) resulted in the prediction that the fricative attribute will be activated in a hyper-clear speaking mode. On a phone level, this change can be interpreted as emphasising the missing /S/ sound.

Continuing with a second example, the figure 5.6 shows the $\Delta(o_t, a)$-scores for the plosive attribute. The data for this figure were extracted in the same way as described for the previous figure. The context is the same word confusion *doubts* vs. *doubt* and the utterances are the same as well. The dashed curve represents the hyperarticulated word, while the solid line shows the data obtained in a normal speaking mode. Similar to the first example, the changes occur in the last third. The $\Delta(o_t, a)$-scores in the hyper-clear speaking mode are now much smaller than for normal speech. That means that the likelihood for being a plosive attribute is decreased. This observation is consistent with the predictions. The *hyper* vector describing the changes between $w_3$ and $v_4$ did predict a deactivated plosive attribute. On a phone

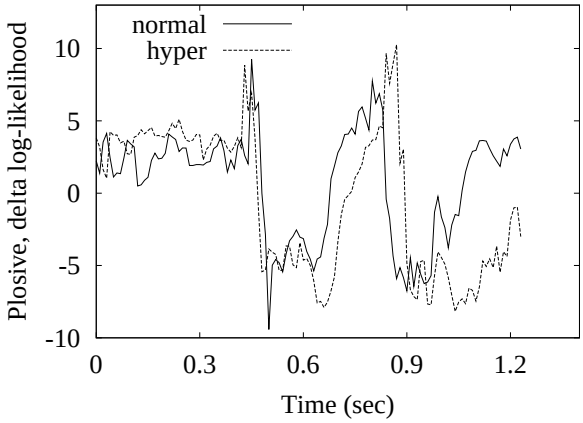


Figure 5.6: $\Delta(o_t, a)$ for attribute *Plosive* while pronouncing *doubts*, normally and hyperarticulated.

level, this change can be interpreted as de-emphasising the /T/ sound. The /T/ sound was actually not wrong, but the /T/ was not the final phone in *doubts*. To indicate there is another phone after /T/, the plosive attribute will be deactivated.

In summary, an algebraic representation of articulatory attributes was presented in this section. The phenomenon of hyperarticulation can be described as a *warping* in an articulatory vector space. The concept of *contrastive attributes* leads to predictions regarding which attributes will be activated or deactivated in order to react to recognition errors. Examples of word confusions reinforce the concept of contrastive attributes.

## 5.4 Statistical Modeling of Acoustic Events

So far, we indicated *why* articulatory attributes could provide a better framework for modeling hyperarticulated speech than pure phone based models. The next point is to discuss *how* articulatory attributes can contribute to a better recognition performance. The question that arises now is, therefore, how we can find a way from CAA's to observable features. There are several requirements: On a very abstract level, the models should capture exactly those features that are relevant for the problem. Task invariant features

should not reach the model level. A further principle for designing information systems is that similar information should be processed in a similar way [Vapnik '98]. Additionally, there must exist efficient training methods for estimating model parameters, and we should not overlook the fact that an efficient decoding algorithm is needed for searching for the best hypothesis with respect to the models.

## Temporal Structure

The temporal structure of a word or a whole utterance can be considered as a trajectory in the vector space $V$. There are several ways for describing such a multi-dimensional trajectory, such as:

1. polynomial : $p(x) = \sum_i \alpha_i x^i$
2. recursion : $Z(k+1) = \Phi Z(k) + U + W(k)$
3. sequence of sampling points

The first thing noted is, if we want to model words using CAA's, the temporal structure does not change compared to a traditional phone based approach. As a consequence, if a word is traditionally modeled as a sequence of phones $p_1...p_n$, then a corresponding representation in the vector space $V$ could consist of $v_1...v_n$. These vectors can be interpreted as *data points* describing a trajectory in the vector space $V$. This would lead us to option three. This approach is quite related to the polynomial proposal. Whether a polynomial is represented as a number of coefficients or as a sequence of sampling points is only a technical question, but it does not change the modeling power.

Li Deng proposed in [Deng '98] a state-space model which is parameterised as a recursive state equation. This concept features a great flexibility and offers an alternative way for representing trajectories. But as mentioned before, there are several requirements for the models. One of them is the need of efficient training and decoding algorithms. To estimate the parameters for the recursive equation, an extended Kalman filter approach was chosen in [Deng '98]. This led to different optimisation criteria for the temporal structure and the emission probabilities. Secondly, plugging in these models into a viterbi decoder will create a series of questions.

Recapitulating the experiments on duration and pronunciation modeling in chapter 4, there is no indication that segments are inserted or deleted in a hyper-clear speaking mode. Hyper-clear speech exhibited longer segments, but the number of segments did not change. This suggests that it is valid to represent the temporal structure of hyperarticulated speech as a sequence of sampling points. Therefore, the temporal structure will be represented as a linear sequence of vectors and the sequence length corresponds to the number of phones for a given word. Thus, the example word *doubts* would look like this:

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ 1 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ 1 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \cdot \\ \cdot \\ 1 \\ 1 \end{pmatrix}$$

**D** **AW** **T** **S**

## Emission Probabilities

The remaining problem is to find a model which computes conditional probabilities $P(o_t|\lambda, v)$ for elements $v$ of the vector space $V$. The underlying model parameters are denoted by $\lambda$. The vector representation of the CAA's suggests separating the conditionals accordingly. The emission probabilities will, therefore, be computed using two levels of conditionals:

1. conditionals $P(o_t|\gamma, a)$ for articulatory attributes $a$
2. conditionals $P(o_t|\lambda, v)$ for $v \in V$

The advantage of this approach is the introduction of parameter sharing across the vectors $v$. The model parameters $\gamma$ for an attribute $\alpha$ will be shared between those vectors relying on the same attribute $\alpha$. The next section will explain in detail how the conditionals $P(o_t|\lambda, a)$ can be estimated using conventional training data. The interesting problem is how to obtain

$P(o_t|\lambda, v)$ based on $P(o_t|\gamma, a)$. Assuming conditional independence, one way to define the probability functions would be:

$$P(o_t|\lambda, v) = \prod_i P(o_t|\gamma_i, v[i])$$

Now, some practical aspects will prevent us from using this definition as it is. In fact, a weighting factor $w_i$ may be introduced to stress certain dimensions. Going to a log-domain to fit the dynamic range, we can define a *score-function* $g$ instead:

$$g(o_t|v) = \sum_i w_i log P(o_t|\gamma_i, v[i])$$

It is obvious that introducing weighting factors will manipulate the probability mass:

$$\int \sum_i P(x|\gamma_i, v[i])^{w_i} dx \neq 1$$

Introducing constraints, such as $\sum_i {w_i}^K = L$ with constants $K$ and $L$ [Hernando '97] will not solve that problem. In fact, the function $g(o_t|v)$ is not a probability density function (pdf) in the log domain. There are two components in a speech recogniser where this non-pdf 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 of the optimisation criterion, the decoder searches for the word sequence with the best score.

From a training point of view, the parameter $\gamma_i$ can be estimated by optimising the ML criterion, since the conditionals $P(o_t|\gamma_i, v[i])$ are valid pdf's. On the other hand, the weighting factors $w_i$ cannot be estimated by maximising the training likelihood. After a few transformation steps, maximising the Kullback-Leibler form would be equal to maximising a sum as:

$$f(w) = \sum_i w_i * k_i$$

with some dimension dependent terms $k_i$. These terms contain the likelihoods for $P(o|\gamma_i, v[i])$, whereby

$$P(o|\gamma_i, v[i]) \leq P(o|\gamma_j, v[j]) \Rightarrow k_i \leq k_j.$$

It is trivial to show that solving that problem would end up in setting $w_i = 1$ for $i$ with the highest likelihood. A second problem will arise, if state dependent weights are used. Since the probability mass is not equal to 1.0 anymore, the acoustic scores of different states cannot be compared. The decoding engine would, therefore, not be able to find the best word sequence.

Since there is not a maximum likelihood solution available, determining the weighting factors consists of a grid search minimising the error rate on a cross-validation set.

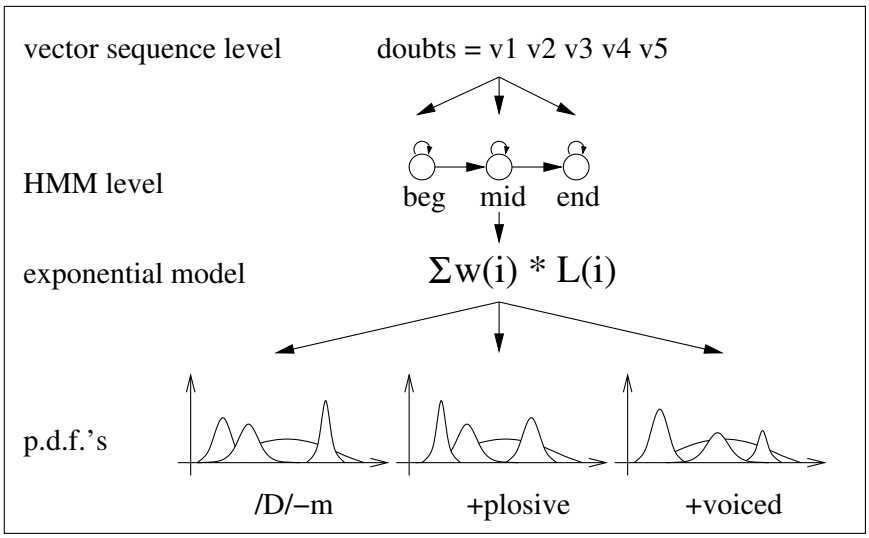


Figure 5.7: Acoustic models for vector elements (example *doubts*).

The overall architecture for computing acoustic scores is depicted in figure 5.7. As shown in this graph, the phones are not completely replaced by CAA's. For example, to compute the acoustic score for the middle state of /D/, the conditionals for /D/-m, +plosive, and +voiced will be computed. The traditional set of context dependent density functions for phonemes remain in this structure. Therefore, the stream weights balance the phoneme models and the articulatory attribute models. This structure allows for plugging in articulatory models in existing traditional phoneme based models.

The acoustic score computation relies on multiple pdf's which are combined on the log-likelihood level $L(i)$.

What is not shown in this graph, but in fact is used, are anti-models. The absence of an attribute needs to be modeled, since the vector space representation is based on activated *and* deactivated attributes. For example, table 5.2 contains deactivated attributes. Suppose an acoustic score for a plosive needs to be computed. Then the probability density functions for all dimensions of the vector space must be evaluated. For any non-plosive attribute, the corresponding pdf must describe the absence of this attribute. Therefore, for each attribute, an anti-model is trained on all data not belonging to this attribute[4]. This allows us to describe the presence and absence of attributes using corresponding probability density functions. That means if there is a set $\mathcal{A}_{pos} = \{a_1...a_n\}$ of attributes, a second set $\mathcal{A}_{neg} = \{\bar{a}_1...\bar{a}_n\}$ is used as well. The field $\mathbb{K}$ is therefore constructed as $\mathbb{K} = 2^{\mathcal{A}_{pos} \cup \mathcal{A}_{neg}}$.

## 5.5 Detection of Articulatory Properties

The overall picture of the system structure is now introduced. The next step is clarifying the details of how to obtain the conditionals $P(o_t|\lambda, v[i])$. There are basically three issues: the input feature space, labeled training data, and classifier topology.

### Feature Space

As a general note to avoid confusion about the terminology, the term *vector* here occurs in different contexts, since there are two vector spaces. The articulatory vector space is denoted by $V$ and is used on the model building level. From a preprocessing point of view, the input features $o_t$ are vectors in the vector space as defined by the front-end.

Given the raw audio data, the input features $o_t$ are transformed in order to eliminate problem invariant information. The essential point is to use the same front-end as for the phone models. As depicted in figure 5.7, the 0th dimension contains the acoustic scores from the phone models. Due to the drift of the total probability masses, it is crucial to ensure that the acoustic scores are comparable across the coefficients. Variance normalisation techniques have, for instance, a high impact on the average acoustic scores

[4]This small detail will significantly increase the training time (sic).

and, therefore, on the probability mass, since the density family is typically based on diagonal covariances[5].

## Training Data

There are several speech corpora available which come together with word transcripts. What is needed to train the conditionals $P(o_t|\lambda, v[i])$, are transcripts on an articulatory attribute level. For each feature vector $o_t$ the corresponding set of attributes needs to be determined. There are multiple assignments possible. For example, a feature vector $o_t$ can be assigned to the plosive class, the voiced class, and the non-vowel class at the same time. There are basically two ways for addressing the problem of training data. The first, and by far more expensive way, makes use of X-ray images [Thimm & Luettin '99]. This allows us to localise the positions of several articulators, such as tongue or jaw. An alternative way is *converting* the word transcripts. Using the function $f : P \rightarrow V$ as defined in the previous section, labeled training data can be obtained as follows:

1. phone alignments
   The phone alignments, or more exactly state alignments, can be computed via the viterbi algorithm using the word transcript and a set of acoustic models. Alternatively, a forward-backward algorithm can be used to generate a list of phones with their posteriori's for each frame.

2. map phones to articulatory attributes
   The phone alignments are then converted to a set of alignments for each articulatory attribute.

The second approach has the advantage that much more training data are available since any speech corpus can be used, but it requires that the mapping function is appropriately defined. Additionally, asynchronous changes of attributes are completely ignored. This is, however, also ignored by the vector representation itself. Thus, the decoding engine processes articulatory attributes in a synchronous way anyway, independently of whether or not the models were trained with asynchronous data.

[5]As the reader may have noticed, this section discusses a few engineering questions which are theoretically not necessary to consider here. The goal of this thesis is, however, also to show that error reductions are achievable. To that end, some practical aspects need to be addressed as well.

## Density family

Gaussian mixture densities provide a well known instrument to model the conditionals $P(o|\gamma,\alpha)$. The same "similar acoustic score"-argument leads to diagonal covariances. The parameterisation is therefore:

$$P(o|\gamma,\alpha) = \sum_i \lambda_i \frac{e^{-\frac{1}{2}(o-\mu)^{-1} diag(\Sigma)^{-1}(o-\mu)}}{\sqrt{(2\pi)^n det(diag(\Sigma))}}$$

## Experimental Setup

The front-end is identical to the setup described in section 3.3. The dimension of the feature vectors is 42. The density functions have a variable number of Gaussians due to the "merge&split"-training. The maximum number of components was set to 48. Three training corpora were investigated. The SWB corpus contains more than 280 hours of conversational telephony speech. The second corpus consists of the first part of the recordings collected with the simulated dialogue system. This database serves as a contrast experiment for the hyperarticulated training data since the set of training speakers are identical. Table 3.1 contains the details for the database for normal and hyperarticulated speech. The test set is the same as used in the previous experiments.

The set of articulatory attributes consists of plosive, nasal, fricative, lateral, approximant, bilabial, labial, labiodental, alveolar, velar, glottal, consonant, voiced, vowel. The likelihood is computed using the corresponding models and anti-models for each frame. The performance is measured as the binary classification accuracy averaged over the number of frames.

## Results

The results for the detection experiments are split into three tables 5.3, 5.4, and 5.5. The experimental setup allows comparisons of the performance across attributes, speaking style, and training corpus.

| attribute | SWB corpus | | HSC-normal | | HSC-hyper | |
|---|---|---|---|---|---|---|
| | normal | hyper | normal | hyper | normal | hyper |
| 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% |

Table 5.3: Detection accuracy for manner of articulation attributes.

| attribute | SWB corpus | | HSC-normal | | HSC-hyper | |
|---|---|---|---|---|---|---|
| | normal | hyper | normal | hyper | normal | hyper |
| 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% |

Table 5.4: Detection accuracy for place of articulation attributes.

| attribute | SWB corpus | | HSC-normal | | HSC-hyper | |
|---|---|---|---|---|---|---|
| | normal | hyper | normal | hyper | normal | hyper |
| voiced | 96% | 96% | 92% | 92% | 86% | 83% |
| consonant | 96% | 93% | 87% | 83% | 88% | 85% |
| all | 85% | 81% | 86% | 81% | 85% | 83% |

Table 5.5: Detection accuracy for global attributes.

## Discussion

### Differences between Attributes

The most adequate comparison between the classification performance of attributes can be done by analysing the fourth column. In that case, we have matched training and test conditions. The models were trained using the normal portion of the HSC training set and evaluated using the normal portion of the HSC test set. The average classification accuracy over all attributes is 86% (table 5.5). If no prior information is used, the performance by chance would be 50%. The statistical models are able to detect articulatory attributes with an acceptable accuracy. The detection performance for manner of articulation varied between 88% for approximants and 93% for fricatives and nasals. The classification performance for place of articulation is more inexact according to table 5.4. It should be noted that the results are based on a *binary* classification. It does not matter, therefore, how many attributes belong to place or manner or articulation.

