Internal Use Only（非公開）

ＴＲ－ＳＬＴ－００８５

Automatic Assessment of the Pronunciation Quality of English Speech Uttered by Non-Natives

Tobias Cincarek and Rainer Gruhn

2004/09/30

概要

This work is about scoring the quality of pronunciation of non-native speakers on utterance and word level. Besides the theoretical background, a literature survey on scoring methods is also given. In the practical part results of the analysis of the human evaluation of the ATR SLT Non-Native English Database are reported. Moreover, experiments for scoring the quality of utterances, automatically assessing the pronunciation skill of non-native speakers and for detecting mispronounced words are carried out.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２TEL：0774－95－1301

Advanced Telecommunications Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

Abstract.

The target of this thesis is the automatic assessment of the pronunciation quality of words and sentences in a second language as well as the automatic assessment of the pronunciation skill of non-native speakers. Possible applications are systems for computer assisted pronunciation training (CAPT).

Non-native English speech from 57 German school children (FAU LME data) and non-native English adult speech from 96 speakers of multiple accent groups (ATR SLT data) are available for experiments. The LME data has annotations on word and speaker level, the SLT data on word and utterance level. Since the same material in the SLT data was evaluated by several human experts w.r.t. pronunciation quality, an analysis of the reliability of the annotations could be carried out. A word mispronunciation model is proposed, which allows the estimation of the mispronunciation probabilities of single phonemes from a statistic of mispronounced words.

Features, which are intended to measure the pronunciation quality of words and sentences are defined and examined. These so-called pronunciation features are mainly calculated from the forced-alignment and the phoneme or word recognition result obtained with a speech recognizer. Additionally, phoneme duration statistics and phoneme confusion matrices are considered. The efficiency of single utterance level features is analysed by means of the correlation to a human evaluation. The data-driven techniques PCA and LDA are examined for feature space transformation. However, for selection of suitable combinations of word and utterance features the floating search algorithm was found to be more effective.

Experiments for the discrimination of correctly pronounced and mispronounced words within a sentence are carried out with the Gaussian classifier and decision trees by combining several features. For both the SLT and LME data a class-wise average recognition rate of 72% is achieved. The accuracy was even 90% for native adult speech.

For scoring utterances the Gaussian classifier and the linear regression are employed. The correlation between human ratings and scores was 0.59 on the utterance level and 0.84 on the speaker level for the SLT data. In the case of the LME data, the correlation was 0.69. An interesting result for children speech is, that the younger the children speaker, the more probable is a low pronunciation quality of utterances.

Kurzfassung.

Diese Arbeit beschäftigt sich mit der automatischen Bewertung der Aussprachequalität von Wörtern und Sätzen in einer Zweitsprache, sowie der automatischen Einschätzung der allgemeinen Aussprachefähigkeit nicht-nativer Sprecher. Mögliche Anwendungen sind Systeme für das computergestützte Lernen von Fremdsprachen (CAPT).

Es stehen die FAU LME Stichprobe mit englischer Sprache von 57 deutschen Schulkindern und die ATR SLT Stichprobe mit englischer Sprache von 96 Erwachsenen verschiedener Akzentgruppen, u.a. Deutsch, Französisch, Japanisch, Chinesisch und Indonesisch zur Verfügung. In der LME Stichprobe gibt es Annotationen auf Wort- und Sprecherebene, in der SLT Stichprobe auf Wort- und Satzebene. Die Wortannotationen bestehen in erster Linie aus Markierungen von falsch ausgesprochenen Wörtern. Äußerungen und Sprecher sind auf einer diskreten Skala von 1 bis 5 bzgl. ihrer Aussprachequalität bewertet. Da gleiches Material der SLT Stichprobe von mehreren menschlichen Bewertern annotiert wurde, konnte zusätzlich eine Analyse der Zuverlässigkeit der Annotationen durchgeführt werden. Dabei wird auch ein Verfahren vorgeschlagen, mit dem die Wahrscheinlichkeit der Falschaussprache von einzelnen Lauten aufgrund der Wortmarkierungen geschätzt werden kann. Es zeigte sich eine hohe Korrelation zwischen den geschätzten Wahrscheinlichkeiten und der Erkennungsrate für jeden einzelnen Laut, sowie zwischen ursprünglichen und geschätzen Wahrscheinlichkeiten des Grades der Falschaussprache von einzelnen Wörtern.

Merkmale, welche die Aussprachequalität von Wörtern und Sätzen messen sollen, werden definiert und untersucht. Diese sog. Aussprache-Merkmale basieren auf dem Ergebnis der Laut- und Worterkennung sowie der erzwungenen Zeitzuordnung einer Äußerung, welche mit Hilfe eines Spracherkenners ermittelt werden. Weiterhin werden Dauerstatistiken von Lauten und Lautverwechslungsmatrizen für falsch und richtig ausgesprochene Wörter zur Merkmalberechnung herangezogen. Die Güte einzelner Merkmale wird u.a. mit Hilfe der Korrelation zu den menschlichen Bewertungen analysiert.

In Experimenten werden die Karhunen-Loéve Transformation (KLT) und die Lineare Diskriminanzanalyse (LDA) als Merkmalraumtransformationen, sowie die alternierende Suche als Merkmalauswahlverfahren verwendet. Letztere erwies sich als besonders geeignet für die Identifikation wichtiger Wort- und Satzmerkmale.

Zur Unterscheidung richtig und falsch ausgesprochener Wörter innerhalb eines Satzes werden der Gauss-Klassifikator und Entscheidungsbäume eingesetzt. Für beide Stichproben wird dabei eine klassenweise gemittelte Erkennungsrate von 72% erreicht. Der Normalverteilungsklassifikator hatte dabei die beste Generalisierungsfähigkeit. Angesichts der Tatsache, dass die menschlichen Bewerter im Durchschnitt nur 58% der falsch ausgesprochenen Wörter erkannt und etwa 8% der als richtig zu behandelnden Wörter als falsch ausgesprochen markiert haben, ist dies ein gutes Ergebnis. Zudem wurden 90% der Äußerungen von erwachsenen englischen Muttersprachlern als richtig ausgesprochen klassifiziert.

Zur Bewertung von Äußerungen wird der Gauss-Klassifikator und die Lineare Regression eingesetzt. Die Ausgabe der linearen Regressionsfunktion musste dabei zusätzlich transformiert werden, um ein akkurates Bewertungsergebnis zu erhalten. Die Korrelation zwischen menschlichen und automatischen Bewertungen war 0.59 auf Äußerungsebene und 0.84 auf

Sprecherebene für die SLT Stichprobe. Im Fall der LME Stichprobe lag die Korrelation auf Sprecherebene bei 0.69. Ein interessantes Ergebnis für native Kindersprache ist, dass je geringer das Alter der Kinder, desto wahrscheinlicher ist eine mangelhafte Aussprachequalität.

Acknowledgements.

The author would like to thank Dr. Satoshi Nakamura and Dr. Seiichi Yamamoto who made the stay at Advanced Telecommunications Research (ATR) International, Spoken Language Translation (SLT) Research Laboratories possible, and Prof. Dr. Heinrich Niemann and Dr. Elmar Nöth who agreed to an external performance of the Diploma thesis. Further acknowledgements go to Dipl.-Inf. Rainer Gruhn, Dipl.-Inf. Christian Hacker and Dipl.-Inf. Stefan Steidl for kindly supervising this thesis.

Contents

Chapter 1

Introduction

1.1 Motivation

English is not the native language of most humans in the world, but it is the language which is taught at schools for secondary education in almost every country. Speaking English is the most common way to communicate with other people when traveling in foreign countries whose official language you are not sufficiently familiar with.

A non-native English speaker, which is somebody whose first language is not English, will pronounce words differently from native speakers. There may be various reasons, e.g. the unfamiliarity with the pronunciation of certain words, the incapability in recognizing certain speech sounds, the difficulty in articulating certain sounds, or the English education regarding pronunciation.

In the last decades there has been remarkable progress in speech recognition technology. A speech recognizer is a device which accepts as input a digitized speech waveform and outputs a word sequence. When this word sequence matches the transcript of the original utterance, the speech recognizer works fine. Recently, practical applications of speech recognition in automatic speech dialog systems such as flight reservation or cinema information services or in robotics for natural human-machine communication emerged.

When building a speech recognizer with a large amount of speech data uttered by speakers of a certain native speaker group, e.g. native American English speakers from the U. S., and testing its performance with speech data from the same group and the evaluation task is feasible, e.g. for read newspaper sentences, high recognition accuracy can be achieved.

However, when trying to recognize English speech of German, Japanese, Indonesian or other non-native speakers, the results can be very poor. For everyone it is very annoying, if he cannot use a speech-controlled timetable information service or automatic booking service, simply because the system cannot properly recognize ones non-native speech.

How to deal with this unsatisfying situation? There are two possibilities: Make speech recognition or humans' English pronunciation better! To do both is of course the silver bullet. The improvement of speech recognition involves approaches like the adaptation of a speech recognizer's acoustic model or pronunciation dictionary. However adaptation methods will not

be considered in this thesis. The main target of this work is to investigate methods for the automatic assessment of non-native speech in terms of pronunciation quality. Such methods may be employed in software systems to support the learner of a foreign language in acquiring the correct pronunciation of speech sounds. Before more details about contributions of this work (cf. Section 1.4) are given, the next two Sections will briefly introduce the technical meaning of the word "pronunciation" and the general purpose and architecture of systems for pronunciation training.

1.2 Aspects of Pronunciation

The meaning of the term "pronunciation" in everyday speech is vague. Here an attempt is made to define the technical meaning of the term by three aspects:

- **Segmental**: closeness of the pronunciation of single phonemes to native speakers
- **Temporal**: speaking rate, duration of phonemes, inter-word pauses
- **Prosodic**: sentence intonation, syllable stress

The segmental aspect may be closest to the meaning in everyday speech. If a non-native speaker's pronunciation of a single phoneme is far away from a native speaker's pronunciation the phoneme will be considered as "mispronounced". The "distance" necessary to declare a phoneme as mispronounced depends on the human listener. Consequently, it is difficult to determine objectively a clear boundary between "correctly pronounced" and "mispronounced" phonemes. Assuming a method to judge phoneme pronunciation is given, a word will be considered as mispronounced, if one or more phonemes are mispronounced, or any phonemes are inserted or deleted.

The temporal aspect of pronunciation can be grasped intuitively. A non-native speaker who is able to utter many words and sentences within a certain time, i.e. has a high speaking rate, may be regarded as fluent. On the other hand, a speaker's skill is regarded as low, if he makes longer pauses before uttering anything, since he needs time to construct a sentence hopefully obeying the foreign language's grammar, or he speaks slowly or even stutters, because he is unsure about the pronunciation of words or even has difficulty to pronounce certain speech sounds.

The prosodic aspect, which, strictly speaking, includes the temporal aspects already mentioned, comprises the sentence intonation, e.g. rising pitch at the end of interrogative sentences, and the lexical stress of words, e.g. syllable "pu" in the word "com-pu-ter".

However for this thesis only the segmental and temporal aspects are examined. No speech database with human annotations for the prosodic aspects were available and bad pronunciation quality with respect to the two former aspects (segmental and temporal) has presumably more detrimental effects to the intelligibility of speech.

1.3 CALL and CAPT

The purpose of a **computer assisted language learning** (CALL) system in general is to make the acquisition of a second language (L2) for the learner more effective. The final goal is to built a system, which does not rely on the presence of a human teacher. The learning of L2 comprises the acquisition of the foreign language's vocabulary, grammar and pronunciation of speech sounds. For this thesis only the aspect pronunciation is considered. Systems which serve the purpose of pronunciation training are referred to as **computer assisted pronunciation training** (CAPT) systems.

A lot of research work on the issue of pronunciation training has already been carried out and first commercial systems for pronunciation training evolved. The development of a CAPT system does not only involve technological but also pedagogical issues. In [NCSB02] Neri et al. analyze existing literature in order to identify the pedagogical requirements for a CAPT system. A bunch of already available CAPT systems is evaluated for certain criteria. In the following a summary is given.

Pronunciation training consists of the three factors input, output and feedback.

- **Input:** e.g. listening to native speech
- **Output:** production of L2 speech, e.g. utter a given sentence or speak freely
- **Feedback:** assessment of pronunciation, e.g. numerical as a score or graphical

The input to the learner may for example be native speech: the student listens to speech samples of native speakers. Another kind of input may be a description of how to pronounce certain speech sounds or a 2D or 3D picture or animation of how to move the articulators to achieve the correct pronunciation.

In order to practice pronunciation the learner produces L2 speech as output. Training may consist of exercises where the learner has to read text prompts, listen and repeat drills, or interactive dialogs, where the user answers freely to questions or reads one out of a choice of given answers. In the study it is emphasized, that pronunciation training should not be limited to the reading of single isolated words, since it is unrealistic and may not lead to the necessary transfer of skills to actual conversation. Consequently practicing dialogs, which are likely to occur in everyday conversation, is recommended.

Since the learner may not be able to perceive whether his pronunciation of speech sounds is correct, feedback from a teacher, who makes the learner aware of his mistakes, is imperative. The meaning of "correct" regarding pronunciation is not straightforward. Its interpretation depends whether the target of training should be an accent-free and native-like pronunciation or just comprehensibility, i.e. neglecting a non-native accent as long as it does not affect intelligibility.

The most common type of feedback provided by human teachers is the recast, i.e. repetition with change and possibly with emphasis of the learners mistake. This kind of feedback has proven to be most effective for the uptake of phonological errors.

Existing CAPT system provide various kinds of feedback. These include

- Waveform, spectrogram, pitch curve, etc.
- Highlighting of words or phonemes
- Pronunciation score

The display of waveform and spectrogram is common but problematic, as usually their interpretation is left to the learner. Even if the student can compare his spectrogram of an utterance to a reference spectrogram of a native speaker, similarity or difference between these does not necessarily allow conclusions about the pronunciation correctness. The display and comparison of pitch or intensity (energy), however, is reasonable, since they are not as variable as waveforms or spectrograms and easier to interpret. Thus prosodic aspects like word stress and sentence intonation can be covered. It remains to address the problem of feedback for the segmental aspects of pronunciation.

Feedback for highlighting words or phonemes, which are classified as mispronounced, as well as one or more pronunciation scores for an utterance is implemented in some CAPT systems. The highlighting of mispronounced segments is a technological challenge. If there are many words, which are in fact correctly pronounced, but are marked as mispronounced and vice versa, the usefulness of such a system will be questioned. Reliability is of great importance, so that the learner is not confused by erroneous feedback. Current systems do not yet meet this requirement sufficiently.

The benefit of pronunciation scores relies on their definition and usage. For example to consider only the fluency of a speaker, which can be measured by the speaking rate, is problematic, since a speaker can improve his score by just speaking faster. Consequently, a different score, which measures overall pronunciation quality, or at least one additional score which measures the segmental quality has to be employed.

In the conclusion recommendations for the building of future CAPT systems are given. A CAPT system should address the learners' needs. The aim of pronunciation training should be intelligibility rather than accent-free pronunciation. The system should provide L2 input from different native speakers and computer animations of, e.g. lip movements. L2 production should be practiced in interactive dialogs, but not only by listen and repeat drills. Real-time feedback should include the scoring of overall comprehensibility and the highlighting of presumably incorrectly pronounced areas. Segmental as well as supra-segmental, i.e. prosodic or temporal, aspects should be considered in assessing pronunciation performance. The highlighting of areas should concentrate on frequent errors, errors which reduce intelligibility, and errors which can be detected robustly.

1.4 Contribution of this Work

The following three topics will be covered in this thesis.

1. Analysis of a human evaluation of non-native speech, which was carried out on the sentence and the word level (Chapter 5)
2. Investigation of methods for the automatic assessment of the pronunciation quality of single words and whole sentences. (Chapters 6 ff.)
3. Find out differences of the characteristics of non-native speech in comparison to native speech.

Issue (1) is a prerequisite for (2), because a human reference is necessary for the validation of a system for automatic assessment of pronunciation. The methods in (2) may be used as part of a CAPT system. The investigation (2) includes an attempt to develop a method for identification of mispronounced words. Issue (3) will be discussed after insight is gained from the results for (2).