### Differences between Speaking Modes

The classification performance can be analysed across the speaking modes by comparing the fourth with the fifth column. The classification accuracy is 5% worse on hyperarticulated speech over all attributes. The impact of hyperarticulation on the detection accuracy is more or less equal for all attributes.

### Differences Between Training Corpora

The first thing noted is that the detection accuracy for normal speech is Independent of the training corpus. The models trained of SWB have 85% on average, training with HSC-normal give 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 and the seventh columns, it can be seen that hyperarticulated training data improves the performance from 81% to 83%. In particular, velar and glottal sounds profit, from these data. On the other hand, the classification whether a sound is voiced or not becomes significantly worse.

## 5.6 Speech Recognition with Vector Models

In this section, the potential of vector models for reducing recognition errors for hyperarticulated speech will be examined. The acoustic score computation for the vector models was already discussed in section 5.4. Given the "acoustic score computer", an efficient decoding engine is needed to search for the string with the best score. As mentioned before, it is not necessary that these scores are real probabilities. The IBIS decoder [Soltau et al. 2001a] is a viterbi decoder based on the concept of linguistic polymorphism. The search network is constructed in a way that isomorphic subgraphs are eliminated. The vector models can just be plugged in the IBIS decoder and the corresponding acoustic scores will be used for the search process.

### Experimental Setup

Two questions need to be addressed to define the experimental setup. The first question is which phone models should be used to serve as a baseline. Secondly, which set of attributes should be used to define the vector space.

#### Phone Models

The experimental setup for model separation in chapter 4 used adapted meeting models as a starting point. A comparison between the SWB and the meeting models has shown (see table 4.9) that the SWB models have lower error rates than the meeting models before adaptation. After adaptation, however, the meeting models give significantly better results. The adaptation is more effective for the meeting models, resulting in an error rate of 18.9% for normal speech and 29.9% for hyperarticulated speech. These results can be attributed to the fact that the SWB models have about 50% more model parameters to estimate. Consequently, the adapted meeting models will be used to set a baseline for validating the vector model concept since they provide a "harder" baseline.

#### Vector Space Basis

The vector space is partitioned into four sub-spaces. For each basis vector, a model and an anti-model is trained. For the full space (manner+place+vowel+global), the space is spanned by 19 basis elements as

| Space | Basis |
|---|---|
| manner | plosive, fricative, lateral, approximant |
| place | alveolar, bilabial, glottal, labiodental, interdental, retroflex |
| vowel | high, mid, low, front, central, back, round |
| global | voiced, consonantal |

Table 5.6: Basis Elements.

shown in table 5.6. The total number of Gaussians is for that case 1216. The number of additional parameters needed for the vector models is, therefore, only a small fraction compared to the phone models.

## Results

A separate system was built for each of this sub-spaces in a first step investigating the capabilities of each attribute group. The baseline is the phone based model set. The full vector space uses all attributes.

| acoustic models | Speaking Style | |
|---|---|---|
| | normal | hyper |
| phone based models | 18.9% | 29.9% |
| manner based vector models | 17.3% | 22.2% |
| place based vector models | 17.5% | 22.3% |
| vowel based vector models | 17.4% | 22.4% |
| global based vector models | 18.2% | 23.2% |
| full vector space | 17.8% | 21.5% |

Table 5.7: Recognition experiments with vector models (results in word error rates).

## Discussion

The results in table 5.7 demonstrate the advantages of vector models for hyperarticulated speech. The error rate is reduced from 29.9% with the phone models to 21.5% with the full vector space. This is an improvement

of more than 28% relative. Moreover, this improvement on hyperarticulated 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%.

The performance for the sub-vector spaces is surprisingly good. The vector space formed by manner of articulation gives most of the gain. This suggests that only a limited number of contrastive attributes are needed to correct a recognition error. The hyperarticulated translation vector is projected down to a sub-space, but the remaining components are sufficient enough for resolving the word confusion. There is no indication that one of these sub-spaces is more important than another for compensating hyperarticulation. The results for all sub-spaces are comparable.

## 5.7 Analysis of Contrastive Attributes

The concept of contrastive attributes introduced above leads to predictions of changes in the articulatory vector spaces. Examples in section 5.3 support this theory. In this section, an analysis of the contrastive attributes will be presented to answer the question whether the predictions really occur in a hyper-clear speaking mode.

In a first step, the predictions need to be computed. To that end, the phone sequences of the confused words were aligned. For example, if *doubts* was uttered and *doubt* was recognised, a dynamic programming technique is used to align the sequences /D/ /AW/ /T/ /S/ and /D/ /AW/ /T/. The alignment procedure produces a set of insertions, deletions, and substitution pairs. The phone substitutions will then be represented in the articulatory vector space to obtain the difference vectors as explained in section 5.3. This alignment is performed for all utterances in the test set. A set of predictions about attribute changes is extracted for each phone unit in each utterance. It should be noted that not all phone occurrences have associated predictions, e.g. correct phone alignment does not produce any predictions. For those phones with predictions, there are 3.5 predicted attribute changes on average.

The statistical models for articulatory attributes can now be used to examine if the predicted changes do occur. As mentioned earlier, models and anti-models are used. Thus, the score function for an attribute $a$ is given by:

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

For each pair of normal and hyperarticulated utterances, the conditional probabilities for the attributes can be computed and $\Delta(o^H, a) - \Delta(o^N, a)$ gives the score difference between the hyperarticulated and normal data for an attribute $a$. The time alignments are obtained by the viterbi algorithm on the true transcripts. The scores are normalised by the number of frames, e.g. a longer duration of a hyperarticulated attribute does not change the score.

| | |
|---|---|
| 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 5.8: Predictions of contrastive attributes.

The table 5.8 contains the results for how often contrastive attributes are correctly predicted. 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 misrecognised word. Keeping this in mind, the results can be interpreted only as a correlation between predicted and observed changes and not as an indicator of the correctness of the predictor models.

The results in table 5.8 show that 51.2% attribute changes occured as predicted. 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%. More details are summarised in the tables 5.9 and 5.10. The prediction is very similar for all place and manner attributes. Only glottal sounds exhibit significantly less predicted changes.

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 “cheating experiment” since the

| attribute | Prediction Probability |
|---|---|
| plosive | 53.4% |
| nasal | 53.0% |
| fricative | 48.1% |
| lateral | 50.6% |
| approximant | 50.7% |

Table 5.9: Predictions of contrastive manner of articulation attributes.

| attribute | Prediction Probability |
|---|---|
| labial | 55.8% |
| bilabial | 55.3% |
| labiodental | 52.7% |
| alveolar | 53.4% |
| velar | 59.2% |
| glottal | 38.9% |

Table 5.10: Predictions of contrastive place of articulation attributes.

contrastive attributes are obtained by an alignment of the confused words. The baseline is the system with the full vector space. The result of this experiment is shown in table 5.11. The error rate improves from 21.5% to 17.0% on the hyperarticulated data. The results on the normal data are only depicted for comparison reasons, but this experiment does not have an effect on those data. Instead of using true transcripts to obtain contrastive attributes, hypotheses from the corresponding normal utterances can be used. In this case, the experiment is no longer a "cheating experiment". As shown in table 5.11, enforcing attributes based on hypotheses leads to a recognition performance of 19.4% error rate. This is an improvement of 9.8% relative.

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. There is a correlation between the observed and the predicted attribute changes. Enforcing contrastive attributes improves the recognition performance significantly.

| Contrastive Attributes | Speaking Style | |
|---|---|---|
| | normal | hyper |
| full vector space | 17.8% | 21.5% |
| enforced attributes (ref) | 17.8% | 17.0% |
| enforced attributes (hyp) | 17.8% | 19.4% |

Table 5.11: Enforcing contrastive attributes (results in word error rate).

## 5.8 Vector Models and Model Selection

So far, hyperarticulated training data are not used in the context of articulatory vector spaces. This is remarkable because this means that hyperarticulation can be compensated mostly without collecting special training data.[6] However, hyperarticulated training data are available for conducting training experiments. Model selection techniques as reported in chapter 4 made efficient use of such training data and led to significant error reduction.

In this section, we integrate the methods from chapter 4 into the articulatory vector space. Since the vector models rely on Gaussian mixture models, we can apply the same model separation technique as presented in chapter 4. The full vector space as described above is used for this experiment. Four combinations were investigated : phone vs. vector models and with or without model selection. The results are shown in table 5.12. The interesting numbers in table 5.12 are the error rates for hyperarticulated speech. The model selection for vector models does not work as well as for the phone models. Only a minor improvement from 21.5% to 20.8% is achieved by the selection of vector models. This is a rather small gain reduction compared to the phone models, where the error rate is decreased from 29.9% to 24.8%. It seems that recognition errors compensated by the articulatory vector space build a super-set of what model selection is able to repair. In summary, training data helps to compensate hyperarticulation as long as the model structure is not changed and invalid model assumptions can be "repaired" to a certain extent by collecting data. The vector models themselves led to significantly better error rates even without using hyperarticulated training data. Thus, using more appropriate models seems to be advantageous over

[6]Moreover, collecting hyperarticulated training data exhibits more difficulties compared to normal data collection procedures, since a special speaking mode is sought.

an approach based on fixing wrong model assumptions by collecting training data.

| model selection | Phone models | | Vector models | |
|---|---|---|---|---|
| | normal | hyper | normal | hyper |
| no | 18.9% | 29.9% | 17.8% | 21.5% |
| yes | 18.0% | 24.8% | 17.5% | 20.8% |

Table 5.12: Vector models and model selection (results in word error rates).

## 5.9 Utterance Combination

Model selection via a likelihood criterion can be viewed as combining the output from different recognition runs: the same utterance is decoded several times with different acoustic models. This approach can be extended as follows. We use here the fact that the hyperarticulated utterance is indeed a repetition of the word sequence spoken previously. This fact suggests combining the decoding output from the normal and the corresponding hyperarticulated utterance. This means that the additional knowledge provided by the hyperarticulated variant might be used to improve the recognition of the normal utterance. This is a different strategy than in all previous experiments. All previous experiments were conducted so as to understand and to compensate hyperarticulated speech. In contrast, utterance combination attempts to improve the recognition of normal speech by using additional information provided by hyperarticulated speech.

The approach chosen is technically speaking simple and is based on a majority voting strategy. Both the normal and the corresponding hyperarticulated utterance will be decoded with two acoustic model sets. The model sets are the same as used in the previous section for model selection. This means that four hypotheses are available for each normal utterance. The final output will be selected by a majority voting without confidences.

The results are shown in table 5.13. The baseline setup uses the vector models with model selection. The error rate is measured on the normal data. The utterance combination makes use of the normal and the corresponding hyperarticulated utterance to determine the output for the normal utterance. The recognition performance is improved by 12.7% relative, namely

| Utterance combination | Vector models |
|---|---|
| no | 17.5% |
| yes | 15.3% |

Table 5.13: Utterance combination (results in word error rates on normal speech).

from 17.5% to 15.3%. It should be noted that this result is quite important with respect to human friendly human-computer speech interfaces. A hyperarticulated repetition can be used to improve the recognition performance significantly.

## 5.10 Summary

In summary, we have shown in this chapter how articulatory attributes can be used for recognition of hyperarticulated speech. The main items are:

1. Hyperarticulation occurs on a sub-phonetic level.
2. The articulatory vector space can be used as a framework for representing articulatory changes.
3. Contrastive attributes explain hyperarticulated variations in an articulatory domain. In 78% of all phone occurrences, at least one attribute changes as predicted by the means of contrastive attributes.
4. Articulatory vector models reduce drastically the recognition errors for hyperarticulated speech. A relative error reduction of 28% is achieved.
5. Model selection does not lead to a major error reduction in the context of articulatory vector spaces. Recognition errors compensated by the articulatory vector space build a super-set of what model selection is able to repair.
6. Utterance combination leads to a significant error reduction for normal speech. The additional use of corresponding hyperarticulated utterances resulted in an improvement of 12%.

# Chapter 6

# A perception study

In chapter 3 we discussed the definition of the term hyperarticulation. We chose the pragmatic, problem-oriented approach. The solution is based on the observation that hyperarticulation occurs as a natural reaction for humans facing recognition errors. The intention when using hyperarticulation is to disambiguate the spoken or intended word from the (mis-)recognised word. We also raised the question of how this approach fits human perception of hyperarticulation. In other words, the question is, how do humans judge the degree of hyperarticulation in our database. This question can be answered by conducting a perception study, which is presented in the following sections.

The data of the perception study serve also to validate the central statements of this thesis from a user point of view as opposed to more abstract criteria, such as word error rate:

1. Hyperarticulation is a huge problem for automatic speech recognition. The error rate increases significantly wherever hyperarticulated speech occurs.

2. The use of hyperarticulated training data reduces the error rate by a certain amount but is not able to solve the problem.

3. The articulatory vector space compensates for hyperarticulation. Acoustic models based on composites of articulatory attributes reduce the error rate for hyperarticulated speech drastically.

## 6.1 Experimental Setup

The experimental setup for the perception study follows [Shriberg et al. '92] and [Hirschberg et al. '99], in which the data's degree of hyperarticulation was labeled independently by two expert human labelers, who were familiar with acoustics and phonetics. Both were native speaker of English. The labeling procedure allowed breaks to split the process into multiple sessions. This was necessary since the whole procedure took about 4-5 times real-time.

The turns were presented in random order to neutralise any prior information for the classification. The random order ensured that the labeler had no information whether the turn had been recorded in error-repair mode or not, which would provide information about the amount of hyperarticulation to be expected.

To further improve the reliability of the procedure, the labeler were allowed to replay the turn as often they want. No discussion or information exchange was allowed between labelers during the perception study, to ensure they were labeling the data independently.


Figure 6.1: Perception Study: User Interface

The user interface for the perception study is depicted in picture 6.1. As indicatd in the figure, there are three possible choices: "Not hyperarticulated", "some hyperarticulated", and "hyperarticulated". This scale is the same as used in the studies by [Shriberg et al. '92] and [Hirschberg et al. '99]. The instructions for the labeler included a description of typical characteristics for hyperarticulation, such as phonological features, slower speaking rate, increased pitch, intonation, or loudness.

## 6.2 Results

To avoid confusion for the different classes and categories, the naming convention applied in the following tables is: The term 'category' refers to the "error-repair" mode as described in chapter 3. There are two categories, 'normal' and 'hyper'. The term 'class' refers to the labels used for the perception study. There are three classes as described in the previous section, which are assigned the numerical values 0 ("not hyper"), 1 ("some hyper"), and 2 ("hyper").

### Raw Data

The results of the perception study are summarised in the following table. Table 6.1 contains the raw data, i.e. the statistics for each class and labeler.

| | class 0 | class 1 | class 2 |
|---|---|---|---|
| labeler 1 | 775 | 803 | 1018 |
| labeler 2 | 1332 | 662 | 602 |

Table 6.1: Counts for each class and labeler

### Interlabeler Agreement

Before we can *use* the data of the perception study we have to *validate* the data of the perception study themselves. Pearson's correlation coefficient can be used to measure whether both labeler assign the scores to the utterances in a consistent way. The following values are obtained:

$$\mu_{normal} = 0.524, \quad t_{normal} = 20.81 \tag{6.1}$$

$$\mu_{hyper} = 0.823, \quad t_{hyper} = 54.97 \tag{6.2}$$

Therefore, the correlation between the labelers is significant at $\alpha = 0.01$.

### Scores per category

We now can compare the labelers' scores with respect to the categories 'normal' and 'hyper', which refer to the "error-repair" mode. The arithmetic

average of class scores is defined as the overall score by combining the scores from each labeler. The following values are obtained:

$$\mu_{normal} = 0.48 \tag{6.3}$$
$$\mu_{hyper} = 1.25 \tag{6.4}$$

These results confirm that the data collected in the error-repair mode exhibit a high degree of hyperarticulation with respect to human perception. The labelers also perceived a small degree of hyperarticulation for the data collected in the normal mode.

## 6.3 Validation

### Statement 1

The first statement to validate is: Hyperarticulation is a huge problem for automatic speech recognition. The error rate increases significantly at hyperarticulated speech.

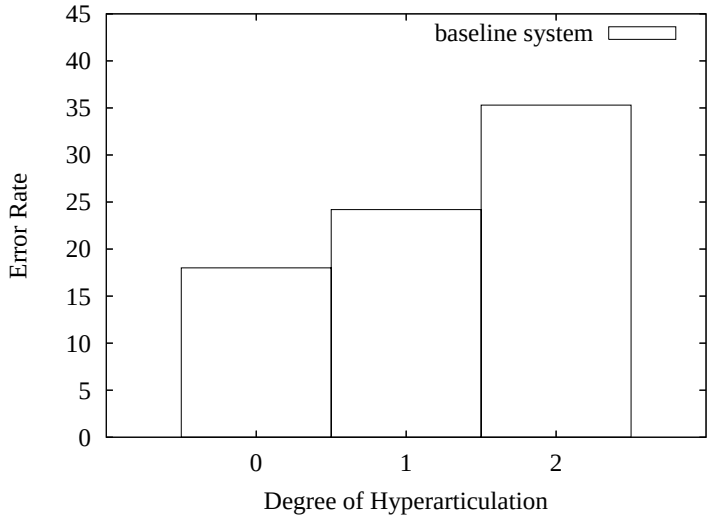


Figure 6.2: Baseline system: error rates with respect to human perception

Table 6.2 shows that the error rate increases drastically with an increasing degree of hyperarticulation. The number of recognition errors has more than doubled at degree 2 compared to degree 0.