A method for automatic assessment of pronunciation should be as universal as possible. Therefore, this thesis aims at a scoring system, that is

- text-independent, i.e. can score any utterance
- independent from a student's first language (L1)
- independent from the target language (L2)

Of course, a system independent from L2 still requires speech data, models and statistics for L2, but the method can work in principle for any target language. Reliability is also an important issue. In general the reliability of automatic assessment will increase, as more speech data is considered at once at the same time, for example when scoring whole utterances. On the other hand for the learner it is very important to localize pronunciation errors, which requires to make judgments from little data. Consequently, the trade-off between more reliable, overall assessments and less reliable, more specific assessment is critical.

1.5 Thesis Outline

Chapter 2 describes the necessary theoretical background for pronunciation scoring and evaluation methods employed and developed in this thesis. A survey about literature on human evaluation of pronunciation and automatic scoring of pronunciation is given in *Chapter 3*. *Chapter 4* provides general information about the speech data used in experiments. English speech data was recorded at ATR from 96 non-native speakers of multiple accent groups. Furthermore each speaker's data was evaluated by English teachers. The results of that analysis are presented in *Chapter 5*. Features employed for word, utterance and speaker level scoring

are defined in *Chapter 6*. Moreover, a preliminary analysis of the usefulness of utterance level features is carried out. *Chapter 7* explains the setup of experiments for utterance and speaker scoring as well as for the detection of mispronounced words. The results of experiments is reported in *Chapter 8*. Further ideas for future work on pronunciation scoring are described in *Chapter 9*. The thesis closes with a summary in *Chapter 10*.

They are several appendices with comprehensive information about the speech data and experimental results. *Appendix A* gives information about the non-native speakers of the ATR SLT non-native adult speech database and the FAU LME non-native children speech database. Background information about the human evaluation of the ATR SLT data can be found in *Appendix B*. Feature distribution plots are arranged in *Appendix C*.

Chapter 2

Fundamentals

This chapter describes briefly the fundamentals about speech, speech transcription, speech recognition, pattern recognition and statistics.

2.1 Speech and Transcription

Speech is the most convenient way for humans to communicate with each other. With the emerging of the speech recognition technology, speech is increasingly used for human-machine interaction. Every human can produce a huge number of speech sounds with his speech organs. The main organs are the vocal tract and the articulators, like the glottis, the lips, the teeth, the tongue, the hard and soft palate and the velum.

Each "configuration" of the speech organs leads to the production of a certain speech sound. For example, if the tongue tip is touching the region in front of the upper incisors and the glottis is pulsating the sound /l/ is generated. The International Phonetic Association proposed a system for classification and transcription of speech sounds. It is intended to be able to transcribe the sounds of any existing language in the world. The speech sound categories (phones) are determined by place and manner of articulation (cf. [IPA99] for details).

Each natural language consists only of a subset of the speech sounds a human can produce. This subset differs for each language. If concerning only a certain language, speech sounds are categorized into phonemes. For example, there are sounds which can be interchanged without changing the meaning of a word. If, however, the meaning changes, the sound belongs to two different phonemes.

Several systems to transcribe speech by means of phones or phonemes exist. The most universal system is the International Phonetic Alphabet (IPA) of the International Phonetic association. Despite the IPA symbols are available in the Unicode character set, they are inconvenient for automatic processing. Consequently, other machine readable systems like SAMPA [SAM] were proposed. SAMPA was initially used for several European languages. In order to cover the whole symbol inventory of the IPA, there is a draft proposal X-SAMPA [Wel] for the necessary extensions.

There is an English specific transcription system called ARPAbet, which was originally used

Table 2.1: *TIMIT and X-SAMPA symbols with examples for American English.*

TIMIT	X-SAMPA	Word, Transcription	
aa	A	pot	pAt
ae	{	pat	p{t
ah	V	cut	kVt
ao	O	cause	kOz
aw	aU	rouse	raUz
ax	@	allow	@laU
axr	@‘	corner	kOrn@‘
ay	aI	rise	raIz
b	b	bin	bIn
ch	tS	chin	tSIn
d	d	din	dIn
dh	D	this	DIs
dx	4	data	deI4@
eh	e	raise	reIz
er	3‘	furs	f3‘z
ey	eI	able	eIb@l
f	f	fin	fIn
g	g	give	gIv
hh	h	hit	hIt
ih	I	pit	pIt
ix	I\	image	ImI\dZ
iy	i	idly	aIdli

TIMIT	X-SAMPA	Word, Transcription	
jh	dZ	change	tSeIndZ
k	k	kin	kIn
l	l	long	lON
m	m	mock	mAk
n	n	knock	nAk
ng	N	thing	TIN
ow	@U	obey	@UbeI
oy	OI	oil	OIl
oh	o	nose	noz
p	p	page	peIdZ
r	r	ring	rIN
s	s	sin	sIn
sh	S	shin	SIn
t	t	tin	tIn
th	T	thin	TIn
uh	U	put	pUt
uw	u	lose	luz
v	v	vain	veIn
w	w	way	weI
y	j	yard	jArd
z	z	zebra	zibr@
zh	Z	asian	eIZ@n

for the TIMIT corpus. There are variants of the ARPAbet employing less symbols and which are used for other English speech corpora like the Wall Street Journal (WSJ). Any of these alphabets are commonly referred to as TIMIT. Table 2.1 lists symbols of the TIMIT alphabet and the corresponding symbols in X-SAMPA with example transcriptions for American English words.

2.2 Speech Recognition

There is a vast literature about speech recognition. Standard references include [RJ93, Sch95, Jel97, GM00]. The task of speech recognition is to invert the speech production process, which is to determine automatically the spoken word sequence from an acoustic speech signal. Figure 2.1 shows the basic architecture of a speech recognizer. Its components are described in the following passages.

The statistical approach has proven to be successful to solve the problem of speech recognition with high performance. Equation (2.1) is fundamental for the statistical approach. The idea is to find the word sequence $\boldsymbol{w}^*$ with the highest probability given a sequence of acoustic

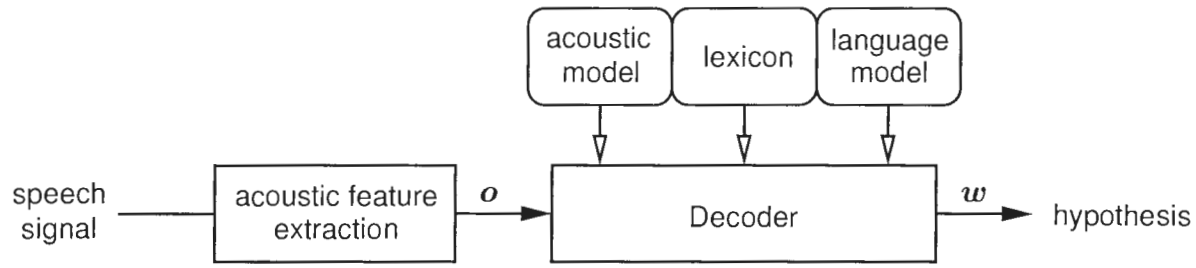


Figure 2.1: *The basic architecture of a speech recognizer.*

observations $\boldsymbol{o}$ and linguistic constraints. This approach separates the problem into the part of acoustic modeling to obtain the acoustic probabilities $P(\boldsymbol{o}|\boldsymbol{w})$, and the part of language modeling to obtain the word sequence probabilities $P(\boldsymbol{w})$.

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{w}|\boldsymbol{o}) = \underset{\boldsymbol{w}}{\operatorname{argmax}} P(\boldsymbol{o}|\boldsymbol{w})P(\boldsymbol{w}) \tag{2.1}$$

Feature extraction. The raw speech signal is an unsuitable representation for speech recognition. For example, it is very prone to interferences by the recording equipment or transmission channel and by noise in the recording environment. Furthermore it contains much unnecessary information about the speaker and his characteristics. Most important for the identification of speech sounds is the vocal tract and its resonance frequencies. The vocal tract's characteristics change when the articulators are moved to produce a certain speech sound. Furthermore, the presence or absence of an excitation by the glottis pulses determines whether the speech sound is voiced or unvoiced. This suggests that the spectrum of the speech signal will be a more suitable representation and that the separation of the vocal tract characteristics from the excitation is important.

The most widely used features for speech recognition are the so-called Mel-Frequency Cepstrum Coefficients (MFCC) and the logarithmic energy together with their first and second derivations. For details about acoustic preprocessing and acoustic feature extraction confer [Pic93]. The result of the acoustic feature extraction is a sequence of feature vectors. Each acoustic feature vector represents the speech signal of a certain time interval. The length of this time interval is commonly set to a value between 10 and 20 milliseconds.

Lexicon and subword units. The lexicon contains one or more entries for each word of the recognition vocabulary. Each entry is a sequence of subword units, e.g. phones or phonemes. The lexicon is also called pronunciation dictionary. Which kind of subword unit to use, depends on the application and the number of data available to train the acoustic model of the recognizer. If the application is only digit recognition, no subword units have to be defined at all. However, as the desired recognition vocabulary increases, it becomes necessary to employ a smaller set of subword units to transcribe each word, because it would be infeasible to build a robust acoustic representation for each word otherwise. If the chosen set of subword units is universal, i.e. the transcription of any word in a language is possible, the recognizer will in principle be able to decode any word even if it was not encountered during training.

The most common subword units are the context-independent monophones and the context-

dependent polyphones. Monophones correspond to single phonetic or phonemic units of the target language. A polyphone is defined by its center unit and a certain left and/or right context. For example the word "and" can be represented as a sequence of three monophones "ax n d" or as the sequence of three triphones "sp-ax+n ax-n+d n-d+sp". Here, "sp" is the phoneme symbol for inter-word pauses. The maximum number of polyphones, e.g. n^2 biphones and n^3 triphones, is much larger than the number of monophones (n), since any combination of the n base units may occur in words in general. The benefit of employing polyphones as subword units is the capability of a more accurate acoustic modeling, which accounts for coarticulation effects. A drawback of using them is the requirement for more training data and they may be less robust to variability. For example, while the recognition performance improves for native speech when switching from monophones to biphones or triphones, adverse results were observed for non-native speech [CGN].

Acoustic model. Most state-of-the-art speech recognition toolkits use the HMM-based approach for speech recognition. The acoustic model consists of one Hidden Markov Model (HMM) for each subword unit and additional models for silence, non-verbals or other kinds of noise. An HMM consists of states, state transitions and one output density per state. Figure 2.2 shows a left-to-right HMM with three states $q_1, \ldots, q_3$ with no state skips, which is typical for the acoustic modeling of one phonetic subword unit. With such an HMM a two-stage stochastic process can be modeled: The first stage of the process is hidden and consists of HMM state transitions $s_t \rightarrow s_{t+1}$, resulting in a sequence of states $\boldsymbol{s} = s_1, \ldots, s_T$. Here, s_t denotes a state variable having any state value q_j. The second stage is observable. Its outcome is a sequence of observations $\boldsymbol{o} = o_1, \ldots, o_T$, one per state $s_t = q_j$.

An HMM λ can be used to calculate an approximation of the probability $P(\boldsymbol{o}|p)$ for an observation sequence $\boldsymbol{o}$, given a subword unit p. It is an approximation since the true duration characteristics of a phoneme and the distribution of HMM state sequence lengths are different [RSS92]. Furthermore, to make computation feasible, the calculation of $P(\boldsymbol{o}|\lambda)$ is simplified by two assumptions to relax dependences:

(a) If the output probability $P(o_t|s_1, \ldots, s_t)$ depends only on the current state s_t, the probability $P(\boldsymbol{o}|\boldsymbol{s})$ is calculated as:

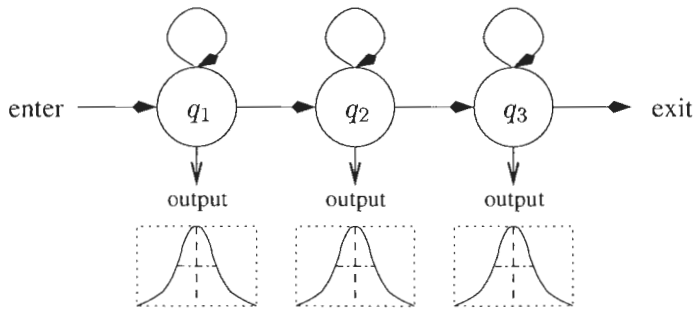


Figure 2.2: *Illustration of a three state HMM with output densities.*

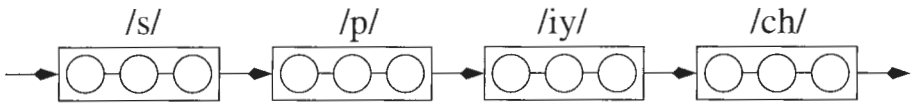


Figure 2.3: *Concatenation of four subword HMMs to a word model.*

$$P(\boldsymbol{o}|\boldsymbol{s},\lambda) = \prod_{t=1}^{T} P(o_t|s_1,\ldots,s_t) \approx \prod_{t=1}^{T} P(o_t|s_t) \tag{2.2}$$

(b) If the probability for a transition from state q_i to q_j depends only on q_i, the computation of the state sequence probability simplifies to

$$P(\boldsymbol{s}|\lambda) = P(s_0) \prod_{t=1}^{T} P(s_t|s_1,\ldots,s_{t-1}) \approx P(s_0) \prod_{t=1}^{T} P(s_t|s_{t-1}) \tag{2.3}$$

The parameters of an HMM, i.e. the state transition probabilities $P(q_i|q_j)$ and the probability density functions to calculate $P(o_t|s_t = q_j)$ are unknown. They can be learned automatically if speech data for each subword unit is available. A maximum likelihood estimation of the parameters is obtained by the Baum-Welch algorithm (cf. [Jel97] for a derivation).

The probability of an acoustic observation $\boldsymbol{o}$ given a subword HMM λ is calculated by summing over all possible sequences of states $\boldsymbol{s}$ through HMM λ (cf. Eq. 2.4). Another often needed value is the probability in Equation (2.5) of the HMM's best path. It is the probability of the state sequence $\boldsymbol{s}^*$ with maximum observation probability. Its calculation will be described in the passage about speech decoding.

$$P(\boldsymbol{o}|\lambda) = \sum_{\boldsymbol{s}} P(\boldsymbol{o}|\boldsymbol{s},\lambda)P(\boldsymbol{s}|\lambda) \approx \sum_{\boldsymbol{s}} P(s_0) \prod_{t=1}^{T} P(s_t|s_{t-1})P(o_t|s_t) \tag{2.4}$$

$$P(\boldsymbol{o},\boldsymbol{s}^*|\lambda) = \max_{\boldsymbol{s}} P(\boldsymbol{o}|\boldsymbol{s},\lambda)P(\boldsymbol{s}|\lambda) \approx \max_{\boldsymbol{s}} P(s_0) \prod_{t=1}^{T} P(s_t|s_{t-1})P(o_t|s_t) \tag{2.5}$$

Subword HMMs only represent single phonetic units. Word models can be obtained by concatenation of several subword HMMs. For example, Figure 2.3 illustrates the construction of an acoustic model for the word "speech" by combining the subword HMMs /s/, /p/, /iy/ and /ch/. The exit transition of HMM /s/ is the enter transition of HMM /p/ a.s.o. at the same time. The probability of an acoustic observation for the word HMM is calculated by summing up the logarithmic probabilities (also called likelihoods) of each subword unit's best path. For a whole sentence - represented as a sequence of word HMMs, which is nothing more than sequence of phoneme HMMs on the next lower level - the computation is analogue.

Language model. The last passage on acoustic modeling revealed how to get an estimate for the acoustic probability $P(\boldsymbol{o}|\boldsymbol{w})$ of subword units, single words or a whole sentence. It remains to determine the probability $P(\boldsymbol{w})$ of an utterance's word sequence $\boldsymbol{w} = (w_1,\ldots,w_m)$. The probability of a word w_i depends on the history of all previous words $w_1,\ldots,w_{i-1}$. A

standard approach is to model this probability by an n-gram which only considers sequences of n words, i.e. a history of only $n-1$ words, to make model estimation and probability computation feasible. For example, in case of the bigram ($n = 2$) the probability of the word sequence $\boldsymbol{w}$ is the product out of the probability $P(w_1)$ that a sentence starts with word w_1 and the bigram probabilities $P(w_i|w_{i-1})$ of all following word pairs (see Equation 2.6 with $k = 1$). For details about statistical language modeling confer [Jel97].

$$P(\boldsymbol{w}) = P(w_1)\prod_{i=2}^{m} P(w_i|w_1,\ldots,w_{i-1}) \approx P(w_1)\prod_{i=2}^{m} P(w_i|w_{i-k},\ldots,w_{i-1}) \tag{2.6}$$

Decoder. Speech decoding is a search problem: find the word sequence $\boldsymbol{w}^*$ with maximum probability for a given sequence of observations $\boldsymbol{o}$. It is a difficult task, since the search space, i.e. a word graph which describes all possible word sequences, is in general almost infinite. For practical applications the search space is reduced by limiting the recognition vocabulary and constraining possible word sequences with a language model.