## Statement 2

The second statement to validate is: The use of hyperarticulated training data reduces the error rate to a certain amount, but is not able to solve the problem.

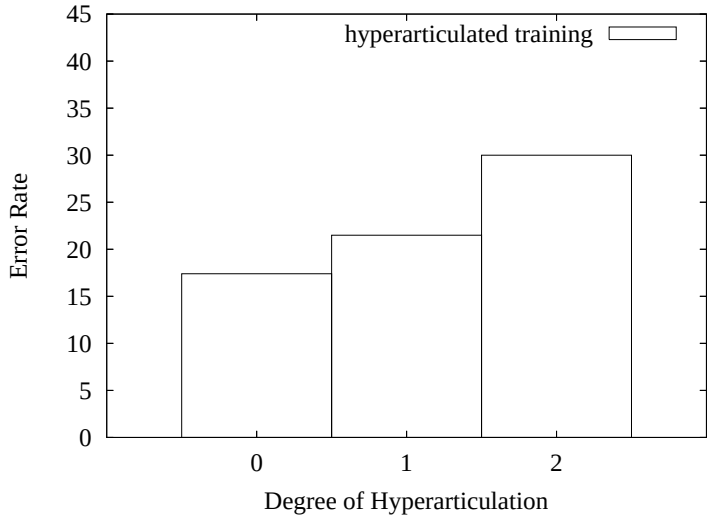


Figure 6.3: Training with hyperarticulated data: error rates with respect to human perception

Comparing the system after training with hyperarticulated data to the baseline system, we observe improved recognition of hyperarticulated utterances. However, the error rate is still much worse at degree 2 (30.0%) compared to degree 0 (17.4%).

## Statement 3

The third statement to validate is: The articulatory vector space compensates for hyperarticulation. Acoustic models based on composites of articulatory attributes reduce the error rate for hyperarticulated speech drastically.

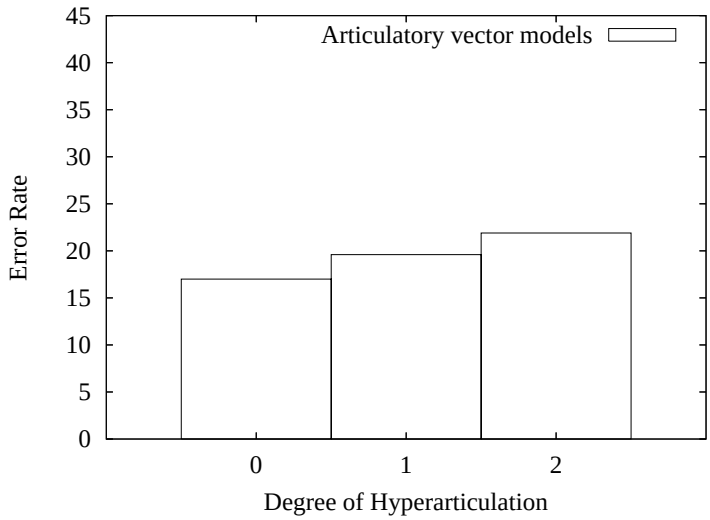


Figure 6.4: Articulatory vector models: error rates with respect to human perception

We observe a drastic improvement for hyperarticulated utterances. The error rate improves from 30.0% to 21.9% when using articulatory vector models. These numbers are in line with results in chapter 5.

Therefore, the central hypotheses outlined at the beginning of this chapter, which were derived through numerical analysis of standard criteria in ASR such as word error rate, have been confirmed by the outcome of the perception study.

# Chapter 7

# Investigations on Portability

All of the experiments described in the last chapters were conducted on a single corpus. Multiple training corpora were used, but all techniques were evaluated on the database for normal and hyperarticulated speech (HSC, table 3.1). In this chapter, experimental results are reported on other corpora, validating the concepts and algorithms developed for compensating hyperarticulated speech. In the first section, the techniques are validated on a different language. In the second section, the SUSAS (speech under simulated and actual stress) corpus is used to extend the work to other speaking modes.

## 7.1 Transfer to Other Languages

In this section, a comparison of hyperarticulated effects in English and German is given. The German corpus was obtained using a similar procedure as described in section 3.2. The recordings were collected with a simulated dialogue system. The subjects were seated in front of a computer and were asked to correct previous recognition errors. The subjects were not told that the system was a simulation only. Since the same setup was used for both data collections, a comparison of hyperarticulated effects across different languages is possible without the danger of uncontrolled side effects. The size of the German corpus is slightly larger than the English one.

A test set of 20 speakers is available, which consists of around 2 hours of speech (table 7.1). The baseline recogniser is derived from the Verbmobil-II evaluation system. Details of this system can be found in

| | speaker | utterances | | speech | |
|---|---|---|---|---|---|
| | | normal | hyper | normal | hyper |
| train | 61 | 5901 | 7309 | 154 min | 235 min |
| test | 20 | 1926 | 2374 | 47 min | 72 min |
| all | 81 | 7827 | 9683 | 202 min | 307 min |

Table 7.1: German Corpus for normal and hyperarticulated speech.

[Soltau et al. 2001b]. The system features state-of-the art acoustic modeling techniques and achieved first ranks in a series of ASR evaluations on the Verbmobil task. Initial experiments show a significant performance degradation for hyperarticulated speech on the German corpus.

| Language | test set | |
|---|---|---|
| | normal | hyper |
| German | 20.4% | 27.1% |
| English | 18.9% | 29.9% |

Table 7.2: Performance degradation at hyperarticulated speech for German and English.

For further experiments, we partitioned the test set into 4 sub-groups according to the error rate. An error rate change of more than 5% was considered as a significant change. The sub-groups are summarised in the table 7.3. A significantly worse recognition performance was observed for 12 out of 20 speakers. For two speakers, a significant improvement was observed, while for six other speakers non-significant changes were observed.

## Phone Duration

An analysis of the phone durations gave results similar to those for the English corpus. The phone duration increased significantly. Furthermore, there is a correlation between phone duration and error rate as shown in table 7.4. Those speakers with a higher error rate in a hyper-clear speaking mode also exhibit 30% higher phone durations. On the other hand, speakers with a better recognition performance do not show higher phone durations.

| speaker group depending on WER | spk | speaking mode | | Δ WER |
|---|---|---|---|---|
| | | normal | hyper | |
| significantly better | 2 | 27.5% | 20.4% | -7.1% |
| significantly worse | 12 | 18.1% | 28.6% | +10.5% |
| slightly better | 3 | 18.6% | 17.5% | +1.1% |
| slightly worse | 3 | 20.7% | 23.1% | -2.4% |

Table 7.3: Sub test groups partitioned according to error rate changes

| speaker group depending on WER | increased duration | | |
|---|---|---|---|
| | voiced | unvoiced | plosives |
| significantly better | 3.9% | -0.4% | -4.2% |
| significantly worse | 25.7% | 31.2% | 24.4% |
| slightly better | 8.2% | 3.9% | 15.2% |
| slightly worse | 17.9% | 22.4% | 17.3% |

Table 7.4: Phone durations versus error rate

## Pitch

To analyse the effect of pitch, a T-Test was performed using a quantile of $\alpha = 0.05$. The test set was partitioned into three groups with respect to the $F0$ mean. The table 7.5 indicates a relation between an increased $F0$ mean and higher error rates. These results confirm those reached on the English corpus.

| $F0$ | speaker | speaking mode | | Δ WER |
|---|---|---|---|---|
| | | normal | hyper | |
| increasing | 8 | 18.8% | 29.3% | 10.5% |
| decreasing | 6 | 17.5% | 18.6% | 1.1% |
| changed not | 6 | 22.2% | 26.4% | 4.2% |

Table 7.5: Word error rate as a function of $F0$ changes

## Model Selection

In the next experiment, the model selection was validated on the German corpus. A separate set of acoustic models was generated for each speaking mode using a regression tree of linear transforms. A likelihood criterion is used to select the appropriate model set. A comparison of the results for the German and English corpus is shown in table 7.6.

| Language | German | | English | |
|---|---|---|---|---|
| | normal | hyper | normal | hyper |
| shared models | 19.7% | 25.7% | 18.9% | 29.9% |
| model selection | 18.5% | 22.0% | 18.0% | 24.8% |

Table 7.6: Model Selection : Comparison of German and English

The relative improvement for model selection on the German corpus (13.4%) is smaller than for the English corpus (17.5%). However, the improvements on both corpora are significant.

## Articulatory Vector Space

The articulatory vector space for the German language is constructed in a way similar to that for the English corpus. Manner of articulation is modeled by 5 dimensions (plosive, fricative, lateral, vibrant, nasal), and the place of articulation occupies four dimensions (labial, glottal, labiodental, velar). A separate attribute is used to distinguish diphthongs. The vowel dimensions are the same as for English. The vector models lead to significant improvements on hyperarticulated speech for both languages. The relative error reduction is 19.8% for German and 28.1% for English.

| Language | German | | English | |
|---|---|---|---|---|
| | normal | hyper | normal | hyper |
| phone models | 19.7% | 25.7% | 18.9% | 29.9% |
| vector models | 16.5% | 20.6% | 17.8% | 21.5% |

Table 7.7: Articulatory Vector Space : Comparison of German and English

## Model Selection in an Articulatory Vector Space

The next experiment investigates the use of model selection in an articulatory vector space. Analogous to phone models, vector models can be separated into normal and hyperarticulated sub-sets. The experimental results in the table below show differences for German and English.

| model selection | German | | English | |
|---|---|---|---|---|
| | normal | hyper | normal | hyper |
| no | 16.5% | 20.6% | 17.8% | 21.5% |
| yes | 16.4% | 16.9% | 17.5% | 20.8% |

Table 7.8: Selection of vector models : Comparison of German and English

The gains from model selection and vector models are fully additive on the German corpus. The selection of phone models led to a 14.4% error reduction, vector models alone gave 19.8%, and all together there was a 34.2% error reduction (from 25.7% to 16.9%)on the German corpus. Moreover, the performance for normal and hyperarticulated speech was now nearly balanced.

## Utterance Combination

To complete the experiments on hyperarticulated speech for German, the portability of utterance combination was investigated. Utterance combination was introduced in the previous chapter to combine knowledge from normal and hyperarticulated speech. It is based on a four-fold majority voting scheme. Both the normal and the corresponding hyperarticulated utterance will be decoded using two acoustic model sets. This means that four hypotheses are available for each normal utterance. The two model sets are the same as in the previous experiment (table 7.8).

| Utterance Combination | German | English |
|---|---|---|
| No | 16.4% | 17.5% |
| Yes | 14.0% | 15.3% |

Table 7.9: Utterance combination for German and English.

As shown in table 7.9, utterance combination reduces the error rate significantly for both languages. The relative improvement is 14.6% for German and 12.5% for English. The better results for German can be attributed to a smaller gap in recognition performance between normal and hyperarticulated speech for German (16.4% for normal, 16.9% for hyper) compared to English (17.5% for normal, 20.8% for hyper).

It can be concluded that hyperarticulation occurs both in German and English and has similar effects. Articulatory vector spaces, model selection, and utterance combination gave significant improvements on hyperarticulated speech for both languages.

## 7.2 Transfer to Other Speaking Modes

The SUSAS (speech under actual and simulated stress) corpus [Hansen et al. '98] allows for studying variations across different speaking styles and emotions. The database contains multiple domains, such as talking styles, stress under workload, and psychiatric analyses. There are multiple domains:

- Talking style domain
  Data from the talking style domain were collected by the Lincoln Laboratory. The speaking modes are : slow, fast, soft, loud, angry, and question. The vocabulary consists of 35 aircraft communication words. The selected words are typically difficult to recognise, e.g. six-fix, white-wide, three-thirty, and eight-eighty. There were nine subjects. Each word was produced 28 times by each subject. The total number of tokens was 8820.

- Stress under workload
  The vocabulary and the speakers are the same as for the talking style domain. The task consists of a "response to a marginally stable, single-pole system". The degree of instability can be adjusted to create different levels of workload. The corpus contains also some data for investigating the Lombard effect.

More details about the corpus can be found in [Hansen et al. '98]. Despite the small vocabulary, the big advantage is that the corpus allows for

investigations across different speaking modes since the test speakers and vocabulary are identical for all speaking modes. These speaking styles exhibit changes in duration, pitch, intensity, and spectrum [Hansen '96].

Since the vector models are designed to capture articulatory changes in different speaking modes, we investigate the effectiveness of our approach on the SUSAS corpus.

## Experimental Setup

The SWB system [Soltau et al. 2003] is used as a baseline, since the SUSAS data are sampled at 8 kHz which fit the SWB models. This system features several acoustic normalisation and adaptation techniques, as well as cross-word contexts and penta-phone models. The parameters were trained using a mixing-up procedure. Furthermore, a maximum mutual information criterion was applied. Phone dependent semi-tied full covariances are used as well. The phone models make use of more than $288,000$ Gaussian densities.

The vector models were built as described in the previous chapter. The full vector space is used and the models are trained on the SWB corpus. The basis of the vector space consists of : plosive, fricative, lateral, approximant, alveolar, bilabial, glottal, labiodental, interdental, retroflex, high, mid, low, front, central, back, round, voiced, and consonantal. No training was performed on the SUSAS corpus. The total number of Gaussians for attribute modeling is $1,216$. The number of additional parameters needed for the vector models is, therefore, only a small fraction compared to the phone models.

## Results

The results are summarised in table 7.10. The categories neutral, slow, fast, soft, angry, loud, and question belong to the first domain of the SUSAS corpus. Moderate (c50) and high (c70) workload stress and the Lombard category originate from the second domain.

## Discussion

There is clear evidence that the vector models perform substantially better than the phone models on most of the categories, in particular for speech under workload stress and slow speech. No special optimisation was performed

| style | phone models | vector models | improvement |
|---|---|---|---|
| neutral | 7.0% | 6.6% | 6.7% |
| slow | 22.5% | 18.4% | 18.2% |
| fast | 12.5% | 12.4% | 0.8% |
| soft | 10.6% | 9.7% | 8.5% |
| loud | 21.1% | 22.2% | -5.2% |
| angry | 28.3% | 25.9% | 8.5% |
| question | 9.7% | 8.9% | 8.2% |
| stress (c50) | 8.4% | 7.0% | 16.6% |
| stress (c70) | 7.0% | 6.0% | 14.3% |
| Lombard | 12.2% | 11.7% | 4.1% |

Table 7.10: Comparison of phone and vector models on the SUSAS corpus (error rates).

for these categories, e.g. the setup is identical to the one for hyperarticulated speech. The results confirm the effectiveness of the vector models with articulatory attributes.

# Chapter 8

# Conclusions

We showed in this thesis that it is important to examine how automatic speech recognition is being used for real world applications. Humans switch to a hyper-clear speaking mode as a natural reaction to resolve word confusions in a dialogue system. Current state-of-the-art acoustic modeling techniques fail to capture hyperarticulated effects due to invalid model assumptions. Therefore, hyperarticulation causes an increased word error rate contrary to the user's expectations. To allow more natural human-to-machine interactions, automatic speech recognition systems must be able to deal with such effects.

To understand why hyperarticulated speech is hard to recognise, we analysed the unique features of this speaking mode. The acoustic-articulatory space of hyperarticulated speech differs from canonical speech, particularly with respect to phone duration, pitch contour, and formant frequencies. As a consequence, the characteristics of hyperarticulated speech will not be covered by the parameters of canonical phone models. The results of the analysis indicate that hyperarticulated effects occur on a sub-phonetic level in an articulatory domain. Therefore, standard acoustic modeling techniques using phones as base units cannot compensate for such effects.

This thesis has presented novel techniques for compensating for hyperarticulated effects in automatic speech recognition. The error rate was reduced by more than 28% using acoustic models based on an articulatory vector space. The vector model consists of probability density functions for each dimension. An exponential combination of the underlying function leads to a score function for the speech events. The articulatory vector space allows the definition of an elegant representation of changes in a hyper-clear speaking

mode. The concept of contrastive attributes explains hyperarticulation as an inversion of those attributes which discriminate between the spoken and the recognised word. This allows us to define a translation vector for modeling hyperarticulated changes from a canonical pronunciation and therefore allows the prediction of hyperarticulated effects. The phenomena of hyperarticulation can then be interpreted as a warping of trajectories in an articulatory vector space. These composites of articulatory and phonetic units can be trained via the Baum-Welch algorithm maximising the training likelihood. The articulatory models can be trained on shared data from different phones and therefore allow a better estimation of speaking mode invariant speech characteristics.