In literature different decoding strategies are proposed. Confer [Sch95] or [Jel97] for several examples. Only the Viterbi decoding and the beam search method shall be introduced here. These decoding strategies are commonly employed in connection with word bigram language models.

The search space is a network of word HMMs as shown in Figure 2.4. Bigram probabilities $P(v|w)$ are associated with each link between the exit state of HMM for word w and the enter state of HMM for word v. The Viterbi search starts from the enter state s_0 of the network and calculates the state sequence $\boldsymbol{s}^*$ to an exit state s_e with maximum probability. The word sequence $\boldsymbol{w}^*$ which corresponds to the likeliest state sequence $\boldsymbol{s}^*$ is taken as recognition result. Viterbi decoding works as follows:

$L = \{s_0\};\ a_0(s_0) = 1.0$

for t in {1,...,T}

$$a_t(s) = \max_{L \ni s' \to s} \{a_{t-1}(s')P(s|s')P(o_t|s)\}$$

$$b_t(s) = \operatorname*{argmax}_{L \ni s' \to s} \{a_{t-1}(s')P(s|s')P(o_t|s)\}$$

$$L' = \{s|s' \to s, s' \in L\};\ L = L'$$

$s' \to s$ refers to all possible transitions from state s' to state s. $a_t(s)$ is the probability of the best path of length t starting from s_0 and ending in s. The state sequence with maximum probability $a_T(s_e)$ can be reconstructed by recursively evaluating $s' = b_t(s); s = s'$ for $t = T$ down to 1 starting with $s = s_e$.

The number of current states in list L becomes too large in practice after a few iterations of the for loop. An approximative solution is the beam search. Instead of keeping all states in L and the corresponding values of $a_t(s)$ and $b_t(s)$, states s' which are endpoints of paths with

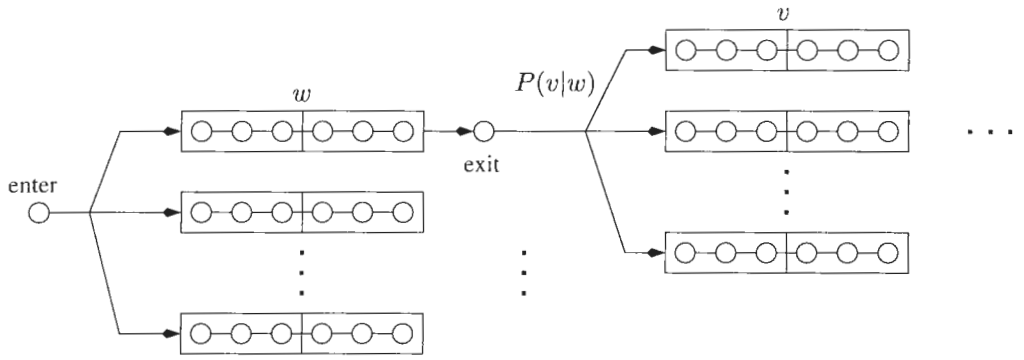


Figure 2.4: *Recognition network with word HMMs and word bigram transition probabilities. The circles in the rectangular boxes represent HMM states.*

a probability $a_t(s')$ smaller than $\max_{s \in L} \frac{a_t(s)}{k}$ are purged from L. The value of k is application dependent and has to be found experimentally [Jel97].

The hypotheses $\boldsymbol{w}$ corresponding to the N best state sequences $\boldsymbol{s}$ ordered by their acoustic probabilities $P(\boldsymbol{o}|\boldsymbol{s})$ are referred to as N-best list. N-best lists are used, for example, to calculate word confidence measures (cf. Section 2.2.1).

The decoder can also be employed to compute the forced-alignment of an utterance. Instead of a large word graph, only a small network is constructed, which consists of sequentially connected word HMMs corresponding to the utterance's known transcription. Additionally, there may be two or more HMMs in parallel, if the lexicon contains pronunciation variants, i.e. there is more than one possible subword sequence for certain words. Results of the forced-alignment are the acoustic likelihood and the duration (in number of frames) of each word or subword unit.

Recognition performance. There are several measures for evaluating the performance of a speech recognizer. They differ with respect to what kind of errors are taken into account. The correct rate

$$\text{Cor} = \frac{\#\text{cor}}{\#\text{ of tokens in reference}} \tag{2.7}$$

only measures the number of correctly recognized tokens (#cor). The accuracy is defined as

$$\text{Acc} = 1 - \frac{\#\text{sub} + \#\text{ins} + \#\text{del}}{\#\text{ of tokens in reference}} \tag{2.8}$$

and takes into account substitutions (#sub), insertions (#ins) and deletions (#del). In this thesis, the normalized minimum-edit-distance (Dist) [DHS01] between the reference and the recognized sequence is employed as feature to measure the recognition performance for pronunciation scoring. It is calculated as

$$\text{Dist} = \frac{\#\text{sub} + \#\text{ins} + \#\text{del}}{\max\{\#\text{ of recognized tokens}, \#\text{ of tokens in reference}\}} \tag{2.9}$$

2.2.1 Confidence Measures

Since speech recognition does not work perfectly, recognition errors occur. For each word of the recognition output it is important to obtain information about its degree of correctness. The degree of correctness can be measured by an estimate of the word posterior probability $P(\boldsymbol{w}|\boldsymbol{o})$. Additionally to classifying recognized words as correct or wrong, there are also attempts to determine whether presumably wrong words are insertions or substitutions [SSN+02]. In the following some approaches from literature to measure the confidence of recognition hypotheses on different levels are introduced. In this thesis it is investigated, whether these measures are also useful for the identification of mispronounced words.

Phoneme correlation technique. Cox et al. [CD02] describe several high-level confidence measures. High-level means that a measure's calculation does not depend on the decoding process or the decoder's architecture, but on final recognition results only. Here the phoneme correlation technique is introduced, for which only the best hypothesis of the word recognizer and the phoneme sequence obtained by free phoneme recognition with phoneme segment time intervals are needed.

Let $\boldsymbol{p} = (p_1, p_2, \ldots)$ be the phoneme symbol sequence corresponding to the best word hypothesis and $\boldsymbol{q} = (q_1, q_2, \ldots)$ be the phoneme symbol sequence corresponding to the phoneme recognition result. The sequences can be obtained on two different levels:

- Frame-level:
  each frame is tagged with the phoneme symbol of the segment it belongs to; for this operation the time interval of each phoneme segment is needed.

- Phoneme-level:
  the two phoneme sequences from word and phoneme recognition are aligned; any phonemes from the phoneme recognition result which cannot be paired are regarded as insertions and will be discarded.

The following definitions are independent from the level concerned. For both correctly $\mathcal{C}$ and incorrectly $\mathcal{I}$ recognized words a phoneme confusion matrix can be estimated on a training data set. Given the prior probabilities $P(\mathcal{C}), P(\mathcal{I})$ and confusion matrix probabilities $P(q_i|p_i, \mathcal{C}), P(q_i|p_i, \mathcal{I})$ for each phoneme pair, a likelihood ratio can be defined:

$$L_i = \frac{P(\mathcal{C}|q_i, p_i)}{P(\mathcal{I}|q_i, p_i)} = \frac{P(q_i|p_i, \mathcal{C})P(\mathcal{C})}{P(q_i|p_i, \mathcal{I})P(\mathcal{I})} \tag{2.10}$$

By summing up all L_i for a word, a confidence measure is obtained. The higher the sum of L_i the higher the possibility that the recognized word is correct.

Word posterior probability. The problem of speech decoding is to find the word sequence with maximum posterior probability (Eq. 2.11) for an observation sequence $\boldsymbol{o}$. However, for decoding itself, only the probabilities $P(\boldsymbol{o}|\boldsymbol{w})$ and $P(\boldsymbol{w})$ are important (cf. Eq. 2.1). The evidence $P(\boldsymbol{o})$ in the denominator need not be calculated because of the argmax operator.

The true value of the posterior probability, depends on the evidence however. $P(\boldsymbol{o})$ can be calculated as the sum of the probabilities $P(\boldsymbol{o}, \boldsymbol{v})$ for all possible word sequences $\boldsymbol{v}$.

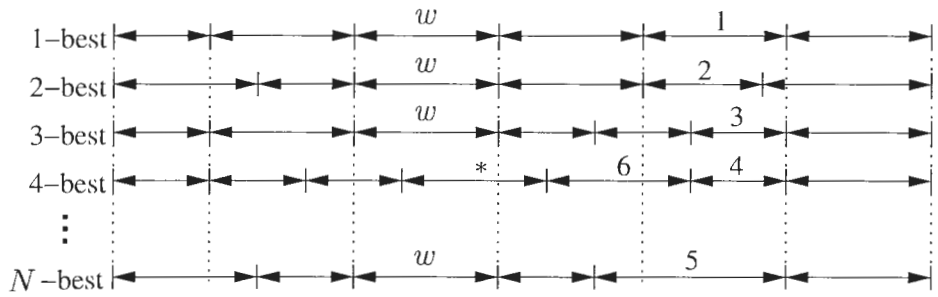


Figure 2.5: *Illustration of N-best lists. Word w appears in several hypothesis at the same time interval. Hence, it may have a high posterior probability.*

$$P(\boldsymbol{w}|\boldsymbol{o}) = \frac{P(\boldsymbol{o}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{o})} = \frac{P(\boldsymbol{o}|\boldsymbol{w})P(\boldsymbol{w})}{\sum_{\boldsymbol{v}} P(\boldsymbol{o}|\boldsymbol{v})P(\boldsymbol{v})} \tag{2.11}$$

Since the number of possible $\boldsymbol{v}$ is in theory almost infinite, the computation of word posterior probabilities is actually a difficult task. A discussion can be found, for example, in [WSMN01].

Here, the aim is to obtain the posterior probability of a particular word w_i in a sequence $\boldsymbol{w} = (w_1, \ldots, w_M)$ delivered by the word recognizer, given the acoustic observations of the whole utterance $\boldsymbol{o}$. Let $[w_i]$ be the word segment with label w_i and its starting and ending time. The posterior probability of w_i may then be defined as:

$$P([w_i]|\boldsymbol{o}) = \sum_{\boldsymbol{v}:\exists j: f([w_i]|[v_j])=1} P(\boldsymbol{v}|\boldsymbol{o}) = \frac{\sum_{\boldsymbol{v}:\exists j: f([w_i]|[v_j])=1} P(\boldsymbol{o}|\boldsymbol{v})P(\boldsymbol{v})}{\sum_{\boldsymbol{v}} P(\boldsymbol{o}|\boldsymbol{v})P(\boldsymbol{v})} \tag{2.12}$$

The function $f(\cdot|\cdot)$ returns 1, if $[w_i]$ and $[v_j]$ overlap, otherwise it returns 0. For the exact calculation of the posterior probability, $f()$ may only return 1, if the starting and ending times of both word hypothesis are identical. However this condition is unpractical. Consequently, $f()$ is usually relaxed to return 1, if there is only an overlap in time of e.g. $H = 75\%$. Since Equation (2.12) is intended to be the sum of posterior probabilities of all sentence hypotheses with overlapping w_i and v_j, H should be set to a value greater than 50%. An example is shown in Figure 2.5. For pairs of the edges 1, 2, 3, 4 and 5, the function $f(\cdot|\cdot)$ should return 1. However $f(1|6)$ should be 0, because e.g. edges 3 and 6, and edges 4 and 6 stretch over completely different acoustic segments.

When using the word graph, the word posterior probabilities can be computed efficiently during decoding with the forward-backward algorithm. They can also be calculated based on the N-best recognition output. If the N-best lists and the word graph are identical, the outcome is the same [WSMN01]. Despite the latter approach is less efficient, it is employed for this work in order to be recognizer independent.

In actual computation based on N-best lists the probabilities $P(\boldsymbol{o}|\boldsymbol{v})$ and $P(\boldsymbol{v})$ are rescaled. The scaling influences the performance of the resulting confidence measure. This is due to the large dynamic range of acoustic likelihoods associated with each hypothesis in the N-best list.

To sum up the probabilities of word hypotheses, likelihoods must be converted into probabilities with the e-function. Since likelihoods are large negative values, an underflow can easily occur if rescaling is not done.

Duration fluctuations. In [Ste01] word confidence measures based on duration ratios and duration fluctuations are investigated. Duration ratio means the quotient of the expected and the actual duration of phonemes or words. Since it describes the relative lengthening or shortening of tokens, it is also used as a relative measure of the speaking rate. Duration fluctuation means, that the duration ratio (speaking rate) changes over time. The incentive to use duration ratios as confidence measures is the observation, that recognition errors increase, if the average duration of phonemes in the training data deviates from durations in the test data.

2.3 Statistical Analysis

Given is a sample, a set of values, which are measurements of a (physical) quantity. Each value is drawn from a basic population with some probability. The basic population consists of all values which may be measured for the quantity. The aim of statistical analysis is to investigate the properties of the quantity, given only the sample. Simple examples for properties are the mean or the variance. Since only a subset of the whole basic population is known, properties can only be estimated. Moreover, measurements can be afflicted with errors, i.e. they may be skewed or censored. Consequently, the question of the accuracy of estimates arises. A standard approach to express the accuracy of an estimate is the indication of a confidence interval.

In the following methods for estimating the reliability of tests, resampling techniques to improve the estimates of statistics and methods for the calculation of certain confidence intervals are introduced briefly.

2.3.1 Reliability

The meaning of reliability actually depends very much on the context in which it is used. For example reliability is a very important concept in psychometrics. That science is concerned with measuring psychological aspects of a person such as knowledge, skills, abilities, or personality [Wik]. In this work the pronunciation skill of non-native speakers is to be assessed.

To obtain information about a person's skill or ability a test has to be carried out. Imagine, for example, a sprint test to measure the time a person needs to run a distance of 100 meters. Another example is the TOIEC or TOEFL test to measure certain aspects of a person's English proficiency. The outcome of such tests are scores, e.g. 10.5 seconds for the sprint test or 850 points for the TOIEC test. These observed scores may deviate from the true scores because of measurement errors:

$$\text{observed } x = \text{true } \tau + \text{error } \epsilon \tag{2.13}$$

Test results should be reliable and valid. "Reliable" can mean, that a test is stable, i.e. test results are reproducible, that it is consistent, i.e. each test item measures the same aspect, or that it is equivalent to another test. "Valid" means, that the test measures those aspects, the designer

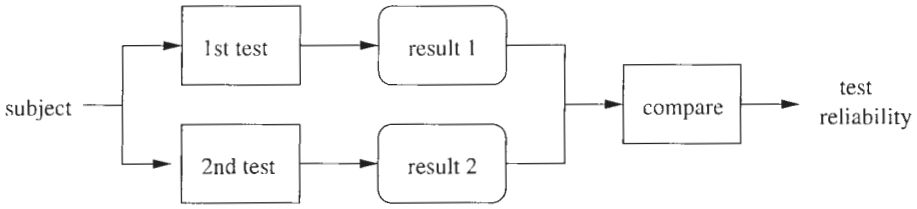


Figure 2.6: *Test reliability.*

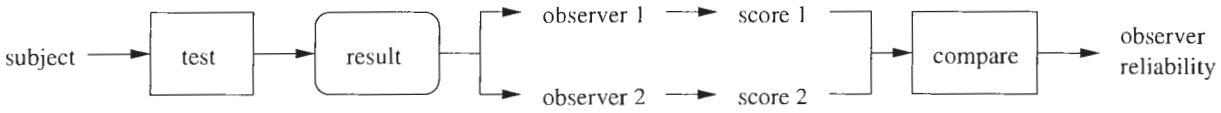


Figure 2.7: *Observer reliability.*

of the test intended to measure. To measure for example the stability a second test has to be carried out (cf. Figure 2.6).