Another important outcome of this thesis is that hyperarticulated speech provides additional knowledge for improving the recognition of normal speech. A combination of normal with corresponding hyperarticulated utterances results in a significant error reduction. The utterance combination is based on a majority voting scheme using multiple utterances and models. The error decreased from 17.5% to 15.3%, a relative improvement of 12.5%. It should be noted that this result is quite important with respect to human-friendly human-computer speech interfaces. A hyperarticulated repetition can be used to improve the recognition performance significantly.

In further experiments, we investigated the efficient use of hyperarticulated training data. Model selection triggered by a likelihood criterion achieved an error reduction of 17%. However, articulatory vector models outperformed model selection significantly. Invalid model assumptions cannot be "repaired" by using hyperarticulated training data. Furthermore, we investigated the use of speaking mode dependent decision trees to capture hyperarticulated effects. These decision trees were trained on normal- and hyper-articulated data. Based on a maximum likelihood criterion, acoustic models can be specialised to a certain speaking style. These experiments showed that models related mainly to the place of articulation were separated into speaking style dependent sub-models. An error reduction of 9% was obtained by these specialised acoustic models.

In order to investigate the capabilities of the compensation techniques, we extended the experiments from English to other languages and speaking styles. On both German and English, similar performance degradations were observed in a hyperarticulated speaking mode. Pitch, formants, and phone durations exhibit similar changes. The articulatory vector space for German is constructed using the same procedure as for English. The vector models

achieved significant improvements on both languages. Experiments on the SUSAS database (speech under simulated and actual stress) confirmed the effectiveness of the developed modeling techniques for several other speaking modes, such as speech under stress, or emotional speech.

# Future Work

Although many questions regarding hyperarticulation and articulatory attributes are addressed in this thesis, several extensions of this work are possible. First, hyperarticulation might also occur in different scenarios. In this thesis, we investigated hyperarticulated effects in the context of error recovery strategies. The effects may vary to a certain extent across different scenarios. For example, an analysis of hyperarticulation in different speaking styles was presented in [Köster 2001], where hyperarticulated effects for words, sentences, and dialogues were studied.

Another interesting question is whether hyperarticulated speech can lead to improved automatic speech recognition. Hyperarticulated speech improves the intelligibility for humans as was demonstrated in [Picheny et al. '86] for hearing impaired people. Therefore, hyperarticulated speech might have the potential to produce lower error rates also for automatic speech recognisers, if the contrast between normal and hyperarticulated speech was useful for achieving improvements. The experiment with enforcing contrastive attributes in chapter 5 suggests that contrastive attributes indeed contain additional information.

From a mathematical point of view, it is not satisfying to work with non-probability density functions in the articulatory vector space. It can not be guaranteed that the probability mass of the combined PDFs sum up to one. A mass normalisation based on fixed weights is not sufficient. What is needed is an integrated solution for estimating the weights and the normalisation. A reliable procedure for weight estimation would also be beneficial for speaker adaptation. For example, the potential use of attribute selection for speaker adaptation was studied in [Metze & Waibel 2003]. Weight estimation in combination with context and mode dependent attributes would provide a more powerful framework for modeling of speech events.

# Appendix A

# Phoneset

| | |
|---|---|
| PHONES | @ +AH +CL +GE +GH +H# +LS +MU +PA +UH +UM AA AE AH AO AW AX AXR AY B CH D DH DX EH ER EY F G HH IH IX IY JH K L M N NG OW P R S SH SIL T TH UH UW V W Y Z |
| NOISES | +AH +CL +GE +GH +H# +LS +MU +PA +UH +UM |
| HUMAN-NOISES | +AH +GH +H# +LS +UH +UM |
| SILENCES | SIL |
| CONSONANT | P B F V TH DH T D S Z SH CH JH K G HH M N NG R Y W L ER DX AXR |
| CONSONANTAL | P B F V TH DH T D S Z SH CH JH K G HH M N NG DX |
| OBSTRUENT | P B F V TH DH T D S Z SH CH JH K G |
| SONORANT | M N NG R Y W L ER AXR DX |
| SYLLABIC | AY EY IY AW OW EH IH AO AE AA AH UW UH IX AX ER AXR |
| VOWEL | AY EY IY AW OW EH IH AO AE AA AH UW UH IX AX |
| DIPHTHONG | AY EY AW OW |
| CARDVOWEL | IY IH EH AE AA AH AO UH UW IX AX |
| VOICED | B D G JH V DH Z M N NG W R Y L ER AY EY IY AW OW EH IH AO AE AA AH UW UH DX AXR IX AX |
| UNVOICED | P F TH T S SH CH K |
| CONTINUANT | F TH S SH V DH Z W R Y L ER |
| DEL-REL | CH JH |
| LATERAL | L |
| ANTERIOR | P T B D F TH S SH V DH Z M N W Y L DX |
| CORONAL | T D CH JH TH S SH DH Z N L R DX |
| APICAL | T D N DX |
| 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 |
| SIBILANT | S SH Z 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 DX |
| ALVEOPALATAL | SH CH JH |
| ALVEOLAR | T D N S Z L SH CH JH DX |
| 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 |
| FLAP | DX |
| NASAL | M N NG |
| FRICATIVE | F V TH DH S Z SH 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 |
| VLS-FR | F TH SH |
| VCD-FR | V DH |
| ROUND | AO OW UH UW 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 | OW EY |
| BF-DIP | AY AW OW |
| Y-DIP | AY EY |
| W-DIP | AW OW |
| ROUND-DIP | 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 Y |
| LQGL-BACK | L R W |

# Appendix B

# Training Data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2481 | 2482 | 2483 | 2484 | 2485 | 2486 | 2487 | 2488 | 2489 | 2490 |
| 2491 | 2492 | 2493 | 2494 | 2495 | 2496 | 2497 | 2498 | 2499 | 2500 |
| 2501 | 2502 | 2503 | 2504 | 2505 | 2506 | 2507 | 2508 | 2509 | 2510 |
| 2511 | 2512 | 2513 | 2514 | 2515 | 2516 | 2517 | 2518 | 2519 | 2520 |
| 2521 | 2522 | 2523 | 2524 | 2525 | 2526 | 2527 | 2528 | 2529 | 2530 |
| 2531 | 2532 | 2533 | 2534 | 2535 | 2536 | 2537 | 2538 | 2539 | 2540 |
| 2541 | 2542 | 2543 | 2544 | 2545 | 2546 | 2547 | 2548 | 2549 | 2550 |
| 2551 | 2552 | 2553 | 2554 | 2555 | 2556 | 2557 | 2558 | 2559 | 2560 |
| 2561 | 2562 | 2563 | 2564 | 2565 | 2566 | 2567 | 2568 | 2569 | 2570 |
| 2571 | 2572 | 2573 | 2574 | 2575 | 2576 | 2577 | 2578 | 2579 | 2580 |
| 2581 | 2582 | 2583 | 2584 | 2585 | 2586 | 2587 | 2588 | 2589 | 2590 |
| 2591 | 2592 | 2593 | 2594 | 2595 | 2596 | 2597 | 2598 | 2599 | 2600 |
| 2601 | 2602 | 2603 | 2604 | 2605 | 2606 | 2607 | 2608 | 2609 | 2610 |
| 2611 | 2612 | 2613 | 2614 | 2615 | 2616 | 2617 | 2618 | 2619 | 2620 |
| 2621 | 2622 | 2623 | 2624 | 2625 | 2626 | 2627 | 2628 | 2629 | 2630 |
| 2631 | 2632 | 2633 | 2634 | 2635 | 2636 | 2637 | 2638 | 2639 | 2640 |
| 2641 | 2642 | 2643 | 2644 | 2645 | 2646 | 2647 | 2648 | 2649 | 2650 |
| 2651 | 2652 | 2653 | 2654 | 2655 | 2656 | 2657 | 2658 | 2659 | 2660 |
| 2661 | 2662 | 2663 | 2664 | 2665 | 2666 | 2667 | 2668 | 2669 | 2670 |
| 2671 | 2672 | 2673 | 2674 | 2675 | 2676 | 2677 | 2678 | 2679 | 2680 |
| 2681 | 2682 | 2683 | 2684 | 2685 | 2686 | 2687 | 2688 | 2689 | 2690 |
| 2691 | 2692 | 2693 | 2694 | 2695 | 2696 | 2697 | 2698 | 2699 | 2700 |
| 2701 | 2702 | 3683 | 3684 | 3685 | 3686 | 3687 | 3688 | 3689 | 3690 |
| 3691 | 3692 | 3693 | 3694 | 3695 | 3696 | 3697 | 3698 | 3699 | 3700 |
| 3701 | 3702 | 3703 | 3704 | 3705 | 3706 | 3707 | 3708 | 3709 | 3710 |

3711 3712 3713 3714 3715 3716 3717 3718 3719 3720
3721 3722 3723 3724 3725 3726 3727 3728 3729 3730
3731 3732 3733 3734 3735 3736 3737 3738 3739 3740
3741 3742 3743 3744 3745 3746 3747 3748 3749 3750
3751 3752 3753 3754 3755 3756 3757 3758 3759 3760
3761 3762 3763 3764 3765 3766 3767 3768 3769 3770
3771 3772 3773 3774 3775 3776 3777 3778 3779 3780
3781 3782 3783 3784 3785 3786 3787 3788 3789 3790
3791 3792 3793 3794 3795 3796 3797 3798 3799 3800
3801 3802 3803 3804 3805 3806 3807 3808 3809 3810
3811 3812 3813 3814 3815 3816 3817 3818 3819 3820
3821 3822 3823 3824 3825 3826 3827 3828 3829 3830
3831 3832 3833 3834 3835 3836 3837 3838 3839 3840
3841 3842 3843 3844 3845 3846 3847 3848 3849 3850
3851 3852 3853 3854 3855 3856 3857 3858 3859 3860
3861 3862 3863 3864 3865 3866 3867 3868 3869 3870
3871 3872 3873 3874 3875 3876 3877 3878 3879 3880
3881 3882 3883 3884 3885 3886 3887 3888 3889 3890
3891 3892 3893 3894 3895 3896 3897 3898 3899 3900
3901 1461 1462 1463 1464 1465 1466 1467 1468 1469
1470 1471 1472 1473 1474 1475 1476 1477 1478 1479
1480 1481 1482 1483 1484 1485 1486 1487 1488 1489
1490 1491 1492 1493 1494 1495 1496 1497 1498 1499
1500 1501 1502 1503 1504 1505 1506 1507 1508 1509
1510 1511 1512 1513 1514 1515 1516 1517 1518 1519
1520 1521 1522 1523 1524 1525 1526 1527 1528 1529
1530 1531 1532 1533 1534 1535 1536 2677 2678 2679
2680 2681 2682 2683 2684 2685 2686 2687 2688 2689
2690 2691 2692 2693 2694 2695 2696 2697 2698 2699
2700 2701 2702 2703 2704 2705 2706 2707 2708 2709
2710 2711 2712 2713 2714 2715 2716 2717 2718 2719
2720 2721 2722 2723 2724 2725 2726 2727 2728 2729
2730 2731 2732 2733 2734 2735 2736 2737 2738 2739
2740 2741 2742 2743 2744 2745 2746 2747 2748 2749
2750 2751 2752 2753 2754 2755 2756 2757 2758 2759
2760 2761 2762 2763 2764 2765 2766 2767 2768 2769
2770 2771 2772 2773 2774 2775 2776 2777 2778 2779
2780 2781 2782 2783 2784 2785 2786 2787 2788 2789

2790 2791 2792 2793 2794 2795 2796 2797 2798 2799
2800 2801 2802 2803 2804 2805 2806 2807 2808 2809
2810 2811 2812 2813 2814 2815 2816 2817 2818 2819
2820 2821 2822 2823 2824 2825 2826 2827 2828 2829
2830 2831 2832 2833 2834 2835 2836 2837 2838 2839
2840 2841 2842 2843 2844 2845 2846 2847 2848 2849
2850 2851 2852 2853 2854 2855 2856 2857 2858 2859
2860 2861 2862 2863 2864 2865 2866 2867 2868 2869
2870 2871 2872 2873 2874 2875 2876 2877 2878 2879
2880 2881 2882 2883 2884 2885 2886 2887 2888 2889
2890 2891 2892 2893 2894 2895 2896 2897 2898 2899
2900 2901 2902 2903 2904 2905 2906 2907 2908 2909
2910 2911 2912 2913 2914 2915 2916 2917 2918 2919
2920 2921 2922 2923 2924 2925 2926 2927 2928 2929
2930 2931 2932 2933 2934 2935 2936 2937 2938 2939
2940 2941 2942 2943 2944 2945 2946 2947 2948 2949
2950 2951 2952 2953 2954 2955 2956 2957 2958 2959
2960 2961 2962 2963 2964 2965 2966 2967 2968 2969
2970 2971 2972 2973 2974 2975 2976 2977 2978 2979
2980 2981 2982 2983 2984 2985 2986 2703 2704 2705
2706 2707 2708 2709 2710 2711 2712 2713 2714 2715
2716 2717 2718 2719 2720 2721 2722 2723 2724 2725
2726 2727 2728 2729 2730 2731 2732 2733 2734 2735
2736 2737 2738 2739 2740 2741 2742 2743 2744 2745
2746 2747 2748 2749 2750 2751 2752 2753 2754 2755
2756 2757 2758 2759 2760 2761 2762 2763 2764 2765
2766 2767 2768 2769 2770 2771 2772 2773 2774 2775
2776 2777 2778 2779 2780 2781 2782 2783 2784 2785
2786 2787 2788 2789 2790 2791 2792 2793 2794 2795
2796 2797 2798 2799 2800 2801 2802 2803 2804 2805
2806 2807 2808 2809 2810 2811 2812 2813 2814 2815
2816 2817 2818 2819 2820 2821 2822 2823 2824 2825
2826 2827 2828 2829 2830 2831 2832 2833 2834 2835
2836 2837 2838 2839 2840 2841 2842 2843 2844 2845
2846 2847 2848 2849 2850 2851 2852 2853 2854 2855
2856 2857 2858 2859 2860 2861 2862 2863 2864 2865
2866 2867 2868 2869 2870 2871 2872 2873 2874 2875
2876 2877 2878 2879 2880 2881 2882 2883 2884 2885

2886 2887 2888 2889 2890 2891 2892 2893 2894 2895
2896 2897 2898 2899 2900 2901 2902 2903 2904 2905
2906 2907 2908 2909 2910 2911 2912 2913 2914 2915
2916 2917 2918 2919 2920 2921 2922 2923 2924 2925
2926 2927 2928 2929 2930 2931 2932 2933 2934 2935
2936 2937 2938 4136 4137 4138 4139 4140 4141 4142
4143 4144 4145 4146 4147 4148 4149 4150 4151 4152
4153 4154 4155 4156 4157 4158 4159 4160 4161 4162
4163 4164 4165 4166 4167 4168 4169 4170 4171 4172
4173 4174 4175 4176 4177 4178 4179 4180 4181 4182
4183 4184 4185 4186 4187 4188 4189 4190 4191 4192
4193 4194 4195 4196 4197 4198 4199 4200 4201 4202
4203 4204 4205 4206 4207 4208 4209 4210 4211 4212
4213 4214 4215 4216 4217 4218 4219 4220 4221 4222
4223 4224 4225 4226 4227 4228 4229 4230 4231 4232
4233 4234 4235 4236 4237 4238 4239 4240 4241 4242
4243 4244 4245 4246 4247 4248 4249 4250 4251 4252
4253 4254 4255 4256 4257 4258 4259 4260 4261 4262
4263 4264 4265 4266 4267 4268 4269 4270 4271 4272
4273 4274 4275 4276 4277 4278 4279 4280 4281 4282
4283 4284 4285 4286 4287 4288 4289 4290 4291 4292
4293 4294 4295 4296 4297 4298 4299 4300 4301 4302
4303 4304 4305 4306 4307 4308 4309 4310 4311 4312
4313 4314 4315 4316 4317 4318 4319 4320 4321 4322
4323 4324 4325 4326 4327 4328 4329 4330 4331 4332
4333 4334 4335 4336 4337 4338 4339 4340 4341 4342
4343 4344 4345 4346 4347 4348 4349 4350 4351 4352
4353 4354 4355 4356 4357 4358 4359 0384 0385 0386
0387 0388 0389 0390 0391 0392 0393 0394 0395 0396
0397 0398 0399 0400 0401 0402 0403 0404 0405 0406
0407 0408 0409 0410 0411 0412 0413 0414 0415 0416
0417 0418 0419 0420 0421 0422 0423 0424 0425 0426
0427 0428 0429 0430 0431 0432 0433 0434 0435 0436
0437 0438 0439 0440 0441 0442 0443 0444 0445 0446
0447 0448 0449 0450 0451 0452 0453 0454 0455 0456
0457 0458 0459 0460 0461 0462 0463 0464 0465 0466
0467 0468 0469 0470 0471 0472 0473 0474 0475 0476
0477 0478 0479 0480 0481 0482 0483 0484 0485 0486

0487 0488 0489 0490 0491 0492 0493 0494 0495 0496
0497 0498 0499 0500 0501 0502 0503 0504 0505 0506
0507 0508 0509 0510 0511 0512 0513 0514 0515 0516
0517 0518 0519 0520 0521 0522 0523 0524 0525 0526
0527 0528 0529 0530 0531 0532 0533 0534 0535 0536
0537 0538 0539 0540 0541 0542 0543 0544 0545 0546
0547 0548 0549 0550 0551 0552 0553 0554 0555 0556
0557 0558 0559 0560 0561 0562 0563 0564 0565 0566
0567 0568 0569 0570 0571 0572 0573 0574 0575 0576
0577 0578 0579 0580 0581 0582 0583 0584 0585 0586
0587 0588 0589 0590 5675 5676 5677 5678 5679 5680
5681 5682 5683 5684 5685 5686 5687 5688 5689 5690
5691 5692 5693 5694 5695 5696 5697 5698 5699 5700
5701 5702 5703 5704 5705 5706 5707 5708 5709 5710
5711 5712 5713 5714 5715 5716 5717 5718 5719 5720
5721 5722 5723 5724 5725 5726 5727 5728 5729 5730
5731 5732 5733 5734 5735 5736 5737 5738 5739 5740
5741 5742 5743 5744 5745 5746 5747 5748 5749 5750
5751 5752 5753 5754 5755 5756 5757 5758 5759 5760
5761 5762 5763 5764 5765 5766 5767 5768 5769 5770
5771 5772 5773 5774 5775 5776 5777 5778 5779 5780
5781 5782 5783 5784 5785 5786 5787 5788 5789 5790
5791 5792 5793 5794 5795 5796 5797 5798 5799 5800
5801 5802 5803 5804 5805 5806 5807 5808 5809 5810
5811 5812 5813 5814 5815 5816 5817 5818 5819 5820
5821 5822 5823 5824 5825 5826 5827 5828 5829 5830
5831 5832 5833 5834 5835 5836 5837 5838 5839 5840
5841 5842 5843 5844 5845 5846 5847 5848 5849 5850
5851 5852 5853 5854 5855 5856 5857 5858 5859 5860
5861 5862 5863 5864 5865 5866 5867 5868 5869 5870
5871 5872 5873 5874 5875 5876 5877 5878 5879 5880
5881 5882 5883 5884 5885 5886 5887 5888 5889 5890
5891 5892 2372 2373 2374 2375 2376 2377 2378 2379
2380 2381 2382 2383 2384 2385 2386 2387 2388 2389
2390 2391 2392 2393 2394 2395 2396 2397 2398 2399
2400 2401 2402 2403 2404 2405 2406 2407 2408 2409
2410 2411 2412 2413 2414 2415 2416 2417 2418 2419
2420 2421 2422 2423 2424 2425 2426 2427 2428 2429

2430 2431 2432 2433 2434 2435 2436 2437 2438 2439
2440 2441 2442 2443 2444 2445 2446 2447 2448 2449
2450 2451 2452 2453 2454 2455 2456 2457 2458 2459
2460 2461 2462 2463 2464 2465 2466 2467 2468 2469
2470 2471 2472 2473 2474 2475 2476 2477 2478 2479
2480 2481 2482 2483 2484 2485 2486 2487 2488 2489
2490 2491 2492 2493 2494 2495 2496 2497 2498 2499
2500 2501 2502 2503 2504 2505 2506 2507 2508 2509
2510 2511 2512 2513 2514 2515 2516 2517 2518 2519
2520 2521 2522 2523 2524 2525 2526 2527 2528 2529
2530 2531 2532 2533 2534 2535 2536 2537 2538 2539
2540 2541 2542 2543 2544 2545 2546 2547 2548 2549
2550 2551 2552 2553 2554 2555 2556 2557 2558 2559
2560 2561 2562 2563 2564 2565 2566 2567 2568 2569
2570 2571 2572 2573 2574 2575 2576 2577 2578 2579
2580 2581 2582 2583 2584 2585 2586 2587 2588 2589
2590 2591 2592 2593 2594 2595 2596 2597 2598 2599
2600 2601 2602 2603 2604 2605 2606 2607 2608 2609
2610 2611 2612 2613 2614 2615 2616 2617 2618 2619
2620 2621 2622 2623 2624 2625 2626 2627 2628 2629
2630 2631 2632 2633 2634 2635 2636 2637 2638 2639
2640 2641 2642 2643 2644 2645 2646 2647 2648 2649
2650 2651 2652 2653 2654 2655 2656 2657 2658 2659
2660 2661 2662 2663 2664 2665 2666 2667 2668 2669
2670 2671 2672 2673 2674 2675 2676 7419 7420 7421
7422 7423 7424 7425 7426 7427 7428 7429 7430 7431
7432 7433 7434 7435 7436 7437 7438 7439 7440 7441
7442 7443 7444 7445 7446 7447 7448 7449 7450 7451
7452 7453 7454 7455 7456 7457 7458 7459 7460 7461
7462 7463 7464 7465 7466 7467 7468 7469 7470 7471
7472 7473 7474 7475 7476 7477 7478 7479 7480 7481
7482 7483 7484 7485 7486 7487 7488 7489 7490 7491
7492 7493 7494 7495 7496 7497 7498 7499 7500 7501
7502 7503 7504 7505 7506 7507 7508 7509 7510 7511
7512 7513 7514 7515 7516 7517 7518 7519 7520 7521
7522 7523 7524 7525 7526 7527 7528 7529 7530 7531
7532 7533 7534 7535 7536 7537 7538 7539 7540 7541
7542 7543 7544 7545 7546 7547 7548 7549 7550 7551

7552 7553 7554 7555 7556 7557 7558 7559 7560 7561
7562 7563 7564 7565 7566 7567 7568 7569 7570 7571
7572 7573 7574 7575 7576 7577 7578 7579 7580 7581
7582 7583 7584 7585 7586 7587 7588 7589 7590 7591
7592 7593 7594 7595 7596 7597 7598 7599 7600 7601
7602 7603 7604 7605 7606 7607 7608 7609 7610 7611
7612 7613 7614 7615 7616 7617 7618 7619 7620 7621
7622 7623 7624 7625 7626 7627 7628 7629 7630 7631
7632 7633 7634 7635 7636 7637 7638 7639 7640 7641
7642 0100 0101 0102 0103 0104 0105 0106 0107 0108
0109 0110 0111 0112 0113 0114 0115 0116 0117 0118
0119 0120 0121 0122 0123 0124 0125 0126 0127 0128
0129 0130 0131 0132 0133 0134 0135 0136 0137 0138
0139 0140 0141 0142 0143 0144 0145 0146 0147 0148
0149 0150 0151 0152 0153 0154 0155 0156 0157 0158
0159 0160 0161 0162 0163 0164 0165 0166 0167 0168
0169 0170 0171 0172 0173 0174 0175 0176 0177 0178
0179 0180 0181 0182 0183 0184 0185 0186 0187 0188
0189 0190 0191 0192 0193 0194 0195 0196 0197 0198
0199 0200 0201 0202 0203 0204 0205 0206 0207 0208
0209 0210 0211 0212 0213 0214 0215 0216 0217 0218
0219 0220 0221 0222 0223 0224 0225 0226 0227 0228
0229 0230 0231 0232 0233 0234 0235 0236 0237 0238
0239 0240 0241 0242 0243 0244 0245 0246 0247 0248
0249 0250 0251 0252 0253 0254 0255 0256 0257 0258
0259 0260 0261 0262 0263 0264 0265 0266 0267 0268
0269 0270 0271 0272 0273 0274 0275 0276 0277 0278
0279 0280 0281 0282 0283 0284 0285 0286 0287 0288
0289 0290 0291 0292 0293 0294 0295 0296 0297 0298
0299 0300 0301 0302 0303 0304 0305 0306 0307 0308
0309 0310 0311 0312 0313 0314 0315 0316 0317 0318
0319 0320 0321 0322 0323 0324 0325 0326 0327 0328
0329 0330 0331 0332 0333 0334 0335 0336 0337 0338
0339 0340 0341 0342 0343 0344 0345 0346 0347 0348
0349 0350 0351 0352 0353 0354 0355 0356 0357 0358
0359 0360 0361 0362 0363 0364 0365 0366 0367 0368
0369 0370 0371 0372 0373 0374 0375 0376 0377 0378
0379 0038 0380 0381 0382 0383 0384 0385 0386 0387

0388 0389 0039 0390 0391 0392 0393 0394 0395 0396
0397 0398 0399 0040 0400 0401 0402 0403 0404 0405
0406 0407 0408 0409 0041 0410 0411 0412 0413 0414
0415 0416 0417 0418 0419 0042 0420 0421 0422 0423
0424 0425 0426 0427 0428 0429 0043 0430 0431 0432
0433 0434 0435 0436 0437 0438 0439 0044 0440 0441
0442 0443 0444 0445 0446 0447 0448 0449 0045 0450
0451 0452 0453 0454 0455 0456 0457 0458 0459 0046
0460 0461 0462 0463 0464 0465 0466 0467 0468 0469
0047 0470 0471 0472 0473 0474 0475 0476 0477 0478
0479 0048 0480 0481 0482 0483 0484 0485 0486 0487
0488 0489 0049 0490 0491 0492 0493 0494 0495 0496
0497 0498 0499 0050 0500 0501 0502 0503 0504 0505
0506 0507 0508 0509 0051 0510 0511 0512 0513 0514
0515 0516 0517 0518 0519 0052 0520 0521 0522 0523
0524 0525 0526 0527 0528 0529 0053 0530 0531 0532
0533 0534 0535 0536 0537 0538 0539 0054 0540 0541
0542 0543 0544 0545 0546 0547 0548 0549 0055 0550
0551 0552 0553 0554 0555 0556 0557 0558 0559 0056
0560 0561 0562 0057 0058 0059 0060 0061 0062 0063
0064 0065 0066 0067 0068 0069 0070 0071 0072 0073
0074 0075 0076 0077 0078 0079 0080 0081 0082 0083
0084 0085 0086 0087 0088 0089 0090 0091 0092 0093
0094 0095 0096 0097 0098 0099 1000 1001 1002 1003
1004 1005 1006 1007 1008 1009 1010 1011 1012 1013
1014 1015 1016 1017 1018 1019 1020 1021 1022 1023
1024 1025 1026 1027 1028 1029 1030 1031 1032 1033
1034 1035 1036 1037 1038 1039 1040 1041 1042 1043
1044 1045 1046 1047 1048 1049 1050 1051 1052 1053
1054 1055 1056 1057 1058 1059 1060 1061 1062 1063
1064 1065 1066 1067 1068 1069 1070 1071 1072 1073
1074 1075 1076 1077 1078 1079 1080 1081 1082 0775
0776 0777 0778 0779 0780 0781 0782 0783 0784 0785
0786 0787 0788 0789 0790 0791 0792 0793 0794 0795
0796 0797 0798 0799 0800 0801 0802 0803 0804 0805
0806 0807 0808 0809 0810 0811 0812 0813 0814 0815
0816 0817 0818 0819 0820 0821 0822 0823 0824 0825
0826 0827 0828 0829 0830 0831 0832 0833 0834 0835

0836 0837 0838 0839 0840 0841 0842 0843 0844 0845
0846 0847 0848 0849 0850 0851 0852 0853 0854 0855
0856 0857 0858 0859 0860 0861 0862 0863 0864 0865
0866 0867 0868 0869 0870 0871 0872 0873 0874 0875
0876 0877 0878 0879 0880 0881 0882 0883 0884 0885
0886 0887 0888 0889 0890 0891 0892 0893 0894 0895
0896 0897 0898 0899 0900 0901 0902 0903 0904 0905
0906 0907 0908 0909 0910 0911 0912 0913 0914 0915
0916 0917 0918 0919 0920 0921 0922 0923 0924 0925
0926 0927 0928 0929 0930 0931 0932 0933 0934 0935
0936 0937 0938 0939 0940 0941 0942 0943 0944 0945
0946 0947 0948 0949 0950 0951 0952 0953 0954 0955
0956 0957 0958 0959 0960 0961 0962 0963 0964 0965
0966 0967 0968 0969 0970 0971 0972 0973 0974 0975
0976 0977 0978 0979 0980 0981 0982 0983 0984 0985
0986 0987 0988 0989 0990 0991 0992 0993 0994 0995
0996 0997 0998 0999 7643 7644 7645 7646 7647 7648
7649 7650 7651 7652 7653 7654 7655 7656 7657 7658
7659 7660 7661 7662 7663 7664 7665 7666 7667 7668
7669 7670 7671 7672 7673 7674 7675 7676 7677 7678
7679 7680 7681 7682 7683 7684 7685 7686 7687 7688
7689 7690 7691 7692 7693 7694 7695 7696 7697 7698
7699 7700 7701 7702 7703 7704 7705 7706 7707 7708
7709 7710 7711 7712 7713 7714 7715 7716 7717 7718
7719 7720 7721 7722 7723 7724 7725 7726 7727 7728
7729 7730 7731 7732 7733 7734 7735 7736 7737 7738
7739 7740 7741 7742 7743 7744 7745 7746 7747 7748
7749 7750 7751 7752 7753 7754 7755 7756 7757 7758
7759 7760 7761 7762 7763 7764 7765 7766 7767 7768
7769 7770 7771 7772 7773 7774 7775 7776 7777 7778
7779 7780 7781 7782 7783 7784 7785 7786 7787 7788
7789 7790 7791 7792 7793 7794 7795 7796 7797 7798
7799 7800 7801 7802 7803 7804 7805 7806 7807 7808
7809 7810 7811 7812 7813 7814 7815 7816 7817 7818
7819 7820 7821 7822 7823 7824 7825 7826 7827 7828
7829 7830 7831 7832 7833 7834 7835 7836 7837 7838
7839 7840 7841 7842 0001 0010 0011 0012 0013 0014
0015 0016 0017 0018 0019 0002 0020 0021 0022 0023

0024 0025 0026 0027 0028 0029 0003 0030 0031 0032
0033 0034 0035 0036 0037 0038 0039 0004 0040 0041
0042 0043 0044 0045 0046 0047 0048 0049 0005 0050
0051 0052 0053 0054 0055 0056 0057 0006 0007 0008
0009 4581 4582 4583 4584 4585 4586 4587 4588 4589
4590 4591 4592 4593 4594 4595 4596 4597 4598 4599
4600 4601 4602 4603 4604 4605 4606 4607 4608 4609
4610 4611 4612 4613 4614 4615 4616 4617 4618 4619
4620 4621 4622 4623 4624 4625 4626 4627 4628 4629
4630 4631 4632 4633 4634 4635 4636 4637 4638 4639
4640 4641 4642 4643 4644 4645 4646 4647 4648 4649
4650 4651 4652 4653 4654 4655 4656 4657 4658 4659
4660 4661 4662 4663 4664 4665 4666 4667 4668 4669
4670 4671 4672 4673 4674 4675 4676 4677 4678 4679
4680 4681 4682 4683 4684 4685 4686 4687 4688 4689
4690 4691 4692 4693 4694 4695 4696 4697 4698 4699
4700 4701 4702 4703 4704 4705 4706 4707 4708 4709
4710 4711 4712 4713 4714 4715 4716 4717 4718 4719
4720 4721 4722 4723 4724 4725 4726 4727 4728 4729
4730 4731 4732 4733 4734 4735 4736 4737 4738 4739
4740 4741 4742 4743 4744 4745 4746 4747 4748 4749
4750 4751 4752 4753 4754 4755 4756 4757 4758 4759
4760 4761 4762 4763 4764 4765 4766 4767 4768 4769
4770 4771 4772 4773 4774 4775 4776 4777 4778 4779
4780 4781 4782 4783 4784 4785 4786 4787 4788 4789
4790 4791 4792 4793 4794 0563 0564 0565 0566 0567
0568 0569 0570 0571 0572 0573 0574 0575 0576 0577
0578 0579 0580 0581 0582 0583 0584 0585 0586 0587
0588 0589 0590 0591 0592 0593 0594 0595 0596 0597
0598 0599 0600 0601 0602 0603 0604 0605 0606 0607
0608 0609 0610 0611 0612 0613 0614 0615 0616 0617
0618 0619 0620 0621 0622 0623 0624 0625 0626 0627
0628 0629 0630 0631 0632 0633 0634 0635 0636 0637
0638 0639 0640 0641 0642 0643 0644 0645 0646 0647
0648 0649 0650 0651 0652 0653 0654 0655 0656 0657
0658 0659 0660 0661 0662 0663 0664 0665 0666 0667
0668 0669 0670 0671 0672 0673 0674 0675 0676 0677
0678 0679 0680 0681 0682 0683 0684 0685 0686 0687