For the evaluation of a speaker's pronunciation skill in this thesis the setting is different, however (cf. Figure 2.7). The test consists of reading a rather difficult set of English sentences. The test result consists only of the recorded utterances without any score. Their interpretation by an expert or a group of experts is needed. These experts (observers) listen to each of the recorded utterances and assign a score to each utterance, which is intended to measure a speaker's pronunciation quality. Whether the test itself is reliable and valid for assessing each speaker's pronunciation skill will not be an issue in this thesis. Since the main topic is automatic pronunciation scoring, the focus will be on a method which can accurately score an already given sample of read sentences. For that it is important to have reference scores for each sentence. The reference scores are assigned by one or more observers. If there are two or more observers, the observers' reliability can be estimated.

In test theory, reliability is defined as the ratio of the true score variance σ_τ^2 and the observed score variance σ_x^2:

$$\text{reliability } \rho_{xx} = \frac{\sigma_\tau^2}{\sigma_x^2} = 1 - \frac{\sigma_\epsilon^2}{\sigma_x^2} \tag{2.14}$$

If there are two parallel observations x_i and x_j and it can be assumed, that the covariances $cov(\tau, \epsilon_i)$ and $cov(\tau, \epsilon_j)$ of the true scores and the measurement errors are 0, the expectations $E[\epsilon_i]$ and $E[\epsilon_j]$ of the errors are 0 and the variances σ_{x_i} and σ_{x_j} are the same, the reliability equals the correlation between x_i and x_j [Veh00]:

$$\rho_{x_i,x_j} = \frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}} = \frac{cov[(\tau + \epsilon_i)(\tau + \epsilon_j)]}{\sigma_{x_i}\sigma_{x_j}} = \frac{\sigma_\tau^2}{\sigma_{x_i}^2} = \rho_{x_i,x_i} \tag{2.15}$$

2.3.2 Correlation

In general, the correlation is used to measure the similarity or dependability of variables or numerical series. The measurement for the self similarity of just one series is called auto-correlation. In case of two different series it is called cross-correlation. Here we are interested in the correlation of pairs of two different series. The correlation $C(X, Y)$ of two number series X and Y is defined as

$$C(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}} \tag{2.16}$$

$cov(X, Y)$ is the covariance of the two series, σ_X and σ_Y the standard deviation of each series. The mean of X is denoted as μ_X, the expectation as $E[X]$. If the probability of each sample x_i of X is uniform, calculation can be simplified:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_X)^2 \sum_{i=1}^{n}(y_i - \mu_Y)^2}} \tag{2.17}$$

$r_{X,Y}$ is called **correlation coefficient**. Its absolute values are between 0 and 1. If $r_{X,Y} = 1$, the points (x_i, y_i) of corresponding values of X and Y lie on a straight line [BS97].

To compare three or more numerical series, which are measurements of the same quantities, the **average correlation** of all possible number series pairs can be used. An alternative is to employ the average **open correlation**. The open correlation is obtained by first averaging all but one series and then calculating the correlation between the averaged series and the excluded series (see e.g. [NFDW00]). If there are K numerical series in total, the open correlation for the k-th series X_k is:

$$C^{(open)}(X_k) = C(X_k, \frac{1}{K-1}\sum_{i \neq k} X_i) \tag{2.18}$$

2.3.3 Resampling Techniques

Resampling is a procedure to generate several samples from an initial given sample. The presumably most widely known and most often applied techniques are jackknife and bootstrap resampling (Figure 2.8).

Jackknife. The jackknife is also known as "leave-one-out". This alternative name is due to the concrete resampling procedure: From a given sample $\mathcal{S}$ with $|\mathcal{S}| = n$ values, n new samples $\mathcal{S}_i$ are produced by removing the i-th value in each new sample. These samples can be used to calculate an estimate of a statistic Θ, which is more accurate than obtaining it directly from the values of the initial set $\mathcal{S}$. The calculation of the jackknife estimate is easy, albeit being computationally intensive: It is defined as the mean of the estimates $\hat{\Theta}_i$ for each sample $\mathcal{S}_i$:

$$\hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\Theta}_i \tag{2.19}$$

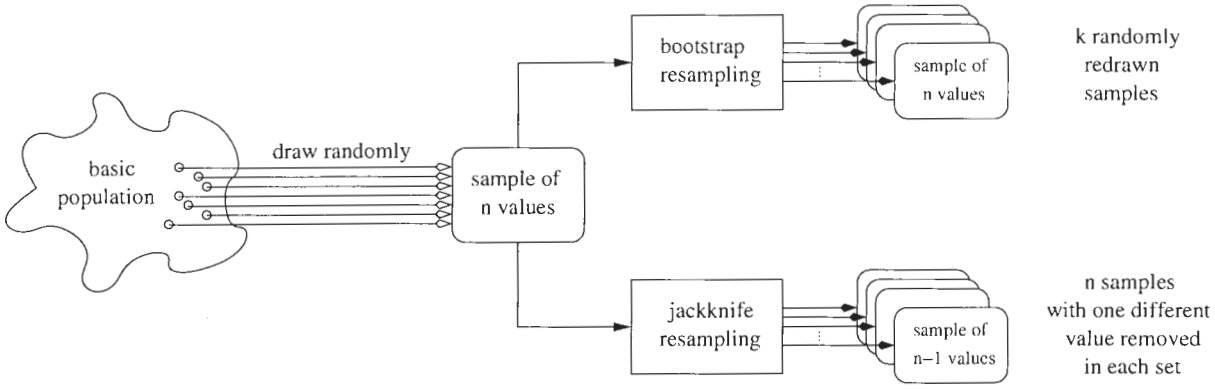


Figure 2.8: *Illustration of jackknife and bootstrap resampling.*

Bootstrap. The bootstrap method [ET93] is not deterministic like the Jackknife procedure: From the initial sample $\mathcal{S}$ with n values, k new samples are produced by randomly drawing n items from $\mathcal{S}$. Since items are selected with replacement, the same item may be drawn twice or more times. k should be set as large as the computational power available allows it. The bootstrap technique can also be employed for estimating a statistic like the Jackknife technique [DHS01]. Furthermore it is commonly used for the estimation of confidence intervals.

A straightforward and simple procedure to obtain a confidence interval is the percentile bootstrap method: Do bootstrap resampling of $\mathcal{S}$ to obtain e.g. $k = 1000$ samples with $|\mathcal{S}|$ items. Estimate the statistic Θ_i for each bootstrap sample $\mathcal{S}_i$. Order the values of Θ_i by their magnitude in ascending order. Let the significance level (cf. Section 2.3.4) be α. The lower bound of the confidence interval is then given by the $\lfloor k\frac{\alpha}{2} \rfloor$-th value, the upper bound by the $\lceil k(1-\frac{\alpha}{2}) \rceil$-th value.

2.3.4 Confidence

A statistic is estimated with a sample. The larger the sample, the more accurate the statistic's estimate will be. Accuracy of an estimate can be described by its confidence interval. The $N\%$ **confidence interval** of a statistic is the range in which the true value of a statistic lies with a probability of $N\%$. In the following the calculation of the confidence interval for the mean and the correlation is explained.

Mean. Given is the sample $\{x_1, \ldots, x_n\}$ of n values measuring the same quantity. Prerequisites are, that the scatter of the basic population is unknown and the values are distributed normal. The confidence interval $[\hat{x} - a, \hat{x} + a]$ of the mean $\hat{x}$ of the sample's values is calculated by

$$\hat{x} \pm a = \hat{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2,n-1} \qquad \hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{x})^2 \tag{2.20}$$

α is the probability of error. It is also called "significance level". The relationship between N and α is $\frac{N}{100} = 1.0 - \alpha$. The quantiles of the Student t-distribution are denoted as $t_{\alpha/2,n-1}$. In

practice the values of $t_{\alpha/2,n-1}$ have to be obtained by table-lookup [BS97].

Correlation. The formula for the calculation of the confidence interval for the mean is simple. If no formula is available or a statistic is complicated, the calculation of the confidence interval can be done via bootstrap resampling. In this work the percentile bootstrap method is used to estimate the confidence interval for the correlation (cf. Section 2.3.3).

2.3.5 Linear Regression

The correlation coefficient is a measure for the dependence of two variables x and y. If such a dependence is found, the next task is to determine the functional relationship $y = f(x)$. In general the problem is to determine the functional relationship $y = f(\boldsymbol{x}) = f(x_1, \ldots, x_K)$ between three or more variables. A pre-requisite here is to describe the relationship as:

$$y = f(x_1, \ldots, x_K) = a_0 + \sum_{i=1}^{K} a_i * g_i(x_1, \ldots, x_K) \tag{2.21}$$

The method to determine the coefficients a_i for known functions g_i is called linear regression. The functions g_i may be arbitrary. Given a set $\{(y^i, \boldsymbol{x}^i)\}$ of N samples the following system $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{a}$ of linear equations can be set up:

$$\begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} = \begin{pmatrix} 1 & g_1(\boldsymbol{x}^1) & \cdots & g_K(\boldsymbol{x}^1) \\ 1 & g_1(\boldsymbol{x}^2) & \cdots & g_K(\boldsymbol{x}^2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & g_1(\boldsymbol{x}^N) & \cdots & g_K(\boldsymbol{x}^N) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_K \end{pmatrix} \tag{2.22}$$

Multiplying each side of Equation (2.22) by $\boldsymbol{G}^T$ leads to the system of normal equations $\boldsymbol{G}^T\boldsymbol{y} = \boldsymbol{G}^T\boldsymbol{G}\boldsymbol{a}$. Cholesky's method is especially suitable to solve this system. See [BS97] for further details.

2.4 Pattern Classification

As literature about pattern classification confer [DHS01] and [Nie03]. Here only the basic concepts and classification methods employed in this work are explained briefly.

Let $\Omega = \{\omega_1, \ldots, \omega_K\}$ be a set of K classes. The classes are mutually disjoint and build a partition of the considered task domain. The task domain consists of a certain kind of patterns, e.g. 2D images or speech signals. A feature vector is denoted as $\boldsymbol{c}$. The features are intended to represent a pattern. They have to be designed in order to capture all relevant information. A numerical classifier for simple classification tasks takes as input a feature vector $\boldsymbol{c}$ and outputs a class label $\omega \in \Omega$ based on a decision rule and parametric models for each class. The model may for example consist of the parameters of a probability density function (e.g. Gaussian), the weights of a neural network (e.g. multi-layer perceptron) or if-then rules (e.g. decision tree). In the following section the classifiers employed in this work are described in more detail.

2.4.1 Classifiers and Decision Rules

In general the aim in classifier design is to optimize the decision rule and model for each class so that the classification risk becomes as small as possible. The risk R is defined as the average costs which go along with the classification of each pattern. It depends on the prior probabilities $p(\omega_\lambda)$ of each class ω_λ, the probabilities $p(\omega_\lambda|\omega_\kappa)$ that a pattern which belongs to class ω_κ is classified as a pattern of ω_λ, and the corresponding costs $r_{\lambda\kappa}$.

$$R = \sum_{\kappa=1}^{K} p(\omega_\kappa) \sum_{\lambda=1}^{K} r_{\lambda\kappa} p(\omega_\lambda|\omega_\kappa) \tag{2.23}$$

Bayes rule. If the costs are 1 for misclassification and 0 for correct classification, it can be shown, that the optimal decision rule is the Bayes rule

$$\kappa = \underset{\lambda\in\{1,\ldots,K\}}{\operatorname{argmax}}\ p(\omega_\lambda|\boldsymbol{c}) = \underset{\lambda\in\{1,\ldots,K\}}{\operatorname{argmax}}\ p(\omega_\lambda)p(\boldsymbol{c}|\omega_\lambda) \tag{2.24}$$

which assigns an unseen pattern to the class with maximum posterior probability. For application of the Bayes rule, the prior probabilities $p(\omega_\kappa)$ and the probability density $p(\boldsymbol{c}|\omega_\lambda)$ must be known. While the former can be obtained easily for each specific task domain, the latter may in general be difficult to determine.

Gaussian classifier. The term Gaussian classifier is employed here for the framework, that there is one model for each class ω_λ, which is based on a single Gaussian density or a Gaussian mixture density. Each model is intended to describe the distribution of feature vectors belonging to class ω_λ by a probability density function $p(\boldsymbol{c}|\omega_\lambda)$. The formula of a multivariate Gaussian density is given in Equation (2.25). d is the feature dimension, $\boldsymbol{\mu}_\lambda$ is the mean vector and $\boldsymbol{\Sigma}_\lambda$ is the covariance matrix of the Gaussian density $\mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)$.

$$\mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_\lambda|}} \exp\left[-\frac{1}{2}(\boldsymbol{c}-\boldsymbol{\mu}_\lambda)^T \boldsymbol{\Sigma}_\lambda^{-1}(\boldsymbol{c}-\boldsymbol{\mu}_\lambda)\right] \tag{2.25}$$

A mixture density is a combination of two or more base densities. The weighting of two or more Gaussians is called Gaussian mixture model (GMM). The GMM is able to approximate any probability density. The higher the number m of mixture components is, the more accurate the distribution of a training sample can be modeled. The weighting coefficients w_i are called mixture weights and must sum up to 1.0 to comply with the law of probability.

$$p(\boldsymbol{c}|\omega_\lambda) = \sum_{i=1}^{m} w_i\, \mathcal{N}(\boldsymbol{c}|\boldsymbol{\mu}_{i\lambda}, \boldsymbol{\Sigma}_{i\lambda}) \tag{2.26}$$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a Gaussian density can easily be obtained by ML estimation. For the mixture density case, however, no closed-form solution to parameter estimation exists. The difficulty is, that for each sample it is unknown which density influenced its generation to what extent. In literature algorithms for learning mixture model parameters can be found. The standard method for GMM training is the expectation-maximization (EM) algorithm [DLR77], which

starts with an initial guess of mixture parameters and successively converges to a maximum-likelihood (ML) estimate in each iteration. The drawbacks of the EM algorithm are, that it finds only suboptimal parameters in general. Furthermore, the number of mixture components must be set manually. The problem of mixture model learning is discussed in [FJ02]. That paper also proposes a method for estimating the density parameters and the number of mixture components at the same time. The new learning algorithm, which will be called FJ algorithm here, is employed to estimate mixture model parameters in experiments.

Decision trees. The prerequisite is, that each pattern is represented as attribute-value pairs, e.g. {color = red, taste = sweet}, or as a real-valued feature vector $\boldsymbol{c}$. A decision tree consists of nodes with questions and leaves with class labels ω and optionally class probabilities $P(\omega)$. Each question tests for an attribute or checks whether a component c_i of the feature vector $\boldsymbol{c}$ lies within a certain range. The test's outcome determines, which branch to follow up to the next question until a leaf node is reached. The class label associated with that leaf is taken as classification result. It is obvious that a decision tree can also be represented as cascaded if-then rules. If a pattern is represented as a n-dimensional vector with real-valued components, the decision tree actually defines regions for each class in $\mathbb{R}^n$.

While the procedure of how to classify a pattern is straightforward, the problem is to find automatically suitable questions for the tree nodes. There is a general framework called **CART** (**C**lassification **A**nd **R**egression **T**ree) for constructing a decision tree from a labeled sample.

The CART learning procedure starts with an initial tree with only one root node, which contains the whole training data. The aim is to partition the whole data into subsets associated with the leaves of the final tree, which are as homogene (pure) as possible w.r.t. to the patterns' class labels. A concrete algorithm requires:

1. a measure of impurity, e.g. leaf entropy
2. a set of questions to test, e.g. intervals for real-valued features
3. a criterion when to stop splitting leaf nodes
4. a method to prune a too large tree
5. a rule to label an impure leaf node

In each iteration, the set of questions is applied to the current leaf. The leaf is split according to the question which reduces impurity most. After the stopping criterion is reached, e.g. impurity or number of patterns associated with a leaf node falls below a threshold, the tree may be pruned. Pruning means to remove or merge leaves, if for example, the classification accuracy increases for a separate validation sample.

Further details can be found in [DHS01]. In this work the `wagon` tool from the Edinburgh Speech Tools Library is employed for CART. The tool can be set up to use either the correct rate or the entropy as impurity criterion for building classification trees, and the correlation or the mean square error (MSE) for construction of regression trees.