0688 0689 0690 0691 0692 0693 0694 0695 0696 0697
0698 0699 0700 0701 0702 0703 0704 0705 0706 0707
0708 0709 0710 0711 0712 0713 0714 0715 0716 0717
0718 0719 0720 0721 0722 0723 0724 0725 0726 0727
0728 0729 0730 0731 0732 0733 0734 0735 0736 0737
0738 0739 0740 0741 0742 0743 0744 0745 0746 0747
0748 0749 0750 0751 0752 0753 0754 0755 0756 0757
0758 0759 0760 0761 0762 0763 0764 0765 0766 0767
0768 0769 0770 0771 0772 0773 0774 3454 3455 3456
3457 3458 3459 3460 3461 3462 3463 3464 3465 3466
3467 3468 3469 3470 3471 3472 3473 3474 3475 3476
3477 3478 3479 3480 3481 3482 3483 3484 3485 3486
3487 3488 3489 3490 3491 3492 3493 3494 3495 3496
3497 3498 3499 3500 3501 3502 3503 3504 3505 3506
3507 3508 3509 3510 3511 3512 3513 3514 3515 3516
3517 3518 3519 3520 3521 3522 3523 3524 3525 3526
3527 3528 3529 3530 3531 3532 3533 3534 3535 3536
3537 3538 3539 3540 3541 3542 3543 3544 3545 3546
3547 3548 3549 3550 3551 3552 3553 3554 3555 3556
3557 3558 3559 3560 3561 3562 3563 3564 3565 3566
3567 3568 3569 3570 3571 3572 3573 3574 3575 3576
3577 3578 3579 3580 3581 3582 3583 3584 3585 3586
3587 3588 3589 3590 3591 3592 3593 3594 3595 3596
3597 3598 3599 3600 3601 3602 3603 3604 3605 3606
3607 3608 3609 3610 3611 3612 3613 3614 3615 3616
3617 3618 3619 3620 3621 3622 3623 3624 3625 3626
3627 3628 3629 3630 3631 3632 3633 3634 3635 3636
3637 3638 3639 3640 3641 3642 3643 3644 3645 3646
3647 3648 3649 3650 3651 3652 3653 3654 3655 3656
3657 3658 3659 3660 3661 3662 3663 3664 3665 3666
3667 3668 3669 3670 3671 3672 3673 3674 3675 3676
3677 3678 3679 3680 3681 3682 1683 1684 1685 1686
1687 1688 1689 1690 1691 1692 1693 1694 1695 1696
1697 1698 1699 1700 1701 1702 1703 1704 1705 1706
1707 1708 1709 1710 1711 1712 1713 1714 1715 1716
1717 1718 1719 1720 1721 1722 1723 1724 1725 1726
1727 1728 1729 1730 1731 1732 1733 1734 1735 1736
1737 1738 1739 1740 1741 1742 1743 1744 1745 1746

1747 1748 1749 1750 1751 1752 1753 1754 1755 1756
1757 1758 1759 1760 1761 1762 1763 1764 1765 1766
1767 1768 1769 1770 1771 1772 1773 1774 1775 1776
1777 1778 1779 1780 1781 1782 1783 1784 1785 1786
1787 1788 1789 1790 1791 1792 1793 1794 1795 1796
1797 1798 1799 1800 1801 1802 1803 1804 1805 1806
1807 1808 1809 1810 1811 1812 1813 1814 1815 1816
1817 1818 1819 1820 1821 1822 1823 1824 1825 1826
1827 1828 1829 1830 1831 1832 1833 1834 1835 1836
1837 1838 1839 1840 1841 1842 1843 1844 1845 1846
1847 1848 1849 1850 1851 1852 1853 1854 1855 1856
1857 1858 1859 1860 1861 1862 1863 1864 1865 1866
1867 1868 1869 1870 1871 1872 1873 1874 1875 1876
1877 1878 1879 1880 1881 1882 1883 1884 1885 1886
1887 1888 1889 1890 1891 1892 1893 1894 1895 1896
1897 1898 1899 1900 1901 1902 1903 1904 1905 1906
1907 1908 1909 1910 1911 1912 1913 6320 6321 6322
6323 6324 6325 6326 6327 6328 6329 6330 6331 6332
6333 6334 6335 6336 6337 6338 6339 6340 6341 6342
6343 6344 6345 6346 6347 6348 6349 6350 6351 6352
6353 6354 6355 6356 6357 6358 6359 6360 6361 6362
6363 6364 6365 6366 6367 6368 6369 6370 6371 6372
6373 6374 6375 6376 6377 6378 6379 6380 6381 6382
6383 6384 6385 6386 6387 6388 6389 6390 6391 6392
6393 6394 6395 6396 6397 6398 6399 6400 6401 6402
6403 6404 6405 6406 6407 6408 6409 6410 6411 6412
6413 6414 6415 6416 6417 6418 6419 6420 6421 6422
6423 6424 6425 6426 6427 6428 6429 6430 6431 6432
6433 6434 6435 6436 6437 6438 6439 6440 6441 6442
6443 6444 6445 6446 6447 6448 6449 6450 6451 6452
6453 6454 6455 6456 6457 6458 6459 6460 6461 6462
6463 6464 6465 6466 6467 6468 6469 6470 6471 6472
6473 6474 6475 6476 6477 6478 6479 6480 6481 6482
6483 6484 6485 6486 6487 6488 6489 6490 6491 6492
6493 6494 6495 6496 6497 6498 6499 6500 6501 6502
6503 6504 6505 6506 6507 6508 6509 6510 6511 6512
6513 6514 6515 6516 6517 6518 6519 6520 6521 6522
6523 6524 6525 6526 6527 6528 6529 6530 6531 6532

6533 6534 6535 6536 6537 6538 6539 6540 0100 0101
0102 0103 0104 0105 0106 0107 0108 0109 0110 0111
0112 0113 0114 0115 0116 0117 0118 0119 0120 0121
0122 0123 0124 0125 0126 0127 0128 0129 0130 0131
0132 0133 0134 0135 0136 0137 0138 0139 0140 0141
0142 0143 0144 0145 0146 0147 0148 0149 0150 0151
0152 0153 0154 0155 0156 0157 0158 0159 0160 0161
0162 0163 0164 0165 0166 0167 0168 0169 0170 0171
0172 0173 0174 0175 0176 0177 0178 0179 0180 0181
0182 0183 0184 0185 0186 0187 0188 0189 0190 0191
0192 0193 0194 0195 0196 0197 0198 0199 0200 0201
0202 0203 0204 0205 0206 0207 0208 0209 0210 0211
0212 0213 0214 0215 0216 0217 0218 0219 0220 0221
0222 0223 0224 0225 0226 0227 0228 0229 0230 0231
0232 0233 0234 0235 0236 0237 0238 0239 0240 0241
0242 0243 0244 0245 0246 0247 0248 0249 0250 0251
0252 0253 0254 0255 0256 0257 0258 0259 0260 0261
0262 0263 0264 0265 0266 0267 0268 0269 0270 0271
0272 0273 0274 0275 0276 0277 0278 0279 0280 0281
0282 0283 0284 0285 0286 0287 0288 0289 0290 0291
0292 0293 0294 0295 0296 0297 0298 0299 0300 0301
0302 0303 0304 0305 0306 0307 0308 0309 0310 0311
0312 0313 0314 0315 0316 0317 0318 0319 0320 0321
0322 0323 0324 0325 0326 0327 0328 0329 0330 0331
0332 0333 0334 0335 0336 0337 0338 0339 0340 0341
0342 0343 0344 0345 0346 0347 0348 0349 0350 0351
0352 0353 0354 0355 0356 0357 0358 0359 0360 0361
0362 0058 0059 0060 0061 0062 0063 0064 0065 0066
0067 0068 0069 0070 0071 0072 0073 0074 0075 0076
0077 0078 0079 0080 0081 0082 0083 0084 0085 0086
0087 0088 0089 0090 0091 0092 0093 0094 0095 0096
0097 0098 0099 5458 5459 5460 5461 5462 5463 5464
5465 5466 5467 5468 5469 5470 5471 5472 5473 5474
5475 5476 5477 5478 5479 5480 5481 5482 5483 5484
5485 5486 5487 5488 5489 5490 5491 5492 5493 5494
5495 5496 5497 5498 5499 5500 5501 5502 5503 5504
5505 5506 5507 5508 5509 5510 5511 5512 5513 5514
5515 5516 5517 5518 5519 5520 5521 5522 5523 5524

5525 5526 5527 5528 5529 5530 5531 5532 5533 5534
5535 5536 5537 5538 5539 5540 5541 5542 5543 5544
5545 5546 5547 5548 5549 5550 5551 5552 5553 5554
5555 5556 5557 5558 5559 5560 5561 5562 5563 5564
5565 5566 5567 5568 5569 5570 5571 5572 5573 5574
5575 5576 5577 5578 5579 5580 5581 5582 5583 5584
5585 5586 5587 5588 5589 5590 5591 5592 5593 5594
5595 5596 5597 5598 5599 5600 5601 5602 5603 5604
5605 5606 5607 5608 5609 5610 5611 5612 5613 5614
5615 5616 5617 5618 5619 5620 5621 5622 5623 5624
5625 5626 5627 5628 5629 5630 5631 5632 5633 5634
5635 5636 5637 5638 5639 5640 5641 5642 5643 5644
5645 5646 5647 5648 5649 5650 5651 5652 5653 5654
5655 5656 5657 5658 5659 5660 5661 5662 5663 5664
5665 5666 5667 5668 5669 5670 5671 5672 5673 5674
5244 5245 5246 5247 5248 5249 5250 5251 5252 5253
5254 5255 5256 5257 5258 5259 5260 5261 5262 5263
5264 5265 5266 5267 5268 5269 5270 5271 5272 5273
5274 5275 5276 5277 5278 5279 5280 5281 5282 5283
5284 5285 5286 5287 5288 5289 5290 5291 5292 5293
5294 5295 5296 5297 5298 5299 5300 5301 5302 5303
5304 5305 5306 5307 5308 5309 5310 5311 5312 5313
5314 5315 5316 5317 5318 5319 5320 5321 5322 5323
5324 5325 5326 5327 5328 5329 5330 5331 5332 5333
5334 5335 5336 5337 5338 5339 5340 5341 5342 5343
5344 5345 5346 5347 5348 5349 5350 5351 5352 5353
5354 5355 5356 5357 5358 5359 5360 5361 5362 5363
5364 5365 5366 5367 5368 5369 5370 5371 5372 5373
5374 5375 5376 5377 5378 5379 5380 5381 5382 5383
5384 5385 5386 5387 5388 5389 5390 5391 5392 5393
5394 5395 5396 5397 5398 5399 5400 5401 5402 5403
5404 5405 5406 5407 5408 5409 5410 5411 5412 5413
5414 5415 5416 5417 5418 5419 5420 5421 5422 5423
5424 5425 5426 5427 5428 5429 5430 5431 5432 5433
5434 5435 5436 5437 5438 5439 5440 5441 5442 5443
5444 5445 5446 5447 5448 5449 5450 5451 5452 5453
5454 5455 5456 5457 5019 5020 5021 5022 5023 5024
5025 5026 5027 5028 5029 5030 5031 5032 5033 5034

5035 5036 5037 5038 5039 5040 5041 5042 5043 5044
5045 5046 5047 5048 5049 5050 5051 5052 5053 5054
5055 5056 5057 5058 5059 5060 5061 5062 5063 5064
5065 5066 5067 5068 5069 5070 5071 5072 5073 5074
5075 5076 5077 5078 5079 5080 5081 5082 5083 5084
5085 5086 5087 5088 5089 5090 5091 5092 5093 5094
5095 5096 5097 5098 5099 5100 5101 5102 5103 5104
5105 5106 5107 5108 5109 5110 5111 5112 5113 5114
5115 5116 5117 5118 5119 5120 5121 5122 5123 5124
5125 5126 5127 5128 5129 5130 5131 5132 5133 5134
5135 5136 5137 5138 5139 5140 5141 5142 5143 5144
5145 5146 5147 5148 5149 5150 5151 5152 5153 5154
5155 5156 5157 5158 5159 5160 5161 5162 5163 5164
5165 5166 5167 5168 5169 5170 5171 5172 5173 5174
5175 5176 5177 5178 5179 5180 5181 5182 5183 5184
5185 5186 5187 5188 5189 5190 5191 5192 5193 5194
5195 5196 5197 5198 5199 5200 5201 5202 5203 5204
5205 5206 5207 5208 5209 5210 5211 5212 5213 5214
5215 5216 5217 5218 5219 5220 5221 5222 5223 5224
5225 5226 5227 5228 5229 5230 5231 5232 5233 5234
5235 5236 5237 5238 5239 5240 5241 5242 5243 1715
1716 1717 1718 2137 2138 2139 2140 2141 2142 2143
2144 2145 2146 2147 2148 2149 2150 2151 2152 2153
2154 2157 2158 2159 2160 2161 2162 2163 2164 2165
2166 2167 2168 2169 2170 2171 2172 2173 2174 2175
2176 2177 2178 2179 2180 2181 2182 2183 2184 2185
2186 2187 2188 2189 2190 2191 2192 2193 2194 2195
2196 2197 2198 2199 2200 2201 2202 2203 2204 2205
2206 2207 2208 2209 2210 2211 2212 2213 2214 2215
2216 2217 2218 2219 2220 2221 2222 2223 2235 2236
2237 2238 1246 1247 1248 1249 1250 1251 1252 1253
1254 1255 1256 1257 1258 1259 1260 1261 1262 1263
1264 1265 1266 1267 1268 1269 1270 1271 1272 1273
1274 1275 1276 1277 1278 1279 1280 1281 1282 1283
1284 1285 1286 1287 1288 1289 1290 1291 1292 1293
1294 1295 1296 1297 1298 1299 1300 1301 1302 1303
1304 1305 1306 1307 1308 1309 1310 1311 1312 1313
1314 1315 1316 1317 1318 1319 1320 1321 1322 1323

1324 1325 1326 1327 1328 1329 1330 1331 1334 1335
1336 1337 1338 1339 1340 1341 1342 1343 1344 1345
1346 1347 1348 1349 1350 1351 1352 1353 1354 1355
1356 1357 1358 1359 1360 1361 1362 1363 1364 1365
1366 1367 1368 1369 1370 1371 1372 1373 1374 1375
1376 1377 1378 1379 1380 1381 1382 1383 1384 1385
1386 1387 1388 1389 1390 1391 1392 1393 1394 1395
1396 1397 1398 1399 1400 1401 1402 1403 1404 1405
1406 1407 1408 1409 1410 1411 1412 1413 1414 1415
1416 1417 1418 1419 1420 1421 1422 1423 1424 1425
1426 1427 1428 1429 1430 1431 1432 1433 1434 1435
1436 1437 1438 1439 1440 1441 1442 1443 1444 1445
1446 1447 1448 1449 1450 1451 1452 1453 1454 1455
1456 1457 1458 1459 1460 0001 0010 0011 0012 0013
0014 0015 0016 0017 0018 0019 0002 0020 0021 0022
0023 0024 0025 0026 0027 0028 0029 0003 0030 0031
0032 0033 0034 0035 0036 0037 0004 0005 0006 0007
0008 0009 5893 5894 5895 5896 5897 5898 5899 5900
5901 5902 5903 5904 5905 5906 5907 5908 5909 5910
5911 5912 5913 5914 5915 5916 5917 5918 5919 5920
5921 5922 5923 5924 5925 5926 5927 5928 5929 5930
5931 5932 5933 5934 5935 5936 5937 5938 5939 5940
5941 5942 5943 5944 5945 5946 5947 5948 5949 5950
5951 5952 5953 5954 5955 5956 5957 5958 5959 5960
5961 5962 5963 5964 5965 5966 5967 5968 5969 5970
5971 5972 5973 5974 5975 5976 5977 5978 5979 5980
5981 5982 5983 5984 5985 5986 5987 5988 5989 5990
5991 5992 5993 5994 5995 5996 5997 5998 5999 6000
6001 6002 6003 6004 6005 6006 6007 6008 6009 6010
6011 6012 6013 6014 6015 6016 6017 6018 6019 6020
6021 6022 6023 6024 6025 6026 6027 6028 6029 6030
6031 6032 6033 6034 6035 6036 6037 6038 6039 6040
6041 6042 6043 6044 6045 6046 6047 6048 6049 6050
6051 6052 6053 6054 6055 6056 6057 6058 6059 6060
6061 6062 6063 6064 6065 6066 6067 6068 6069 6070
6071 6072 6073 6074 6075 6076 6077 6078 6079 6080
6081 6082 6083 6084 6085 6086 6087 6088 6089 6090
6091 6092 6093 6094 6095 6096 6097 6098 6099 6100

6101 6102 6103 6104 6105 6106 6107 6108 6109 6110
6111 6112 6113 6114 2939 2940 2941 2942 2943 2944
2945 2946 2947 2948 2949 2950 2951 2952 2953 2954
2955 2956 2957 2958 2959 2960 2961 2962 2963 2964
2965 2966 2967 2968 2969 2970 2971 2972 2973 2974
2975 2976 2977 2978 2979 2980 2981 2982 2983 2984
2985 2986 2987 2988 2989 2990 2991 2992 2993 2994
2995 2996 2997 2998 2999 3000 3001 3002 3003 3004
3005 3006 3007 3008 3009 3010 3011 3012 3013 3014
3015 3016 3017 3018 3019 3020 3021 3022 3023 3024
3025 3026 3027 3028 3029 3030 3031 3032 3033 3034
3035 3036 3037 3038 3039 3040 3041 3042 3043 3044
3045 3046 3047 3048 3049 3050 3051 3052 3053 3054
3055 3056 3057 3058 3059 3060 3061 3062 3063 1026
1027 1028 1029 1030 1031 1032 1033 1034 1035 1036
1037 1038 1039 1040 1041 1042 1043 1044 1045 1046
1047 1048 1049 1050 1051 1052 1053 1054 1055 1056
1057 1058 1059 1060 1061 1062 1063 1064 1065 1066
1067 1068 1069 1070 1071 1072 1073 1074 1075 1076
1077 1078 1079 1080 1081 1082 1083 1153 1154 1155
1156 1157 1158 1159 1160 1161 1162 1163 1164 1165
1166 1167 1168 1169 1170 1171 1172 1173 1174 1175
1176 1177 1178 1179 1180 1181 1182 1183 1184 1185
1186 1187 1188 1189 1190 1191 1192 1193 1194 1195
1196 1197 1198 1199 1200 1201 1202 1203 1204 1205
1206 1207 1208 1209 1210 1211 1212 1213 1214 1215
1216 1217 1218 1219 1220 1221 1222 1223 1224 1225
1226 1227 1228 1229 1230 1231 1232 1233 1234 1235
1236 1237 1238 1239 1240 1241 1242 1243 1244 1245
0591 0592 0593 0594 0595 0596 0597 0598 0599 0600
0601 0602 0603 0604 0605 0606 0607 0608 0609 0610
0611 0612 0613 0614 0615 0616 0617 0618 0619 0620
0621 0622 0623 0624 0625 0626 0627 0628 0629 0630
0631 0632 0633 0634 0635 0636 0637 0638 0639 0640
0641 0642 0643 0644 0645 0646 0647 0648 0649 0650
0651 0652 0653 0654 0655 0656 0657 0658 0659 0660
0661 0662 0663 0664 0665 0666 0667 0668 0669 0670
0671 0672 0673 0674 0675 0676 0677 0678 0679 0680

0681 0682 0683 0684 0685 0686 0687 0688 0689 0690
0691 0692 0693 0694 0695 0696 0697 0698 0699 0700
0701 0702 0703 0704 0705 0706 0707 0708 0709 0710
0711 0712 0713 0714 0715 0716 0717 0718 0719 0720
0721 0722 0723 0724 0725 0726 0727 0728 0729 0730
0731 0732 0733 0734 0735 0736 0737 0738 0739 0740
0741 0742 0743 0744 0745 0746 0747 0748 0749 0750
0751 0752 0753 0754 0755 0756 0757 0758 0759 0760
0761 0762 0763 0764 0765 0766 0767 0768 0769 0770
0771 0772 0773 0774 0775 0776 0777 0778 0779 0780
0781 0782 0783 0784 0785 0786 0787 0788 0789 0790
0791 0792 0793 0794 0795 0796 0797 0798 0799 0800
0801 0802 0803 0804 0805 0806 0807 0808 0809 6541
6542 6543 6544 6545 6546 6547 6548 6549 6550 6551
6552 6553 6554 6555 6556 6557 6558 6559 6560 6561
6562 6563 6564 6565 6566 6567 6568 6569 6570 6571
6572 6573 6574 6575 6576 6577 6578 6579 6580 6581
6582 6583 6584 6585 6586 6587 6588 6589 6590 6591
6592 6593 6594 6595 6596 6597 6598 6599 6600 6601
6602 6603 6604 6605 6606 6607 6608 6609 6610 6611
6612 6613 6614 6615 6616 6617 6618 6619 6620 6621
6622 6623 6624 6625 6626 6627 6628 6629 6630 6631
6632 6633 6634 6635 6636 6637 6638 6639 6640 6641
6642 6643 6644 6645 6646 6647 6648 6649 6650 6651
6652 6653 6654 6655 6656 6657 6658 6659 6660 6661
6662 6663 6664 6665 6666 6667 6668 6669 6670 6671
6672 6673 6674 6675 6676 6677 6678 6679 6680 6681
6682 6683 6684 6685 6686 6687 6688 6689 6690 6691
6692 6693 6694 6695 6696 6697 6698 6699 6700 6701
6702 6703 6704 6705 6706 6707 6708 6709 6710 6711
6712 6713 6714 6715 6716 6717 6718 6719 6720 6721
6722 6723 6724 6725 6726 6727 6728 6729 6730 6731
6732 6733 6734 6735 6736 6737 6738 6739 6740 6741
6742 6743 6744 6745 6746 6747 6748 6749 6750 6751

# Appendix C

# Test Data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1411 | 1412 | 1413 | 1414 | 1415 | 1416 | 1417 | 1418 | 1419 | 1420 |
| 1421 | 1422 | 1423 | 1424 | 1425 | 1426 | 1427 | 1428 | 1429 | 1430 |
| 1431 | 1432 | 1433 | 1434 | 1435 | 1436 | 1437 | 1438 | 1439 | 1440 |
| 1441 | 1442 | 1443 | 1444 | 1445 | 1446 | 1447 | 1448 | 1449 | 1450 |
| 1451 | 1452 | 1453 | 1454 | 1455 | 1456 | 1457 | 1458 | 1459 | 1460 |
| 1461 | 1462 | 1463 | 1464 | 1465 | 1466 | 1467 | 1468 | 1469 | 1470 |
| 1471 | 1472 | 1473 | 1474 | 1475 | 1476 | 1477 | 1478 | 1479 | 1480 |
| 1481 | 1482 | 1483 | 1484 | 1485 | 1486 | 1487 | 1488 | 1489 | 1490 |
| 1491 | 1492 | 1493 | 1494 | 1495 | 1496 | 1497 | 1498 | 1499 | 1500 |
| 1501 | 1502 | 1503 | 1504 | 1505 | 1506 | 1507 | 1508 | 1509 | 1510 |
| 1511 | 1512 | 1513 | 1514 | 1515 | 1516 | 1517 | 1518 | 1519 | 1520 |
| 1521 | 1522 | 1523 | 1524 | 1525 | 1526 | 1527 | 1528 | 1529 | 1530 |
| 1531 | 1532 | 1533 | 1534 | 1535 | 1536 | 1537 | 1538 | 1539 | 1540 |
| 1541 | 1542 | 1543 | 1544 | 1545 | 1546 | 1547 | 1548 | 1549 | 1550 |
| 1551 | 1552 | 1553 | 1554 | 1555 | 1556 | 1557 | 1558 | 1559 | 1560 |
| 1561 | 1562 | 1563 | 1564 | 1565 | 1566 | 1567 | 1568 | 1569 | 1570 |
| 1571 | 1572 | 1573 | 1574 | 1575 | 1576 | 1577 | 1578 | 1579 | 1580 |
| 1581 | 1582 | 1583 | 1584 | 1585 | 1586 | 1587 | 1588 | 1589 | 1590 |
| 1591 | 1592 | 1593 | 1594 | 1595 | 1596 | 1597 | 1598 | 1599 | 1600 |
| 1601 | 1602 | 1603 | 1604 | 1605 | 1606 | 1607 | 1608 | 1609 | 1610 |
| 1611 | 1612 | 1613 | 1614 | 1615 | 1616 | 1617 | 1618 | 1619 | 1620 |
| 1621 | 1622 | 1623 | 1624 | 1625 | 1626 | 1627 | 1628 | 1629 | 1630 |
| 1631 | 1632 | 1633 | 1634 | 1635 | 1636 | 1637 | 1638 | 1639 | 1640 |
| 1641 | 1642 | 1643 | 1644 | 1645 | 1646 | 1647 | 1648 | 1649 | 1650 |
| 1651 | 1652 | 1653 | 1654 | 1655 | 1656 | 1657 | 1658 | 1659 | 1660 |

1661 1662 1663 1664 1665 1666 1667 1668 1669 1670
1671 1672 1673 1674 1675 1676 1677 1678 1679 1680
1681 1682 1683 1684 1685 1686 1687 1688 1689 1690
1691 1692 1693 1694 1695 1696 1697 1698 1699 1700
1701 1702 1703 1704 1705 1706 1707 1708 1709 1710
1711 1712 1083 1084 1085 1086 1087 1088 1089 1090
1091 1092 1093 1094 1095 1096 1097 1098 1099 1100
1101 1102 1103 1104 1105 1106 1107 1108 1109 1110
1111 1112 1113 1114 1115 1116 1117 1118 1119 1120
1121 1122 1123 1124 1125 1126 1127 1128 1129 1130
1131 1132 1133 1134 1135 1136 1137 1138 1139 1140
1141 1142 1143 1144 1145 1146 1147 1148 1149 1150
1151 1152 1153 1154 1155 1156 1157 1158 1159 1160
1161 1162 1163 1164 1165 1166 1167 1168 1169 1170
1171 1172 1173 1174 1175 1176 1177 1178 1179 1180
1181 1182 1183 1184 1185 1186 1187 1188 1189 1190
1191 1192 1193 1194 1195 1196 1197 1198 1199 1200
1201 1202 1203 1204 1205 1206 1207 1208 1209 1210
1211 1212 1213 1214 1215 1216 1217 1218 1219 1220
1221 1222 1223 1224 1225 1226 1227 1228 1229 1230
1231 1232 1233 1234 1235 1236 1237 1238 1239 1240
1241 1242 1243 1244 1245 1246 1247 1248 1249 1250
1251 1252 1253 1254 1255 1256 1257 1258 1259 1260
1261 1262 1263 1264 1265 1266 1267 1268 1269 1270
1271 1272 1273 1274 1275 1276 1277 1278 1279 1280
1281 1282 1283 1284 1285 1286 1287 1288 1289 1290
1291 1292 1293 1294 1295 1296 1297 1298 1299 1300
1301 1302 1303 1304 1305 1306 1307 1308 1309 1310
1311 1312 1313 1314 1315 1316 1317 1318 1319 1320
1321 1322 1323 1324 1325 1326 1327 1328 1329 1330
1331 1332 1333 1334 1335 1336 1337 1338 1339 1340
1341 1342 1343 1344 1345 1346 1347 1348 1349 1350
1351 1352 1353 1354 1355 1356 1357 1358 1359 1360
1361 1362 1363 1364 1365 1366 1367 1368 1369 1370
1371 1372 1373 1374 1375 1376 1377 1378 1379 1380
1381 1382 1383 1384 1385 1386 1387 1388 1389 1390
1391 1392 1393 1394 1395 1396 1397 1398 1399 1400
1401 1402 1403 1404 1405 1406 1407 1408 1409 1410

6115 6116 6117 6118 6119 6120 6121 6122 6123 6124
6125 6126 6127 6128 6129 6130 6131 6132 6133 6134
6135 6136 6137 6138 6139 6140 6141 6142 6143 6144
6145 6146 6147 6148 6149 6150 6151 6152 6153 6154
6155 6156 6157 6158 6159 6160 6161 6162 6163 6164
6165 6166 6167 6168 6169 6170 6171 6172 6173 6174
6175 6176 6177 6178 6179 6180 6181 6182 6183 6184
6185 6186 6187 6188 6189 6190 6191 6192 6193 6194
6195 6196 6197 6198 6199 6200 6201 6202 6203 6204
6205 6206 6207 6208 6209 6210 6211 6212 6213 6214
6215 6216 6217 6218 6219 6220 6221 6222 6223 6224
6225 6226 6227 6228 6229 6230 6231 6232 6233 6234
6235 6236 6237 6238 6239 6240 6241 6242 6243 6244
6245 6246 6247 6248 6249 6250 6251 6252 6253 6254
6255 6256 6257 6258 6259 6260 6261 6262 6263 6264
6265 6266 6267 6268 6269 6270 6271 6272 6273 6274
6275 6276 6277 6278 6279 6280 6281 6282 6283 6284
6285 6286 6287 6288 6289 6290 6291 6292 6293 6294
6295 6296 6297 6298 6299 6300 6301 6302 6303 6304
6305 6306 6307 6308 6309 6310 6311 6312 6313 6314
6315 6316 6317 6318 6319 6979 6980 6981 6982 6983
6984 6985 6986 6987 6988 6989 6990 6991 6992 6993
6994 6995 6996 6997 6998 6999 7000 7001 7002 7003
7004 7005 7006 7007 7008 7009 7010 7011 7012 7013
7014 7015 7016 7017 7018 7019 7020 7021 7022 7023
7024 7025 7026 7027 7028 7029 7030 7031 7032 7033
7034 7035 7036 7037 7038 7039 7040 7041 7042 7043
7044 7045 7046 7047 7048 7049 7050 7051 7052 7053
7054 7055 7056 7057 7058 7059 7060 7061 7062 7063
7064 7065 7066 7067 7068 7069 7070 7071 7072 7073
7074 7075 7076 7077 7078 7079 7080 7081 7082 7083
7084 7085 7086 7087 7088 7089 7090 7091 7092 7093
7094 7095 7096 7097 7098 7099 7100 7101 7102 7103
7104 7105 7106 7107 7108 7109 7110 7111 7112 7113
7114 7115 7116 7117 7118 7119 7120 7121 7122 7123
7124 7125 7126 7127 7128 7129 7130 7131 7132 7133
7134 7135 7136 7137 7138 7139 7140 7141 7142 7143
7144 7145 7146 7147 7148 7149 7150 7151 7152 7153

```
7154 7155 7156 7157 7158 7159 7160 7161 7162 7163
7164 7165 7166 7167 7168 7169 7170 7171 7172 7173
7174 7175 7176 7177 7178 7179 7180 7181 7182 7183
7184 7185 7186 7187 7188 4360 4361 4362 4363 4364
4365 4366 4367 4368 4369 4370 4371 4372 4373 4374
4375 4376 4377 4378 4379 4380 4381 4382 4383 4384
4385 4386 4387 4388 4389 4390 4391 4392 4393 4394
4395 4396 4397 4398 4399 4400 4401 4402 4403 4404
4405 4406 4407 4408 4409 4410 4411 4412 4413 4414
4415 4416 4417 4418 4419 4420 4421 4422 4423 4424
4425 4426 4427 4428 4429 4430 4431 4432 4433 4434
4435 4436 4437 4438 4439 4440 4441 4442 4443 4444
4445 4446 4447 4448 4449 4450 4451 4452 4453 4454
4455 4456 4457 4458 4459 4460 4461 4462 4463 4464
4465 4466 4467 4468 4469 4470 4471 4472 4473 4474
4475 4476 4477 4478 4479 4480 4481 4482 4483 4484
4485 4486 4487 4488 4489 4490 4491 4492 4493 4494
4495 4496 4497 4498 4499 4500 4501 4502 4503 4504
4505 4506 4507 4508 4509 4510 4511 4512 4513 4514
4515 4516 4517 4518 4519 4520 4521 4522 4523 4524
4525 4526 4527 4528 4529 4530 4531 4532 4533 4534
4535 4536 4537 4538 4539 4540 4541 4542 4543 4544
4545 4546 4547 4548 4549 4550 4551 4552 4553 4554
4555 4556 4557 4558 4559 4560 4561 4562 4563 4564
4565 4566 4567 4568 4569 4570 4571 4572 4573 4574
4575 4576 4577 4578 4579 4580 6755 6756 6757 6758
6759 6760 6761 6762 6763 6764 6765 6766 6767 6768
6769 6770 6771 6772 6773 6774 6775 6776 6777 6778
6779 6780 6781 6782 6783 6784 6785 6786 6787 6788
6789 6790 6791 6792 6793 6794 6795 6796 6797 6798
6799 6800 6801 6802 6803 6804 6805 6806 6807 6808
6809 6810 6811 6812 6813 6814 6815 6816 6817 6818
6819 6820 6821 6822 6823 6824 6825 6826 6827 6828
6829 6830 6831 6832 6833 6834 6835 6836 6837 6838
6839 6840 6841 6842 6843 6844 6845 6846 6847 6848
6849 6850 6851 6852 6853 6854 6855 6856 6857 6858
6859 6860 6861 6862 6863 6864 6865 6866 6867 6868
6869 6870 6871 6872 6873 6874 6875 6876 6877 6878
```