2.4.2 Feature Selection and Feature Transformation

Patterns are represented by features. The more features there are, the more accurate the representation of the pattern is going to be. On the other hand, the extraction of features is usually costly and the required amount of data for robust classifier training increases exponentially with the number of features. The solution would be to use the n best features available. The set of n best features has the property, that there is not another set with n or less features, which has a smaller classification risk. However, it is in general impossible to determine this set [Nie03].

In practice methods for feature selection and feature transformation are employed to reduce the dimension of the feature vector $\boldsymbol{c}$. The floating search algorithm is a heuristic selection procedure which often yields a set of good features. The principal component analysis (PCA) and linear discriminant analysis (LDA) are methods for feature transformation, which optimize the feature space with respect to certain criterions.

Floating search algorithm. Here only a rough description of the algorithm is given. The detailed algorithm can be found in [Nie03]. First some criterion to evaluate the quality of a feature set has to be defined. A reasonable criterion is for example the classification error of a test set or the classification risk. The algorithm works as follows:

1. Start with an empty feature set S.
2. Add relatively best feature to S.
3. Remove relatively worst feature from S, if quality of feature set becomes better than best set S' with $|S| - 1 = |S'|$ features so far.
4. Stop if a predefined number of features is used, else go to 2.

PCA. The principal component analysis (PCA) is an analytic method to obtain feature components, which are uncorrelated and ordered by the magnitude of their variances. It can be employed to calculate a basis of each k-dimensional subspace which is the best approximation of the original n-dimensional feature space ($k < n$) w.r.t. the mean square error (MSE).

From a training sample of feature vectors, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are calculated. The first k Eigenvectors ordered descending by the magnitude of the Eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ are a basis of the k-dimensional subspace. Original features vectors $\boldsymbol{c}$ are mapped into the PCA space by application of the Karhunen-Loéve transform:

$$\mathrm{KLT}(\boldsymbol{c}) = \boldsymbol{A}^T(\boldsymbol{c} - \boldsymbol{\mu}) \tag{2.27}$$

The columns of matrix $\boldsymbol{A}$ are the Eigenvectors belonging to the k largest Eigenvalues [DHS01].

LDA. Although PCA finds the best approximation of the original feature space in the MSE sense, it may happen that those dimensions, which are very important for the discrimination of certain classes are eliminated. The linear discriminant analysis (LDA) avoids that by seeking those directions, which are most effective for discrimination. The LDA transformation is given

as a matrix $\boldsymbol{W}$, which maximizes the between-class scatter $\boldsymbol{S}_B$ and minimizes the within-class scatter $\boldsymbol{S}_W$:

$$\boldsymbol{S}_W = \sum_{i=1}^{K} \sum_{\boldsymbol{c} \in \omega_i} (\boldsymbol{c} - \boldsymbol{\mu}_i)(\boldsymbol{c} - \boldsymbol{\mu}_i)^T \qquad \boldsymbol{S}_B = \sum_{i=1}^{K} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{2.28}$$

$\boldsymbol{\mu}_i$ is the mean vector of samples $\boldsymbol{c}$ of class ω_i, $\boldsymbol{\mu}$ is the mean of the whole sample and n_i is the number of samples available for class ω_i.

The rows of matrix $\boldsymbol{W}$ are the Eigenvectors corresponding to the largest Eigenvalues of the solution of the Eigenvalue problem in Equation (2.29). Like PCA, the dimension of the resulting feature space can be reduced by setting up $\boldsymbol{W}$ with only the first k Eigenvectors.

$$\boldsymbol{S}_W^{-1} \boldsymbol{S}_B \boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i \tag{2.29}$$

Chapter 3

Literature Survey

This chapter gives an overview of selected publications on human evaluation of pronunciation and automatic pronunciation scoring. Subjective and objective measures of pronunciation are introduced. Subjective measures are assessments of pronunciation made by humans. Measures which are calculated automatically on the speech signal with speech processing tools are called objective. By combining several objective measurements an automatic assessment of pronunciation quality can be obtained. Some methods from literature which effectively combined multiple pronunciation scores are presented.

3.1 Definition of Technical Terms

In literature about human evaluation of pronunciation and pronunciation scoring, different terms are used to refer to the same thing. Here, an attempt is made to unify the usage of technical terms for this thesis.

Units. Speech is made up of phones. By narrowing the view on speech to a certain natural language, it can be abstracted from certain sets of phones - the allophones - to phonemes. Phoneme strings form words, which may be separated by short pauses. Concrete phonemes or words are also called **tokens**. Strings of word tokens make up a sentence/utterance. Words may be separated by short pauses and utterances by silence. Consequently, the **phoneme-level**, the **word-level**, the **sentence/utterance-level** and the **speaker-level** are differentiated. On each level certain assessments of speech qualities can be made. Aspects of low-level units inevitably influence aspects of high-level units.

Pronunciation. The term pronunciation will be used for a wide range of speech quality aspects. These include not only segmental aspects like the concrete pronunciation of a word as a string of phonemes together with their realization, but also supra-segmental aspects like the speaking rate, pauses between phonemes or words, lexical stress of a word, sentence intonation, etc. (cf. Section 1.2)

Evaluator. A human being who listens to speech data of non-native speakers and assigns a label of pronunciation quality to certain evaluation units is called **evaluator** or **rater**. When reporting about the labeling procedure, the evaluator may also be called **annotator**.

Human Ratings. The labels of pronunciation quality made by humans are referred to as **ratings**. Ratings may be discrete, e.g. integer values from one to five, or continuous, e.g. values which lie inside a fixed interval $[a; b]$. Depart from that, the term phoneme-level or word-level **marking** is employed, if the label is binary, i.e. 1 if the token is mispronounced, and 0 if the token's pronunciation is correct.

Scoring. The process of automatic assessment of speech qualities by a machine is called **pronunciation scoring** or just scoring. The result of the automatic assessment are **scores** for certain aspects of pronunciation. A score may be based on only one or several features describing speech quality.

3.2 Performance Measures

There are several measures for comparing human evaluations and scoring algorithms. For example, Witt el al. [WY00] define four measures either for the consistency of phoneme markings, which are also meaningful for word markings, or the similarity of utterance ratings assigned by two or more human evaluators. A descriptive explanation of each measure together with its calculation formula are given here. The transcription vector with n components is denoted as $\boldsymbol{y} = (y_1, \ldots, y_n)$. There are as many components as tokens in an utterance. A component y_i is 0 if the corresponding token is not marked, i.e. it is correctly pronounced, and 1 if it is marked as mispronounced.

Strictness. How many of the tokens presented are marked as mispronounced? Since the decision to mark a token is subjective, the number of marked tokens may vary depending on how strict an evaluator is in correcting mispronounced tokens. Furthermore, the boundary between correct and wrong pronunciation may not be clear in some cases. The strictness is defined as the ratio between the number of rejected tokens and the total number of tokens.

$$S = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3.1}$$

Agreement. To what extent do two annotations of mispronounced tokens differ? The measure $A_{j,k}$ gives the relative share of tokens with the same annotation for a pair of evaluators (j, k). The more the evaluators agree with each other, the closer $A_{j,k}$ is to the value 1.0.

$$A_{j,k} = 1 - \frac{1}{n} \sum_{i=1}^{n} |y_i^j - y_i^k| \tag{3.2}$$

Correlation. The concept of correlation is introduced in Section 2.3.1. It is used to measure the reliability of scoring results by means of calculating the correlation between machine-based pronunciation scores and human ratings. Furthermore it is employed to assess the similarity of rating behavior of multiple human evaluations for the same material (see e.g. [NFDW00]).

Another example can be found in [WY00]. Witt el al. measure the overall similarity of rejection counts of mispronounced tokens marked by two evaluators. The rejection correlation is calculated as:

$$C_{j,k}^{(phone\_reject)} = \frac{\sum_{m=1}^{M}(c_m^j - \mu^j)(c_m^k - \mu^k)}{\sqrt{\sum_{m=1}^{M}(c_m^j - \mu^j)^2 \sum_{m=1}^{M}(c_m^k - \mu^k)^2}} \tag{3.3}$$

Here c_m^k denotes the number of rejections of phoneme m annotated by evaluator k. μ^k is the mean number of phoneme rejections by evaluator k, i.e. the sum of all rejections counts of evaluator k divided by the number of different phonemes M.

3.3 Human Evaluation

Human evaluation of pronunciation quality can be carried out on any level. The costs for evaluation on all levels are usually very high. Fortunately, it is possible to obtain a rating for a unit higher in the hierarchy, e.g. word or sentence, if annotations are available for a lower level only, e.g. phoneme or word, by combining the lower level ratings. In the following sections, a survey of methods for direct evaluation on sentence or phoneme level is given.

3.3.1 Sentence Level

Cucchiarini et al. [CSB00, CSBB00] report about human evaluation and automatic scoring of pronunciation in read and spontaneous speech. Here, only the part about human evaluation of read speech will be summarized, since spontaneous speech is out of the scope of this thesis.

From 60 non-native, 16 native and four native standard speakers two sets of five phonetically rich sentences (about 60 seconds of speech per speaker) were recorded. The target language is Dutch. Evaluation is carried out by one group of phoneticians and two groups of speech therapists. Pronunciation quality is assessed by four different aspects: **Overall pronunciation** (OP), **Segmental quality** (SQ), **Fluency** (FL) and **Speaking rate** (SR).

The grading scale for OP, SQ and FL ranged from 1 to 10. For SR a scale from -5 to +5 is used. The evaluation of OP, which was supposed to be an overall assessment of pronunciation, is carried out separately from the evaluation of the specific aspects of pronunciation, SQ, FL and SR. No specific instructions about the meaning of each scale are given to the evaluators. In order to put the human ratings on a common basis, each evaluator listened to uniform speech material from five different non-native speakers before proper evaluation started.

Table 3.1 shows the correlation between the four pronunciation measures. The correlation between the measures segmental quality (SQ) and overall pronunciation (OP) as well as between the measures fluency (FL) and speaking rate (SR) is highest. Least correlation was found between speaking rate (SR) and overall pronunciation (OP) or segmental quality (SQ), respectively. The authors point out that the high correlation between SQ and OP is in line with the general idea of pronunciation meaning "the degree of correctly articulating individual speech sounds". The fact that correlation among the four measures OP, SQ, FL and SR is rather high, is interpreted as

Table 3.1: *Correlation between different aspects of pronunciation and overall pronunciation score. Table is taken from [CSBB00]*

Correlation	Segmental Quality	Fluency	Speaking Rate
Overall Pronunciation	0.90	0.78	0.67
Segmental Quality		0.78	0.61
Fluency			0.88

evidence that the dependency between the different aspects of pronunciation are actually high, rather than that the raters have not correctly used the different metrics.

Measurements of inter-rater and intra-rater reliability were made. Reliability measured by Cronbach's α [Cro51] were reported to be satisfactory ($0.72 < \alpha < 0.99$). Furthermore, correlations between the ratings of different expert groups were shown to be high ($0.84 < r < 0.96$) after normalization by subtraction of mean and division by standard deviation is applied.

Neumeyer and Franco et al. [NFDW00] also report about human evaluation of whole utterances. In contrast to Cucchiarini et al. only one metric to measure the overall pronunciation quality is employed. For the study non-native French speech spoken by 100 natives from the U. S. is recorded. Each subject reads 5,089 sentences from newspapers with 14 words on average. Each sentence is rated on a scale from 1 to 5 by five teachers, who are native speakers of French. The average inter-rater correlation calculated on a common set of 342 sentences is 0.65 on the sentence and 0.8 on the speaker-level. The values for the average inter-rater open correlation are 0.76 and 0.87 respectively. Intra-rater correlation is 0.77 for 130 sentences evaluated twice by each rater.

3.3.2 Phone Level

Witt et al. [WY00] report about human evaluation and scoring of non-native pronunciation on the phone level. The database available consists of non-native English speech of ten students (6 female, 4 male) from several countries with an intermediate proficiency level. Each subject reads 40 phonetically balanced sentences and 80 sentences from "Penguin Readers", a text written for English teaching purposes with a limited vocabulary and simple grammar. Six trained phoneticians, who are native speakers of English, annotated insertions, deletions and substitutions of phones. All six evaluators labeled a set of 20 calibration utterances of one non-native speaker. The remaining speech material was split up, so that the material of each speaker was annotated by only one evaluator. The phoneticians were instructed to transcribe each phone as closely as possible to non-native speech sounds. Additionally each word and each sentence was graded on a scale from 1 to 4. However, only the annotations on the phone level were used for experiments.

The study reports acceptable values for correlation of phoneme rejections, correlation of annotations and agreement between phoneticians for the calibration sentences: About 10-25% of the phonemes were marked as mispronounced depending on each annotator's strictness. When considering all pairs of two different evaluators, the average values of the performance measures

were: agreement $A = 0.91$, phoneme rejection correlation $C^{(phone\_reject)} = 0.78$ and cross-correlation of annotations $C = 0.47$.

3.4 Pronunciation Scoring

In order to assess the pronunciation of a non-native speaker automatically, scoring metrics for pronunciation quality must be defined. The automatic assessment may take place on each of the different levels, i.e. phoneme-level, word-level or sentence-level. Pronunciation scores which have a high correlation with human ratings are favored to use them for a scoring method in a CAPT system (cf. Section 1.3). In the following sections several metrics for scoring pronunciation are introduced together with experimental results of correlation analysis.

The flow chart for pronunciation score extraction on the phoneme-level is shown in Figure 3.1. It illustrates all important processing steps, which have to be carried out to extract different kinds of pronunciation scores: The speech signal of the utterance to be analyzed is fed into a speech recognizer, which performs phoneme recognition and computes the forced-alignment, i.e. the segmentation into words and or phonemes together with an acoustic score of each token. Pronunciation scores are then calculated based on the segmentation, the recognition result, acoustic scores and statistics (e.g. phoneme duration) estimated on native speech data.

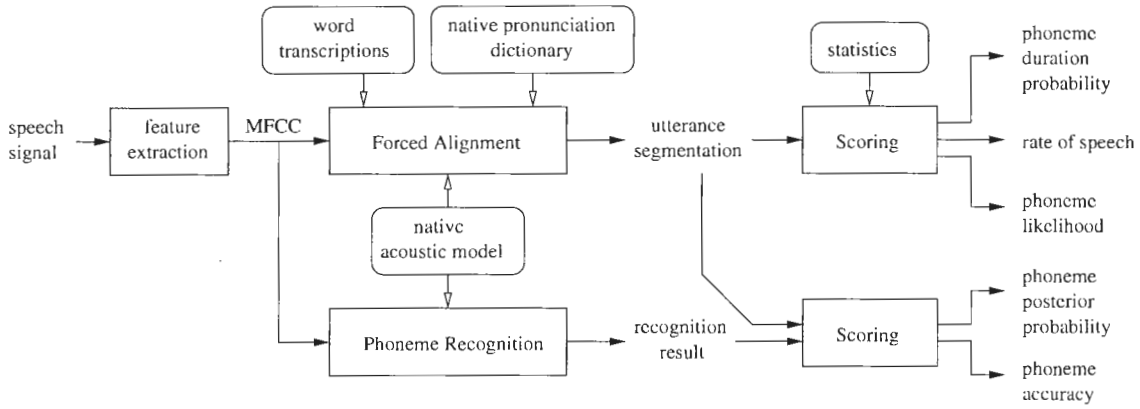


Figure 3.1: *Procedure for pronunciation scoring on the phoneme-level.*

3.4.1 Acoustic Scores

In literature three kinds of acoustic scores are employed to define segmental pronunciation scores. These acoustic scores can be calculated from the speech recognition result and the forced alignment.

- The likelihood $L(\boldsymbol{o}|\lambda)$ of observation $\boldsymbol{o}$ given acoustic model λ,
- the posterior likelihood $L(\lambda|\boldsymbol{o})$ of observation $\boldsymbol{o}$,

- the likelihood ratio $L(\boldsymbol{o}|\lambda_1) - L(\boldsymbol{o}|\lambda_2)$ of observation $\boldsymbol{o}$ given two models λ_1 and λ_2.

In the following passages, examples for the usage of each kind of acoustic score are given. Witt et al. [WY00] propose the **goodness of pronunciation** (GOP) metric for *scoring the pronunciation quality of single phonemes*. Their aim is to locate phonetic pronunciation errors, assess the closeness of each phoneme's pronunciation to a native speaker and identify systematic deviations in pronunciation. The GOP score together with a rejection threshold for each target language phoneme is applied to reject mispronounced phonemes. In the following a brief description of the basic GOP algorithm is given.

The utterance to be scored is segmented into phonemes by the force-aligning with a native acoustic model, the word level transcription and a native pronunciation dictionary. The acoustic model represents each phoneme q as one HMM λ_q. For each phoneme segment $\boldsymbol{x}$, the acoustic likelihood $\log P(\boldsymbol{x}|\lambda_q)$ for each phoneme model can be obtained. From these likelihoods the **segment posterior score** $S^{(gop)}(p|\boldsymbol{x})$ is calculated for the reference phoneme p. The definition of the GOP score is:

$$S^{(gop)}(p|\boldsymbol{x}) = \frac{1}{t}\log P(p|\boldsymbol{x}) = \frac{1}{t}\log\frac{P(\boldsymbol{x}|\lambda_p)P(p)}{\sum\limits_{q\in Q} P(\boldsymbol{x}|\lambda_q)P(q)} \approx \frac{1}{t}\log\frac{P(\boldsymbol{x}|\lambda_p)}{\max\limits_{q\in Q} P(\boldsymbol{x}|\lambda_q)} \tag{3.4}$$

Here t denotes the number of frames of segment $\boldsymbol{x}$ and Q is the set of all phonemes in the target language. The approximation in Equation (3.4) is obtained by assuming that the prior probabilities $P(q)$ of all phonemes $q \in Q$ are equal and substituting the segment evidence $P(\boldsymbol{x})$ by the probability $P(\boldsymbol{x}|\lambda_q)$ of the phoneme q with maximum segment likelihood $\log P(\boldsymbol{x}|\lambda_q)$. In practice, the calculation of the GOP score is carried out in the following way:

1. Obtain the forced-alignment (A) with the utterance transcription and a pronunciation dictionary.
2. Carry out free phoneme recognition (B) on the utterance, i.e. the recognition network is a phoneme loop.
3. Determine the likelihood of each frame for the phoneme model corresponding to the segment label in A and B.
4. Calculate the GOP score for each phoneme segment p in A. The nominator of Equation (3.4) is calculated by summing up the likelihoods of frames belonging to phoneme segment p in A the denominator by summing up the corresponding frame likelihoods of B.

Frame likelihood are easily obtained by dividing the segment likelihood with the segment length, measured in number of frames. The higher the value of $S^{(gop)}$ is, the closer is the pronunciation of p to the native sound. In order to reject or accept a speech sound, a score threshold can be employed. The simplest approach is to use a uniform global threshold. The value of this threshold depends on the desired strictness. However, the authors do not give further

Table 3.2: *Performance of human evaluation and the basic GOP algorithm; from [WY00]*

Performance	Agreement A	Correlation $C^{(phone\_reject)}$	Correlation C
Human Evaluators	0.91	0.78	0.47
GOP (global τ)	0.89	0.71	0.46
GOP (individual $\tau_p^{(a)}$)	0.88	0.57	0.46
GOP (individual $\tau_p^{(b)}$)	0.89	0.76	0.48

details about how to choose the global threshold effectively. Further considerations concentrate on the development of phoneme-specific thresholds. Since the distribution of GOP scores is different for each phoneme and class of phonemes, e.g., vowels vs. fricatives, two approaches to obtain phoneme-specific thresholds are examined:

- Linear combination of mean, standard deviation of the GOP score $S^{(gop)}$ and an additive constant

$$\tau_p^{(a)} = \mu_{S(p)} + \alpha\sigma_{S(p)} + \beta \tag{3.5}$$

- Phoneme rejection counts of human evaluators

$$\tau_p^{(b)} = \log \frac{1}{K} \sum_{k=1}^{K} \frac{c_k(p)}{\sum_{q \in Q} c_k(q)} \tag{3.6}$$

K denotes the number of speakers and $c_k(q)$ the number of rejections of phoneme q for speaker k. Q is the phoneme set of the target language.

Phoneme rejection with the GOP algorithm is evaluated with a global and individual rejection thresholds for each phoneme by comparing its performance to a human annotation, as it was described in section 3.3.2.

The performance of the basic GOP method with a global threshold is reported to be lower than the human evaluation in terms of the performance measures agreement $A = 0.89$, phoneme rejection count correlation $C^{(phone\_reject)} = 0.71$ and cross-correlation of annotations $C = 0.46$. The correlation values increase to $C^{(phone\_reject)} = 0.76$ and $C = 0.48$ respectively for individual thresholds $\tau_p^{(b)}$ based on the human rejection counts for each phoneme. However, performance is worse for individual thresholds $\tau_p^{(a)}$ based on mean and standard deviation of GOP scores (cf. Table 3.2).

In [WY97] Witt et al. report about evaluation results with a modified GOP metric $S^{(gop')}$ which also incorporates probabilities of phoneme strings in contrast to the previous simplifying assumption of equal prior probabilities. However, no increase in scoring accuracy could is achieved.

Neumeyer and Franco et al. [NFDW00, FNDR00] employ several kinds of duration-normalized likelihood scores for a *sentence-based pronunciation scoring* method.

With the forced-alignment of the target utterance, a likelihood score $L(\boldsymbol{x}) = \log P(\boldsymbol{x}|\lambda_p)$ can be calculated for each segment $\boldsymbol{x}$ given its phoneme label p and the corresponding HMM λ_p. $L(\boldsymbol{x})$ is the logarithmic probability of the best path through the HMM λ_p for segment $\boldsymbol{x}$. Since this acoustic score depends on the segment length, it is normalized by the segment duration. A score for a whole sentence $\boldsymbol{u} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is obtained by summing up likelihood scores of all utterance segments $\boldsymbol{x}_i$. Figure 3.2 illustrates the calculation of the utterance scores defined in Equation 3.7.

$$S^{(sent)}_{global}(\boldsymbol{u}) = \frac{\sum_{i=1}^{n} L(\boldsymbol{x}_i)}{\sum_{i=1}^{n} t_i} \qquad S^{(sent)}_{local}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \frac{L(\boldsymbol{x}_i)}{t_i} \tag{3.7}$$

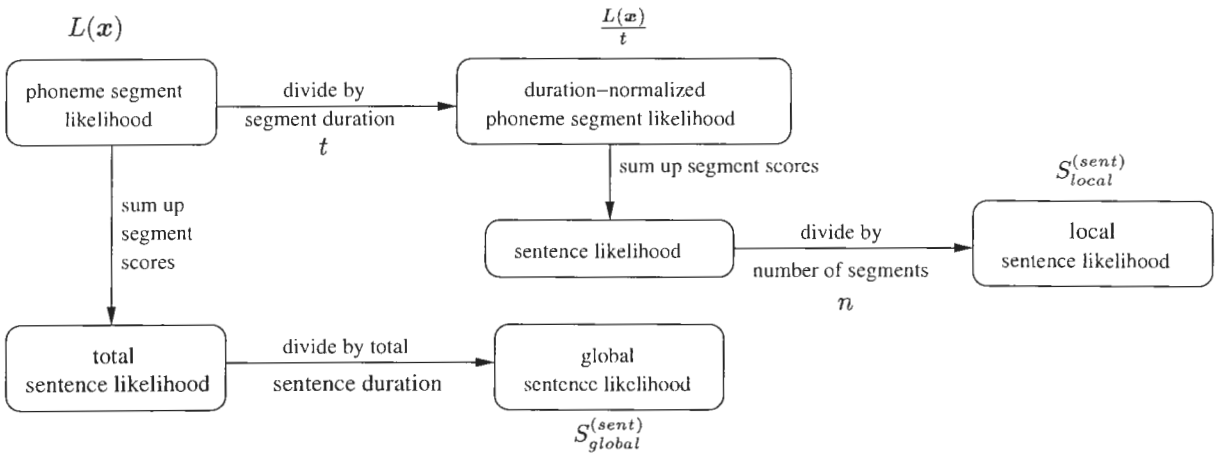


Figure 3.2: *Relationship between the likelihood scores of Equation (3.7).*

Almost similar to the GOP measure employed by Witt et al. a **frame-based phoneme posterior probability** $P(p|x_i)$ is formulated. Denoting the set of all phonemes with Q, the prior probability of phoneme q with $P(q)$ and the probability of the current observation x_i with $P(x_i|q)$, it is defined as:

$$P(p|x_i) = \frac{P(x_i|p)P(p)}{\sum_{q \in Q} P(x_i|q)P(q)} \tag{3.8}$$

By summing up the logarithm of $P(p|x_i)$ for all frames of the segment $\boldsymbol{x}$ and dividing by the number of frames t of segment $\boldsymbol{x}$ a duration-normalized segment score $S^{(segment)}(\boldsymbol{x})$ can be calculated. A **sentence posterior score** is then available via

$$S^{(sent)}_{post}(\boldsymbol{u}) = \frac{1}{n}\sum_{j=1}^{n} \frac{S^{(segment)}_{post}(\boldsymbol{x}_j)}{t_j} \qquad S^{(segment)}_{post}(\boldsymbol{x}) = \sum_{i=1}^{t} \log P(p|x_i) \tag{3.9}$$

The posterior probability $P(p|x_i)$ is calculated in the same way as for the GOP measure, which was already described in previous passages. However, the calculation of the posterior

probability employed for computation of $S_{post}^{(sent)}$ is more precise, since no approximation of the denominator (see Equation 3.4) is done.

The difference $L_1 - L_2$ of two log-likelihoods L_1 and L_2 is called **likelihood ratio** (LR). L_1 and L_2 are scores for a *single speech frame* or a *whole phoneme or word segment.* The LR is employed to compute a pronunciation score from two lattices for the same utterance. The lattices may originate from a forced-alignment or from word or phoneme recognition. Furthermore different acoustic models may have been used for alignment or decoding. The subtraction of L_2 from L_1 can also be regarded as normalization of score L_1. The dynamic range of acoustic scores is very large. This brings about the problem that the acoustic score for one segment influences the scores of other segments, e.g. when calculating a sum score for a whole utterance. The LR alleviates this problem, because the likelihoods L_1 and L_2 are of the same order of magnitude and the LR is likely to have smaller values.

For example, Neumeyer et al. [NFDW00] report that the rating to score correlation increases if using the likelihood ratio $S_{CD/CI}^{(ratio)}(\boldsymbol{u})$, which is defined in Equation (3.10), instead of the local average likelihood $S_{local}^{(sent)}(\boldsymbol{u})$. $L_{CD}(\boldsymbol{x})$ is the likelihood of segment $\boldsymbol{x}$ for the context-dependent (CD) model, $L_{CI}(\boldsymbol{x})$ for the context-independent (CI) model.

$$S_{CD/CI}^{(ratio)}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}\frac{L_{CD}(\boldsymbol{x}_i) - L_{CI}(\boldsymbol{x}_i)}{t_i} \tag{3.10}$$

Nakagawa et al. [NMN03] also employs several likelihood ratio scores, e.g. the ratio of a native and a non-native acoustic score. The rating-score correlation could be improved from 0.30 for a traditional duration-normalized likelihood score to 0.50 for the LR.

3.4.2 Rate of Speech

The rate of speech (ROS) measures how many speech tokens a speaker produces during a certain time interval. The rate of speech in terms of phonemes $R^{(phon)}$ is defined as the average number of phonemes articulated every second [NFDW00], the word-based rate of speech $R^{(word)}$ as the average number of words per second. Which measure is preferable depends on the application. For example, an important normalization of phoneme durations can be achieved by multiplication with $R^{(phon)}$ (cf. Section 3.4.3).

There are other measures similar to rate of speech, which are defined in [CSB97] and [CSBB00]. For example, the phonation-time ratio and the articulation rate. While the rate of speech is defined as the ratio of number of phones and duration of speech with pauses, articulation rate is defined as the ratio of the number of phones and the duration of speech without pauses. The phonation-time ratio is the ratio of the duration of speech without pauses and the duration of speech with pauses. Experimental results in [CSBB00] showed, that the correlation between these ROS measures and the aspect fluency (FL) and speaking rate (SR) of pronunciation is highest, varying between 0.8 and 0.9. The correlation for segmental quality (SQ) is between 0.6 and 0.7.

3.4.3 Duration and Timing Scores

Duration of tokens and pauses between tokens in non-native speech may significantly differ from native speech. This phenomenon is caused by the unfamiliarity of a non-native speaker with certain words or speech sounds. Consequently, a non-native speaker needs more time to think about how to articulate [NFDW00].

A phoneme duration statistic is estimated from the phoneme durations obtained by forced-alignment of native speech data with a native acoustic model. The duration probability $P_{dur}^{(phon)}(t|p, \boldsymbol{x})$ of a phoneme p with duration t is modeled with a histogram function. The sentence level **duration score** $S_{dur}^{(sent)}(\boldsymbol{u})$ is defined by the sum of logarithmic duration probabilities of each phoneme segment $\boldsymbol{x}$ of the utterance $\boldsymbol{u}$.

$$S_{dur}^{(sent)}(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \log P_{dur}^{(phon)}(f(t_i)|p_i, \boldsymbol{x}_i) \tag{3.11}$$

$f(t)$ is a function for duration normalization. For a text-independent normalization Neumeyer et al. propose to multiply the duration t of each phoneme segment $\boldsymbol{x}$ by the rate of speech R, leading to the Equation $f(t) = t * R^{(phon)}$.

The typical length of syllables is different for each language. Neumeyer et al. [NFDW00] examined this phenomenon for scoring. They define syllabic periods as the time interval between the center vowels of two syllables. These syllable durations are also normalized by multiplication with the rate of speech. Scores based on probabilities of syllable durations are referred to as timing scores.

3.4.4 Recognition Accuracy

Recognition performance for non-native speech is usually lower than for native speech when recognition is performed with a native acoustic model. The closer a non-native speaker's pronunciation is to standard pronunciation, the better speech recognition will work and the higher recognition accuracy that can be expected. Consequently, recognition accuracy may be a good indicator of a non-native speakers pronunciation proficiency [NFDW00].

3.4.5 Other Prosodic Features

In order to further improve the scoring based on the three most approved features, posterior phoneme likelihood, phoneme duration and timing scores, Teixeira et al. [TFS+00, TFS+01] investigated pronunciation scoring with prosodic features. The features they employ are based on the fundamental frequency f_0, pause statistics and also include lexical stress information. Except for some pause-related features, e.g. duration of the first or second longest inter-sentence pause or time interval between two pauses, the correlation between each prosodic feature and human ratings is less than 0.3. Moreover, when combining the three most approved features with one or more of the new prosodic features only very slight improvements in the human rating to score correlation can be observed.

Table 3.3: *Comparison of several pronunciation scores; from [NFDW00].*

Correlation of scores with ratings	Sentence level	Speaker level
Global likelihood $S_{global}^{(sent)}$	0.18	0.31
Local likelihood $S_{local}^{(sent)}$	0.29	0.48
Posterior likelihood $S_{post}^{(sent)}$	0.52	0.84
Phoneme recognition accuracy	0.40	0.47
Duration score $S_{dur}^{(sent)}$	0.41	0.86
Human evaluation	0.65	0.80

3.4.6 Correlation with Human Ratings

Neumeyer et al. [NFDW00] analyzed the correlation between human ratings and the different scores explained in this section. The analysis was based on 100 non-native speakers and 30 utterances per speaker. Results are summarized in Table 3.3. On the sentence level, correlation of any scores with the human ratings is lower than the average inter-rater correlation. The posterior likelihood and the phoneme duration score lead to the best result for both sentence level and speaker level correlation. The correlation values for these two scores are even better than the speaker level inter-rater correlation (though not significantly better w.r.t. to the number of speakers). Adding more scores like global likelihood or phoneme recognition rate did not improve correlation with human ratings.

3.5 Combination of Scores and Classification

From section 3.4.1 through section 3.4.4 several scores (e.g. likelihood, duration) for automatic assessment of pronunciation on different levels (e.g. phoneme, sentence) were introduced. To build a robust system for automatic pronunciation scoring, scores, which proved to have a high correlation with human ratings, have to be combined.

When regarding scores as features and scoring as classification, the task is to construct a classifier which is able to assign one out of several discrete levels of overall pronunciation quality to each token like a human rater who selects a grade e.g. from 1 to 5. Franco et al. [FNDR00] define the problem of automatic pronunciation assessment as estimation of a random variable which measures a particular or the overall pronunciation skill.