6879 6880 6881 6882 6883 6884 6885 6886 6887 6888
6889 6890 6891 6892 6893 6894 6895 6896 6897 6898
6899 6900 6901 6902 6903 6904 6905 6906 6907 6908
6909 6910 6911 6912 6913 6914 6915 6916 6917 6918
6919 6920 6921 6922 6923 6924 6925 6926 6927 6928
6929 6930 6931 6932 6933 6934 6935 6936 6937 6938
6939 6940 6941 6942 6943 6944 6945 6946 6947 6948
6949 6950 6951 6952 6953 6954 6955 6956 6957 6958
6959 6960 6961 6962 6963 6964 6965 6966 6967 6968
6969 6970 6971 6972 6973 6974 6975 6976 6977 6978
4795 4796 4797 4798 4799 4800 4801 4802 4803 4804
4805 4806 4807 4808 4809 4810 4811 4812 4813 4814
4815 4816 4817 4818 4819 4820 4821 4822 4823 4824
4825 4826 4827 4828 4829 4830 4831 4832 4833 4834
4835 4836 4837 4838 4839 4840 4841 4842 4843 4844
4845 4846 4847 4848 4849 4850 4851 4852 4853 4854
4855 4856 4857 4858 4859 4860 4861 4862 4863 4864
4865 4866 4867 4868 4869 4870 4871 4872 4873 4874
4875 4876 4877 4878 4879 4880 4881 4882 4883 4884
4885 4886 4887 4888 4889 4890 4891 4892 4893 4894
4895 4896 4897 4898 4899 4900 4901 4902 4903 4904
4905 4906 4907 4908 4909 4910 4911 4912 4913 4914
4915 4916 4917 4918 4919 4920 4921 4922 4923 4924
4925 4926 4927 4928 4929 4930 4931 4932 4933 4934
4935 4936 4937 4938 4939 4940 4941 4942 4943 4944
4945 4946 4947 4948 4949 4950 4951 4952 4953 4954
4955 4956 4957 4958 4959 4960 4961 4962 4963 4964
4965 4966 4967 4968 4969 4970 4971 4972 4973 4974
4975 4976 4977 4978 4979 4980 4981 4982 4983 4984
4985 4986 4987 4988 4989 4990 4991 4992 4993 4994
4995 4996 4997 4998 4999 5000 5001 5002 5003 5004
5005 5006 5007 5008 5009 5010 5011 5012 5013 5014
5015 5016 5017 5018 3902 3903 3904 3905 3906 3907
3908 3909 3910 3911 3912 3913 3914 3915 3916 3917
3918 3919 3920 3921 3922 3923 3924 3925 3926 3927
3928 3929 3930 3931 3932 3933 3934 3935 3936 3937
3938 3939 3940 3941 3942 3943 3944 3945 3946 3947
3948 3949 3950 3951 3952 3953 3954 3955 3956 3957

3958 3959 3960 3961 3962 3963 3964 3965 3966 3967
3968 3969 3970 3971 3972 3973 3974 3975 3976 3977
3978 3979 3980 3981 3982 3983 3984 3985 3986 3987
3988 3989 3990 3991 3992 3993 3994 3995 3996 3997
3998 3999 4000 4001 4002 4003 4004 4005 4006 4007
4008 4009 4010 4011 4012 4013 4014 4015 4016 4017
4018 4019 4020 4021 4022 4023 4024 4025 4026 4027
4028 4029 4030 4031 4032 4033 4034 4035 4036 4037
4038 4039 4040 4041 4042 4043 4044 4045 4046 4047
4048 4049 4050 4051 4052 4053 4054 4055 4056 4057
4058 4059 4060 4061 4062 4063 4064 4065 4066 4067
4068 4069 4070 4071 4072 4073 4074 4075 4076 4077
4078 4079 4080 4081 4082 4083 4084 4085 4086 4087
4088 4089 4090 4091 4092 4093 4094 4095 4096 4097
4098 4099 4100 4101 4102 4103 4104 4105 4106 4107
4108 4109 4110 4111 4112 4113 4114 4115 4116 4117
4118 4119 4120 4121 4122 4123 4124 4125 4126 4127
4128 4129 4130 4131 4132 4133 4134 4135 1916 1917
1918 1919 1920 1921 1922 1923 1924 1925 1926 1927
1928 1929 1930 1931 1932 1933 1934 1935 1936 1937
1938 1939 1940 1941 1942 1943 1944 1945 1946 1947
1948 1949 1950 1951 1952 1953 1954 1955 1956 1957
1958 1959 1960 1961 1962 1963 1964 1965 1966 1967
1968 1969 1970 1971 1972 1973 1974 1975 1976 1977
1978 1979 1980 1981 1982 1983 1984 1985 1986 1987
1988 1989 1990 1991 1992 1993 1994 1995 1996 1997
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007
2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
2018 2019 2020 2021 2022 2023 2024 2025 2026 2027
2028 2029 2030 2031 2032 2033 2034 2035 2036 2037
2038 2039 2040 2041 2042 2043 2044 2045 2046 2047
2048 2049 2050 2051 2052 2053 2054 2055 2056 2057
2058 2059 2060 2061 2062 2063 2064 2065 2066 2067
2068 2069 2070 2071 2072 2073 2074 2075 2076 2077
2078 2079 2080 2081 2082 2083 2084 2085 2086 2087
2088 2089 2090 2091 2092 2093 2094 2095 2096 2097
2098 2099 2100 2101 2102 2103 2104 2105 2106 2107
2108 2109 2110 2111 2112 2113 2114 2115 2116 2117

2118 2119 2120 2121 2122 2123 2124 2125 2126 2127
2128 2129 2130 2131 2132 2133 2134 2135 2136 7189
7190 7191 7192 7193 7194 7195 7196 7197 7198 7199
7200 7201 7202 7203 7204 7205 7206 7207 7208 7209
7210 7211 7212 7213 7214 7215 7216 7217 7218 7219
7220 7221 7222 7223 7224 7225 7226 7227 7228 7229
7230 7231 7232 7233 7234 7235 7236 7237 7238 7239
7240 7241 7242 7243 7244 7245 7246 7247 7248 7249
7250 7251 7252 7253 7254 7255 7256 7257 7258 7259
7260 7261 7262 7263 7264 7265 7266 7267 7268 7269
7270 7271 7272 7273 7274 7275 7276 7277 7278 7279
7280 7281 7282 7283 7284 7285 7286 7287 7288 7289
7290 7291 7292 7293 7294 7295 7296 7297 7298 7299
7300 7301 7302 7303 7304 7305 7306 7307 7308 7309
7310 7311 7312 7313 7314 7315 7316 7317 7318 7319
7320 7321 7322 7323 7324 7325 7326 7327 7328 7329
7330 7331 7332 7333 7334 7335 7336 7337 7338 7339
7340 7341 7342 7343 7344 7345 7346 7347 7348 7349
7350 7351 7352 7353 7354 7355 7356 7357 7358 7359
7360 7361 7362 7363 7364 7365 7366 7367 7368 7369
7370 7371 7372 7373 7374 7375 7376 7377 7378 7379
7380 7381 7382 7383 7384 7385 7386 7387 7388 7389
7390 7391 7392 7393 7394 7395 7396 7397 7398 7399
7400 7401 7402 7403 7404 7405 7406 7407 7408 7409
7410 7411 7412 7413 7414 7415 7416 7417 7418 1000
1001 1002 1003 1004 1005 1006 1007 1008 1009 1010
1011 1012 1013 1014 1015 1016 1017 1018 1019 1020
1021 1022 1023 1024 1025 0810 0811 0812 0813 0814
0815 0816 0817 0818 0819 0820 0821 0822 0823 0824
0825 0826 0827 0828 0829 0830 0831 0832 0833 0834
0835 0836 0837 0838 0839 0840 0841 0842 0843 0844
0845 0846 0847 0848 0849 0850 0851 0852 0853 0854
0855 0856 0857 0858 0859 0860 0861 0862 0863 0864
0865 0866 0867 0868 0869 0870 0871 0872 0873 0874
0875 0876 0877 0878 0879 0880 0881 0882 0883 0884
0885 0886 0887 0888 0889 0890 0891 0892 0893 0894
0895 0896 0897 0898 0899 0900 0901 0902 0903 0904
0905 0906 0907 0908 0909 0910 0911 0912 0913 0914

```
0915  0916  0917  0918  0919  0920  0921  0922  0923  0924
0925  0926  0927  0928  0929  0930  0931  0932  0933  0934
0935  0936  0937  0938  0939  0940  0941  0942  0943  0944
0945  0946  0947  0948  0949  0950  0951  0952  0953  0954
0955  0956  0957  0958  0959  0960  0961  0962  0963  0964
0965  0966  0967  0968  0969  0970  0971  0972  0973  0974
0975  0976  0977  0978  0979  0980  0981  0982  0983  0984
0985  0986  0987  0988  0989  0990  0991  0992  0993  0994
```

# Bibliography


Andreou A., Kamm T., Cohen J. (1994). Experiments in Vocal Tract Normalization. In *Proceedings of the CAIP Workshop*, 1994.

Brandt S. (1975). *Datenanalyse*. B.I.-Wissenschaftsverlag.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole.

Buntine W., Niblett T. (1992). A further Comparison of Splitting Rules for Decision-Tree Induction. *Machine Learning*, Vol. 8.

Denecke M. (2002). *Generische Aktionsmuster für aufgabenorientierte Dialogmuster*. PhD Thesis, University of Karlsruhe, Germany, 2002.

Deng L. (1998). A Dynamic, Feature-based Approach to the Interface between Phonology and Phonetics for Speech Modeling and Recognition. *Speech Communication*, Vol. 30.

Eide E. (2001). Distinctive Features for Use In an Automatic Speech Recognition System. In *Proceedings of the Eurospeech*, Aalborg, Denmark, 2001.

Fant G. (1960). *Acoustic Theory of Speech Production*. Mouton & Co.

Finke M., Geutner P., Hild H., Kemp T., Ries K., Westphal M. (1997). The Karlsruhe-Verbmobil Speech Recognition Engine. In *Proceedings of the ICASSP*, Munich, Germany, 1997.

Fügen C., Rogina I. (2000). Integrating Dynamic Speech Modalities into Context Decision Trees. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

Gales M.J.F. (1999). Semi-Tied Covariance Matrices for Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, Vol. 2.

Gauvain J., Lee C. (1994). Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, Vol. 2.

Godfreq J.J., Holliman E., MaxDaniel J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, San Francisco, USA, 1992.

Gopinath R. (1998). Maximum Likelihood Modeling with Gaussian Distributions for Classification. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

Hansen J., Bou-Ghazale S., Sarikaya R., Pellom B. (1998). Getting Started with the SUSAS: Speech under Simulated and Actual Stress. Technical Report RSPL-98-10, Duke University, Department of Electrical Engineering, 1998.

Hansen J. (1996). Analysis and Compensation of Speech under Stress and Noise for Robustness in Speech Recognition. *Speech Communications*, Vol. 20(2).

Hernando J. (1997). 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.

Hirschberg J., Litman D., Swerts M. (1999). Prosodic Cues to Recognition Errors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Keystone, USA, 1999.

Hirschberg J., Litman D., Swerts M. (2000). Gernalizing Prosodic Prediction of Speech Recognition Errors. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

Hirschberg J., Litman D., Swerts M. (2001). Identifying User Corrections Automatically in Spoken Dialgue Systems. In *North American*

*Chapter of the Association for Computational Linguistics (NAACL)*, Pittsburgh, USA, 2001.

Holtzapfel H., Fügen C., Denecke M., Waibel A. (2002). Integrating Emptional Cues into a Framework for Dialoge Management. In *Proceedings of the International Conference on Multimodal Interfaces*, USA, 2002.

Holtzapfel H. (2003). Emotionen als Parameter der Dialogverarbeitung. Master's thesis, University of Karlsruhe, Germany, 2003.

Humphries J. (1997). *Accent Modelling and Adaptation in Automatic Speech Recognition.* PhD Thesis, University of Cambridge, 1997.

International Phonetic Association (1999). *Handbook of the International Phonetic Association.* Cambridge University Press.

Johnson D. E., Oles F. J., Zhang T., Goetz T. (2002). A Decision-Tree-based Symbolic Rule Induction System for Text Categorization. *IBM System Journal, Special Issue on AI*, Vol. 3.

Kienast M., Paeschke A., Sendlmeier W. (1999). Articulatory Reduction in Emotional Speech. In *Proceedings of the Eurospeech*, Budapest, Hungary, 1999.

Kirchhoff K., Fink G., Sagerer G. (2000). Conversational Speech Recognition using Acoustic and Articulatory Features. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

Kirchhoff K. (1999). *Robust Speech Recognition Using Articulatory Information.* PhD Thesis, Universität Bielefeld (Germany), 1999.

Köster S. (2001). Acoustic-Phonetic Characteristics of Hyperarticulated Speech for Different Speaking Styles. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, 2001.

Ladefoged P. (1975). *A Course in Phonetics.* Harcourt College.

Leggetter C.J. (1995). *Improving Acoustic Modelling for HMMs using Linear Transforms.* PhD Thesis, Cambridge University, England, 1995.

Lemmetty S. (1999). Review of Speech Synthesis Technology. Master's thesis, Helsinki University of Technology, Finland, http://www.acoustics.hut.fi, 1999.

Levow G. (1998). Characterizing and Recognizing Spoken Corrections in Human-Computer Dialogue. In *Proceedings of the International Conference on Computational Linguistics*, Montreal, Canada, 1998.

Lindblom B., Brownlee S., Davis B., Moon S.-J. (1992). Speech Transforms. *Speech Communication*, Vol. 11.

Medan Y., Yair E., Chazan D. (1991). Super Resolution Pitch Determination of Speech Signals. *IEEE Transactions on Signal Processing*, Vol. 39.

Metze F., Waibel A. (2002). A Flexible Stream Architecture for ASR using Articulatory Features. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, USA, 2002.

Metze , Waibel A. (2003). Using Articulatory Information for Speaker Adaptation. In *Proceedings of the Automatic Speech and Recognition Workshop (ASRU)*, St. Thomas, USA, 2003.

Morgan N., Fossler-Lussier E. (1998). Combining Multiple Estimators of Speaking Rate. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, http://www.icsi.berkeley.edu/ftp/global/pub/speech/morgan, 1998.

Oviatt S. (1998). The CHAM Model of Hyperarticulate Adaptation During Human-Computer Error Resolution. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

Pearl J. (1988). *Probabilistic Reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Picheny M., Durlach N., Braida L. (1986). Speaking Clearly for the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech. *Journal of Speech and Hearing Research*, Vol. 29.

Picheny M. (1981). *Speaking Clearly for the Hard of Hearing*. PhD Thesis, Massachusetts Institute of Technology, USA, 1981.

Pickert J. P. (2000). *The American Heritage Dictionary of the English Language*. Houghton Mifflin.

Press W., Flannery B., Teukolsky S., Vetterling W. (1988). *Numerical Recipes in C*. Cambridge University Press.

Rabiner L. (1978). *Digital Processing of Speech Signals*. Prentice-Hall.

Rogina I., Waibel A. (1994). Learning state-dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, 1994.

Rogina I. (1997). Automatic Architecture Design by Likelihood-Based Context Clustering with Crossvalidation. In *Proceedings of the Eurospeech*, Rhodos, Greece, 1997.

Shriberg E., Wade E., Price P. (1992). Human Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction. In *Proceedings of the 5th DARPA Workshop on Speech and Natural Language*, USA, 1992.

Soltau H., Waibel A. (1998). On the Influence of Hyperarticulated Speech on the Recognition Performance. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

Soltau H., Waibel A. (2000a). Acoustic Models for Hyperarticulated Speech. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

Soltau H., Waibel A. (2000b). Specialized Acoustic Models for Hyperarticulated Speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

Soltau H., Metze F., Fügen C., Waibel A. (2001a). A One Pass Decoder based on polymorphic linguistic context assignment. In *Proceedings of the Automatic Speech and Recognition Workshop (ASRU)*, Madonna di Campiglio, Italy, 2001.

Soltau H., Schaaf T., Metze F., Waibel A. (2001b). The ISL Evaluation System for Verbmobil-II. In *Proceedings of the ICASSP*, Salt Lake City, USA, 2001.

Soltau H., Metze F., Waibel A. (2002a). Compensating for Hyperarticulation by Modeling Articulatory Properties. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, USA, 2002.

Soltau H., Yu H., Metze F., Fügen C., Pan Y., Jou S. (2002b). ISL Meeting Recognition. In *Rich Transcription Workshop*, Vienna, USA, 2002.

Soltau H., Yu H., Metze F., Fügen C., Jin Q., Jou S. (2003). The ISL Evaluation System for RT-03 CTS. In *Rich Transcription Workshop*, Boston, USA, 2003.

Suhm B. (1998). *Multimodal Interactive Error Recovery for Speech User Interfaces*. PhD Thesis, University of Karlsruhe, Germany, 1998.

Thimm G., Luettin J. (1999). Extraction of Articulators in X-Ray Images. In *Proceedings of the Eurospeech*, Budapest, Hungary, 1999.

Vapnik V. (1998). *Statistical Learning Theory*. Wiley.

Woodland P., Povey D. (2002). Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. *Computer Speech and Language*, Vol. 6.

Wordnet (2003). *Wordnet, a lexical database for the English language*. Internet, http://www.cogsci.princeton.edu.