When denoting the pronunciation rating of a token with h and the different scores with s_i, the task is to find a functional dependency

$$h = f(s_1, \ldots, s_m) \tag{3.12}$$

between the scores and the human ratings. Since in reality this kind of mapping is usually not perfect, the goal is to find at least a good approximation $h \approx f(s_1, \ldots, s_m)$ or a function f which correlates well with h. When applying the mean square error (MSE) criterion to Equation (3.12), the following optimization problem has to be solved:

$$f^* = \min_f E[h - f(s_1, \ldots, s_m)]^2 \qquad h^* = E[h|s_1, \ldots, s_m] \tag{3.13}$$

The solution is the conditional expected value h^* of the human rating h given the scores s_i. In the following, several approaches employed by Franco et al. [FNDR00] to determine $f(s_1, \ldots, s_m)$ are explained in more detail. They differ in the assumptions which can be made about the score distributions or about the relationship between human ratings and scores.

Distribution Estimation. The conditional expectation can directly be calculated via the conditional probabilities $P(h|s_1, \ldots, s_m)$. Applying the Bayes rule to $P(h|s_1, \ldots, s_m)$ in Equation (3.14) leads to Equation (3.15) with the probability density functions $P(s_1, \ldots, s_m|h)$.

$$h^* = E[h|s_1, \ldots, s_m] = \sum_h h * P(h|s_1, \ldots, s_m) \tag{3.14}$$

$$P(h|s_1, \ldots, s_m) = \frac{P(s_1, \ldots, s_m|h)P(h)}{P(s_1, \ldots, s_m)} = \frac{P(s_1, \ldots, s_m|h)P(h)}{\sum_g P(s_1, \ldots, s_m|g)P(g)} \tag{3.15}$$

Since these probability functions are not known a priori, they have to be estimated. One possibility is to approximate these functions with discrete probability distributions. Therefore the score space spanned by the scores s_i is discretized by vector or scalar quantization (VQ). The probabilities $P(V(s_1, \ldots, s_m)|h)$ can be obtained by counting the number of occurrences of each VQ index $V(s_1, \ldots, s_m)$ for each human rating h and then dividing each count by the number of occurrences of each corresponding human rating.

Linear Combination. Assuming there is a linear relationship between the human ratings or the scores are distributed normal, the conditional expected value of h is a linear combination of the scores s_i. Their weightings a_i and the bias b can be obtained by linear regression (cf. Section 2.3.5).

$$h = a_1 s_1 + a_2 s_2 + \cdots + a_M s_M + b \tag{3.16}$$

Artificial Neural Networks. In the general case, i.e. any kind of score distribution and any kind of relationship between human ratings and scores, a nonlinear function $f(s_1, \ldots, s_m)$ has to be estimated. Neural networks with a sufficient number of hidden layers and suitable neuron activation functions are in general capable of approximating any function [DHS01]. The input to the network are the scores s_i. For the number of outputs there are two choices: Only one output neuron for all values of h, or one output neuron for each human rating value, i.e. if there are five human grades there will be five output neurons.

Regression tree. An alternative to estimate the distributions $P(h|s_1, \ldots, s_m)$ of the human grades as described before is to employ a regression tree. Such a tree can be constructed with the CART framework (cf. Section 2.4.1).

Results. Franco et al. [FNDR00] carry out numerous experiments for different combinations of scores with the approaches described in the last passages. Their objective was sentence-level scoring on a discrete scale from 1 to 5. The criterion for selection of the scores to be combined

is a high correlation of these scores with human ratings and for each additional score a low correlation with already employed scores. Important results are summarized in Table 3.4.

Best results were achieved with a neural network leading to a correlation of 0.64 between scores and human ratings, which is almost as high as the inter-rater correlation of 0.65 (cf. Table 3.3). The performance of score combination with regression trees and discrete distributions was slightly lower than neural networks but still acceptable. Furthermore, construction of distribution-based or regression-tree-based classifier is much less computationally intensive and needs less experimentation than neural network training and design.

Table 3.4: *Combination of several scores lead to a higher correlation between scores and human ratings. The objective was sentence-level scoring on a discrete scale from 1 to 5. Best performance was achieved with a neural network; from [FNDR00].*

Scores	Combination Method	Correlation
Posterior score (P)	None	**0.58**
Duration score (D)	None	0.47
Timing score (T)	None	0.35
P + D + T	Linear Combination	0.59
P + D + T	Discrete Distribution	0.62
P + D + T	Regression Tree	0.62
P + D + T	Neural Network	**0.64**

3.6 Speaker Scoring

Recently Minematsu [Min04] proposed a method for obtaining a speaker score directly from a speaker-dependent acoustic model without using any information from sentence or lower level scoring. The motivation of this approach is the argument, that speech recognition technology is still unreliable and a more robust scoring paradigm should be employed for a CALL system.

The scoring method is based on a distortion measure. Initially, a simple speaker dependent acoustic model is constructed from 60 phonetically balanced sentences. The model consists of 3-state HMMs with single mixture Gaussian distributions of spectral features. The distance between two phonemes can be defined via the Bhattacharyya distances [Nie03] between the two model distributions. Minematsu shows, that the Bhattacharyya distance is invariant against any affine transformations, i.e. rotation, shear and shift, of the underlying spectral feature space. The distances between each phoneme model pair define a so-called "universal structure". This universal structure can be calculated for each native and non-native speaker.

A structural distortion measure can be defined between two structures. To measure this distortion the structures have to be aligned. The center of gravity of the structures is calculated by shifting one structure that the two centers of gravity fall together and rotating them until the sum of angles between each pair of edges becomes minimal. The structural distortion can then be expressed as the sum of edge pair length differences.

This structural distortion measure had a correlation of -0.88 with human ratings for a set of 200 non-native Japanese speakers of English. While for inter-speaker distances based on the average HMM distance there were almost no differences for American-Japanese and American-American speaker pairs, the distribution of American-Japanese inter-speaker structural distortions for vowels were different from American-American distortions.

3.7 Conclusion

In previous work on pronunciation scoring, which is described in the last sections, methods for detecting mispronounced phonemes, and methods for the automatic assessment of the pronunciation quality of utterances and the pronunciation skill of speakers have been developed. Promising results are achieved, i.e. the agreement between human raters and the automatic method w.r.t. mispronounced tokens is high, and there is also a rating to score correlation comparable to the inter-rater correlation at the utterance level.

The pronunciation scores with the highest correlation to human ratings for the overall pronunciation quality of utterances are the sentence posterior score, the recognition accuracy, and duration and timing scores. Scores based on prosodic features only had a low correlation with the human ratings.

Investigations in previous work are carried out for adult speech of either many adult speakers having the first language in common or few adult speakers with different first languages. For this thesis two databases are available: The ATR SLT non-native speech database with English speech data with about 100 non-native English speakers data from several accent groups and the FAU LME non-native children speech corpus contains speech of German children reading English texts.

Since there are annotations available on the word, utterance or speaker level, the main tasks in this work are the detection of mispronounced words, scoring the pronunciation quality of utterances and obtain an assessment of a speaker's pronunciation skill. Additionally, the human annotations of the ATR SLT database are analyzed in detail.

Chapter 4

Speech Data

A speech database with native English speech for building the acoustic model of a word recognizer, and a non-native English speech database with human annotations regarding pronunciation quality are necessary for conducting pronunciation scoring experiments. This chapter describes each database. Results of overall analysis for the human labels of the ATR SLT non-native speech data are also reported.

4.1 Non-Native Speech Data Collected at ATR SLT

Speakers and Contents. English speech from non-native and a few native speakers was collected at the Spoken Language Translation Research Laboratories (SLT) of ATR. The subjects are from several countries with different mother tongues. The first language of most speakers was Japanese, Chinese, German, French and Indonesian (cf. Table 4.1). About half of the subjects are or were members of ATR at the time of the recording. The remaining subjects in the database were hired from an agency. Each subject had to read a uniform set of about 150 English sentences. There are 25 utterances with credit card numbers, 48 phonetically rich sentences (TIMIT SX set) and six hotel reservation dialogs with 73 sentences in total (see overview in Table 4.2). The total duration of all SX sentences of the 96 non-native speakers is 6.4 hours.

Table 4.1: *Distribution of speakers' first languages in the ATR SLT database.*

NNS-DB	# Male	# Female	Age
Chinese	16	2	21-52
French	15	1	21-42
German	15	0	23-43
Indonesian	15	1	24-43
Japanese	15	9	21-45
other	5	2	31-42
total	81	15	21-52

Table 4.2: *Detailed contents for each speaker of the ATR SLT non-native English database.*

Set	Contents	# Words	# Utterances
SX	48 phonetically rich sentences from the TIMIT database	395	48
TAC22012	Hotel reservation dialogs	252	19
TAS12008		104	9
TAS12010		144	12
TAS22001		162	10
TAS32002		182	13
demo02		70	10
DIGITS	Credit card numbers	200	25

There are also recordings of some sentence sets for a few native speakers in the database. The data of these speakers are used for validation of the pronunciation scoring system.

Recording Conditions. Recording was carried out in an acoustic booth with reverberant-free walls. As recording equipment a Sennheiser close-talking microphone (HMD-410) and a Sony DAT Recorder (DTC-2000ES and TCD-D100) were employed. The speech data was downsampled to 16-kHz with 16-bit precision.

Each sentence a subject had to read was displayed on a computer screen. The subject started and stopped the recording by himself. The recording was supervised by ATR staff, including the author. Each subject reads each sentence usually only once for recording. However, he/she was asked to repeat the recording of a sentence, when he/she completely misread a word, forgot to utter a word, made too long pauses between words or noise during recording. Furthermore, a subject was allowed to repeat the recording of a sentence, if he/she was not satisfied with the recorded utterance.

Evaluators and annotations. For a pronunciation scoring system human made labels are necessary as a reference in order to validate the system. 15 English teachers were hired via an agency each for seven hours (six hours working time, one hour break). All teachers were native English speakers from North America. Further information about each annotator, e.g. English teaching experience, can be found in Appendix B. Each evaluator had to listen to 1,152 utterances (48 TIMIT sentences times 24 non-native speakers) in order to assign a utterance-level rating from 1 (best) to 5 (worst) in terms of pronunciation and fluency to each utterance and mark any words which are mispronounced. In total the speech data of 96 non-native speakers was evaluated, i.e. since there are 15 evaluators and each evaluator annotated the data of 24 non-native speakers, each speaker is assessed by three to four evaluators. The next passages will explain details and results of the human evaluation procedure.

Human evaluation procedure. Each evaluator was given an instruction sheet as it is shown in Appendix B. In the beginning, each evaluator had to listen to a uniform set of 22 calibration sentences. This set consists of 22 different sentences of 22 different non-native speakers. The evaluator had to assign a rating from 1 to 5 to each calibration sentence considering pronunciation

and fluency. The selection criterion for the speakers was their recognition rate in the hotel reservation tasks TAS22001 and TAS32002 with a native acoustic model. For each of the five major first languages, one speaker with the highest, one speaker with the lowest and one speaker with a medium recognition rate were selected. Furthermore, one sentence of one speaker each of the remaining first languages (Bulgarian, Spanish, Portuguese, Hindi, Sinhalese, Hungarian and Korean) was included. Since the evaluators were asked to assign level 1 to the best utterances and level 5 to the worst utterances and to use each level at least once, they had an anchor for future ratings.

In the next step 48 sentences of 24 speakers were presented separately to the evaluator. The order of presentation was determined randomly and was different for each evaluator. The evaluator had first to listen to an utterance, mark mispronounced words and finally select a level of proficiency. For these utterance level ratings, the evaluators were instructed to ignore sentence intonation, for marking of words to consider phonetic pronunciation errors but to ignore wrong lexical stress. The evaluator was not allowed to go back to already processed utterances and change previously assigned labels. See Appendix B for screen shots of the web interface used for human evaluation. Results of the human evaluation are discussed in Chapter 5.

4.2 FAU LME Children Speech Corpus

The corpus contains speech data from 57 German children. There are 26 male and 31 female children. The age of the children is between 10 and 15. Each child reads English sentences. A large number of sentences consist only of single words. The sentences are made from a vocabulary of 1,077 words. There are 4,630 utterances or 3.4 hours of speech in total.

The corpus has been annotated by one German student of Anglistics. Instead of rating each utterance separately, the annotator assigned only one overall rating of pronunciation to each speaker. Most of the speaker level ratings are discrete on a scale from 1 to 5. There are only three ratings of the form "a-b". These ratings will be treated as $\frac{a+b}{2}$. The distribution of the ratings is shown in Figure 4.1. The ratings are almost distributed normal with a mode of 3.

Furthermore the annotator marked mispronounced words and transcribed insertions, substitutions and deletions of words, non-speech noise and garbage words. A statistic of the relative number of words considered mispronounced for each speaker is also shown in Figure 4.1. The distribution is different from the distribution of the ratings, but there is still a correlation

Table 4.3: *Correlation between the relative number of marked or substituted words (MisRatio), the speaker level human rating (HumRating), the word accuracy (WordAcc) with a unigram language model, and the phoneme accuracy (PhonAcc) with a zerogram language model (FAU LME data).*

Correlation	HumRating	WordAcc	PhonAcc
MisRatio	+0.67	-0.52	-0.61
HumRating		-0.49	-0.49

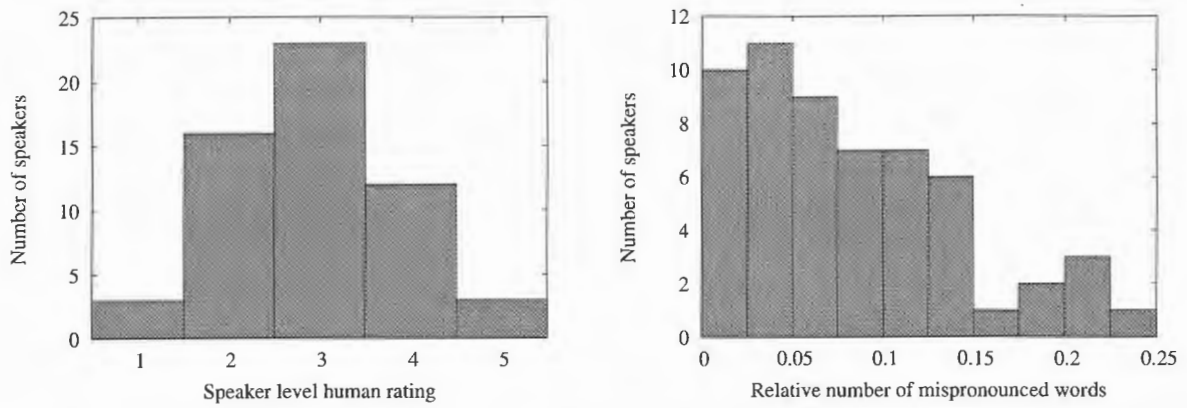


Figure 4.1: *Distribution of the ratings and the relative share of marked or substituted words for each speaker (FAU LME data).*

of 0.67 between the relative number of words marked and the speaker rating (cf. Table 4.3). The correlation between the speaker rating and the word or phoneme accuracy are both -0.49. These values of the correlation coefficient are lower than for ATR SLT data, as the results in Section 6.2.4 will demonstrate.

Any word which is marked as mispronounced or which is transcribed as an insertion or substitution is considered as candidate for the class $\mathcal{X}$ of mispronounced words in the experiments. Furthermore garbage words, e.g. words which are aborted during reading, are also members of class $\mathcal{X}$.

4.3 Wall Street Journal (WSJ) Corpus

The Wall Street Journal Corpus was collected in two phases: the pilot project CSR-WSJ0 in 1991 and the main project CSR-WSJ1 from 1992 to 1993. The collection has been sponsored by the Advanced Research Projects Agency (ARPA) and the Linguistic Data Consortium (LDC) and has been carried out by MIT, Texas Instruments and SRI International. WSJ1 contains about 73 hours (approx. 78,000 utterances) of speech for training and 8 hours (approx. 8,200 utterances) of speech for testing purposes. Most of the training data is read speech style. However most of the test data are from spontaneous dictation by journalists.

For this thesis only read speech data from non-journalists subjects is employed for training and evaluation of a speaker-independent acoustic model. The data was recorded with a Sennheiser close-talking head-mounted microphone. The speakers are from North America and read texts from the Wall Street Journal. The selected training set consists of about 30,000 utterances of 200 speakers from WSJ1 and 7,200 utterances from WSJ0.

In order to evaluate the performance of the acoustic model built with the training data, the Hub2 test set was used. It comprises 20 utterances of ten native speakers each. The test set is designed for a 5000 word bigram closed-vocabulary grammar.

4.4 WSJ Cambridge Corpus

The WSJCAM0 corpus is derived from the Wall Street Journal text corpus. It consists of British English speech recorded from 140 native speakers uttering about 110 utterances each. There are 92 training speakers. The age of most training speakers is between 18 and 28. Only 16 speakers are of age 29 or older. There are more than 18 hours of training speech data in total. Recordings were carried out in an acoustically isolated room and made with a Sennheiser close-talking head-mounted microphone [RFP+94].

4.5 TIMIT Corpus

The TIMIT corpus was sponsored by the Defense Advanced Research Project Agency (DARPA) and set up by MIT, SRI and Texas Instruments. The corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from eight major dialect regions of the U.S. These 10 sentences are from three different sets:

- SA: dialect sentences, which are meant to expose the dialectal variants of speakers
- SX: phonetically compact sentences, which are designed to provide a good coverage of pairs of phones
- SI: phonetically diverse sentences, which are selected from the Brown Corpus and the Playwrights Dialog in order to add diversity in sentence types and phonetic contexts

There are two SA, five SX and three SI sentences for each speaker. In this thesis the corpus is used in order to build a phoneme bigram grammar which is used to calculate the probability of phoneme strings obtained by unconstrained phoneme recognition of non-native speech data.

4.6 PF_STAR BE Children Speech Corpus

The goal of the PF_STAR project [PFS] is to establish baselines for automatic speech recognition performance for children's speech in a range of European languages [DWR]. Read and spontaneous Italian, German, Swedish and British English (BE) speech was collected during the project. The following description is only for the read speech part of the British English corpus.

There are 159 subjects from 4 years of age up to 14. Each subject reads 25 digit triples, a list of 40 single words and longer sentences, which are taken from a set of 460 phonetically rich SCRIBE sentences. The SCRIBE sentences are the anglicized version of the TIMIT sentences. The SCRIBE sentences were divided into two sets of difficult degree. Children up to 8 years read only one list of ten sentences, and the older children two lists of ten sentences. The corpus comprises 14 hours of read speech in total.

The training data set of the corpus consists of 7.4 hours of speech from 86 children. It is employed in this work for training a monophone acoustic model for recognition of children speech and as native validation data for pronunciation scoring of children speech.

Chapter 5

Analysis of Human Annotations

In Section 5.1 results from the analysis of the human evaluation for the SX sentences of the ATR SLT non-native database are reported. Inter-rater correlation for the utterance level discrete ratings, the agreement between annotators with regard to the marking of mispronounced words and statistics of word markings are investigated. The confidence of the human ratings for the utterance and speaker level is examined in Section 5.2.

In Section 5.3 a word mispronunciation model is proposed. With that model it is possible to calculate the mispronunciation probability of phonemes if only information about mispronounced words (i.e. markings of mispronounced words) and the phonemic transcription of words is available.

5.1 Results of Human Evaluation

In order to analyze to what extent the raters agreed with each other regarding utterance ratings and the marking of mispronounced words, the performance measures introduced in Section 3.2 are employed.

Correlation analysis for word, utterance and speaker level. Tables 5.1, 5.2 and 5.3 show the inter-rater correlation for word markings, utterance level and speaker level ratings respectively. The speaker level ratings were obtained by averaging the ratings assigned to the 48 phonetically rich sentences available for each speaker.

The inter-rater correlation for word markings is between 0.16 and 0.52. Its average for all evaluator pairs was 0.34. This value, which is even lower than the correlation of 0.47 which was reported for an evaluation on phoneme level in a previous study (cf. Section 3.3.2), indicates that there was hardly a common basis among most of the evaluators regarding the notion of "mispronounced words". The decision to mark a word as mispronounced is obviously very subjective.

The utterance level correlation ranged between a very low value of 0.28 and an acceptable value of 0.65. The average correlation for all rater pairs was 0.49. In previous work of Franco and Neumeyer et al. an average inter-rater correlation of 0.65 was reported (cf. Section 3.4.6). The reasons for these comparably low correlations could be one or more of the following:

Table 5.1: *Word level inter-rater correlation for the ATR SLT data.*

EvalID	8	12	16	20
8	+1.00	+0.34	+0.40	+0.28
12		+1.00	+0.23	+0.16
16			+1.00	+0.36

EvalID	3	7	11	15
3	+1.00	+0.47	+0.36	+0.34
7		+1.00	+0.37	+0.35
11			+1.00	+0.31

EvalID	5	9	13	17
5	+1.00	+0.44	+0.43	+0.23
9		+1.00	+0.37	+0.21
13			+1.00	+0.20

EvalID	2	6	10
2	+1.00	+0.41	+0.52
6		+1.00	+0.42
10			+1.00

Table 5.2: *Utterance level inter-rater correlation for the ATR SLT data.*

EvalID	8	12	16	20
8	+1.00	+0.37	+0.63	+0.64
12		+1.00	+0.40	+0.39
16			+1.00	+0.62

EvalID	3	7	11	15
3	+1.00	+0.55	+0.48	+0.44
7		+1.00	+0.47	+0.39
11			+1.00	+0.41

EvalID	5	9	13	17
5	+1.00	+0.53	+0.45	+0.39
9		+1.00	+0.39	+0.40
13			+1.00	+0.28

EvalID	2	6	10
2	+1.00	+0.61	+0.65
6		+1.00	+0.60
10			+1.00

1. The raters were asked to assign one rating for the two different aspects pronunciation and fluency. The subjective importance of these aspects may be different for each evaluator.

2. Disagreement between evaluators concerning the term "mispronounced". Although the evaluation instructions are fairly clear ("consider only phonetic errors and ignore lexical stress"), some raters reported after evaluation that they were unsure, whether to mark only words which are not intelligible or all words which are pronounced phonetically incorrect.

3. Different usage of the grading scale despite initial calibration with a set of 22 utterances balanced with respect to the pronunciation quality of the corresponding speakers. For example, one rater (ID 16) reported, he changed his rating behavior during evaluation. Furthermore, each evaluator's experience in teaching English may have influenced the grading behavior.

4. There was at least one evaluator (ID 12) who did not diligently evaluate all utterances, since he was found out skipping at least some of the utterances, i.e. he assigned any rating without listening to the utterance in order to finish his work quickly. This finding could also apply to some of the other evaluators with low correlation values, but is rather unlikely, since raters' performance was monitored.

Table 5.3: *Speaker level inter-rater correlation for the ATR SLT data.*

EvalID	8	12	16	20
8	+1.00	+0.85	+0.94	+0.97
12		+1.00	+0.88	+0.85
16			+1.00	+0.94

EvalID	3	7	11	15
3	+1.00	+0.94	+0.85	+0.96
7		+1.00	+0.90	+0.90
11			+1.00	+0.84

EvalID	5	9	13	17
5	+1.00	+0.93	+0.95	+0.95
9		+1.00	+0.88	+0.86
13			+1.00	+0.89

EvalID	2	6	10
2	+1.00	+0.94	+0.91
6		+1.00	+0.94
10			+1.00

A further aspect for reasons (1) and (3) is, that the evaluators were instructed to consider aspects like "strong non-native accent", "mispronounced words", "long between-word pauses", "stuttering", etc. as guideline, but it is unknown which characteristic of non-native speech influences a human evaluator to what extent in his/her rating decision.

Nevertheless, there were six evaluator pairs with a correlation greater or equal to 0.6, which indicates that the evaluation procedure is of acceptable quality compared to results reported in literature. There was a certain extent of agreement among raters about the notion of pronunciation and the grading scale. An analysis of evaluator's subjectiveness in Section 6.2.4 will support this hypothesis.

The correlation between the mean rating of all but one evaluator and the remaining evaluator for a speaker or an utterance is referred to as open correlation (cf. Section 3.2). This performance measure can be calculated for each rater. The utterance level inter-rater open correlation is between 0.45 and 0.70, the speaker level open correlation between 0.88 and 0.98. The average value for the open correlation is 0.60 on the utterance level and 0.94 on the speaker level. These values may be taken as a reference for the accuracy of an automatic pronunciation scoring system. Such a system can be considered to work reliably if its scores have a correlation to averaged-ratings as high as the open correlation.

Relationship between utterance ratings and word markings. The evaluators marked

Table 5.4: *Correlation between the number of marked words and the discrete sentence level rating, and strictness for each evaluator. Most evaluators are strongly influenced by mispronounced words in their rating decision (ATR SLT data).*

Evaluator ID	2	3	5	6	7	8	9
Correlation	+0.77	+0.65	+0.64	+0.72	+0.65	+0.60	+0.45
Strictness	0.14	0.10	0.05	0.12	0.18	0.04	0.04
Evaluator ID	10	11	13	15	16	17	20
Correlation	+0.71	+0.83	+0.36	+0.50	+0.63	+0.86	+0.85
Strictness	0.06	0.09	0.04	0.17	0.09	0.29	0.28

Table 5.5: *Most frequently mispronounced words ranked by relative and absolute frequency in the ATR SLT corpus. The values in the third columns is the absolute number of how often the word was marked as mispronounced by any of the evaluators for any speaker. The values in the second columns are obtained when dividing the absolute frequency by the frequency of each word in the utterance transcriptions and the number of speakers.*

Word	relative	absolute
EXTRA	1.00	96
EXPOSURE	1.00	96
EXAM	1.00	96
BOX	1.00	96
MIRAGE	0.91	87
CENTRIFUGE	0.85	82
BUGLE	0.85	82
FRANTICALLY	0.84	81
OASIS	0.77	74
PURCHASE	0.75	72

Word	relative	absolute
THE	0.24	562
WERE	0.47	179
OIL	0.58	111
A	0.12	108
EXTRA	1.00	96
EXPOSURE	1.00	96
EXAM	1.00	96
BOX	1.00	96
MIRAGE	0.91	87
CENTRIFUGE	0.85	82

mispronounced words before they made their decision for the sentence level ratings. Consequently, the absolute number of words marked may strongly influence a human rater. The correlation between this number and the sentence level rating was calculated and is shown in Table 5.4 separately for each evaluator. The average correlation is 0.63, supporting the initial suggestion. Furthermore the higher the strictness (Eq. 3.1) of a rater is, the higher word-markings to sentence-rating correlation seems to be, as there is a correlation of 0.73 between these two values.

Mispronounced words. How many words an annotator marked as mispronounced depends on his strictness (Equation 3.1). While some annotators will mark a word if it is unintelligible, others will mark a word if just one phoneme was pronounced incorrectly. Each evaluator's strictness was calculated by dividing the number of marked words by the total number of words in the presented utterances, which is 24 speakers times 395 words equal to 9,480 words. The comparison of strictness values of two different rater groups is not reasonable, since the raters of each group evaluated utterances of different speakers with presumably different pronunciation skills. Consequently, the number of mispronounced words varies.

The ten most often marked words are shown in Table 5.5. The complete mispronunciation index of words as well as the ratio of words which are marked as mispronounced for each non-native speaker can be found in Appendix A. From the listing it is apparent that words which consist of many phonemes are more likely to be marked as short words.

Reference ratings. In order to obtain more robust pronunciation labels, which reflect the true pronunciation rating, the annotations of several human evaluators have to be combined effectively. For the sentence level ratings the combination can easily be carried out by averaging the ratings of all evaluators for the same utterance, since they consist of integer values on a linear scale from 1 to 5 and it may be assumed, that the human measurements are distributed normally.

Table 5.6: *Each evaluators strictness and inter-rater open correlation for word markings, utterance level ratings, and speaker level ratings obtained by averaging the utterance level ratings (ATR SLT data).*

EvalID	Strictness	Word level	Utter level	Speaker level
2	0.14	+0.55	+0.70	+0.94
3	0.10	+0.51	+0.61	+0.96
5	0.05	+0.45	+0.60	+0.98
6	0.06	+0.47	+0.67	+0.96
7	0.09	+0.53	+0.59	+0.95
8	0.04	+0.44	+0.70	+0.97
9	0.04	+0.42	+0.57	+0.91
10	0.12	+0.57	+0.70	+0.94
11	0.18	+0.45	+0.56	+0.88
12	0.02	+0.28	+0.45	+0.88
13	0.04	+0.40	+0.46	+0.93
15	0.17	+0.43	+0.50	+0.93
16	0.09	+0.46	+0.70	+0.95
17	0.29	+0.27	+0.44	+0.92
20	0.28	+0.37	+0.70	+0.96
Avg.	0.11	+0.44	+0.60	+0.94

The final speaker level ratings are obtained by averaging the utterance level ratings available for one speaker. In Section 5.2 a statistic of the confidence interval of the mean utterance ratings is computed. The ratings of the unreliable evaluator with ID 12 are not taken into account.

Reference markings. The reference markings of mispronounced words for a word level scoring method have to be determined. Since the strictness of many raters seem to be rather low, the markings of each evaluator group were unified, i.e. a word of any utterance of any non-native speaker which was marked by at least one evaluator is considered as mispronounced. The annotations of the unreliable evaluator with ID 12 are not taken into account.

Evaluator specific analysis. Table 5.6 summarizes each rater's strictness and each rater's open correlation on word level, utterance level and speaker level. The higher the level of annotation, the higher the inter-rater correlation becomes. This may be not astonishing, since at a higher level a rater's decision is based on more information than on a lower level, which makes his assessment more reliable.

5.2 Confidence of Human Ratings

In this section the confidence of the mean human ratings μ at the utterance level and at the speaker level is analyzed. Since there are at maximum four evaluators for each utterance, the reliability of the average utterance level ratings may be questioned. To measure this reliability,

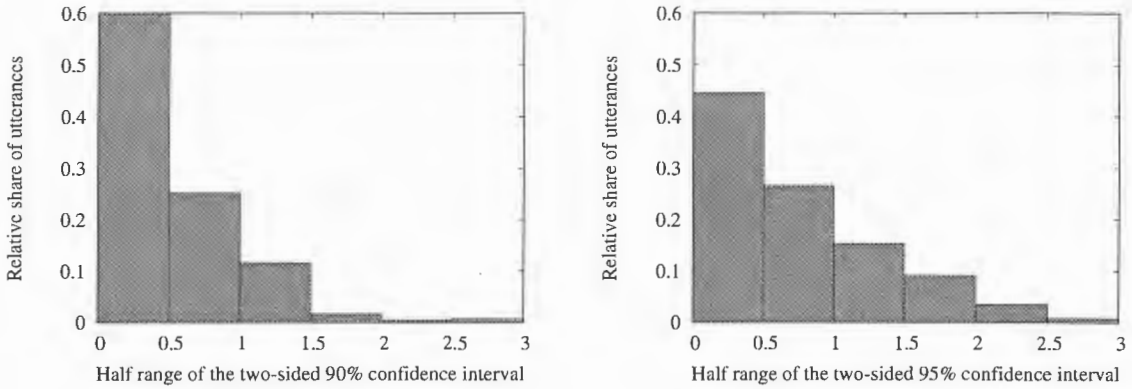


Figure 5.1: *Histogram distribution of the half ranges of the two-sided 90% (left) and 95% (right) confidence intervals for the 4608 utterance level mean ratings (ATR SLT data).*

a statistic of the confidence interval of all utterance level ratings is employed. The calculation of the confidence interval for the mean is explained in Section 2.3.4. Prerequisites were, that the scatter of the basic population is unknown and its values are distributed normal. For human ratings we may assume a normal distribution. A test for normal distribution (e.g. Chi-square χ^2 test) cannot be carried out, because the sample size is too small.

A two-sided confidence interval $[\mu - a, \mu + a]$ can be calculated for each utterance. Figure 5.1 shows the histogram of the values of $|a|$, which is the half range of the confidence interval.

Since the average half range of the 90% confidence intervals is 0.52, the simple averaging of utterance level ratings seems to be very unreliable. At the speaker level, the situation is much better as Figure 5.2 illustrates. The average half range of the 99% confidence intervals is only 0.05. Consequently, the speaker level average ratings are fairly reliable.

To deal with the unsatisfying result for the utterance level rating confidence, the jackknife resampling technique can be used to obtain more satisfactory estimates for the mean ratings (cf. Section 2.3.3). Instead of the traditional scatter of utterance ratings, the square root of

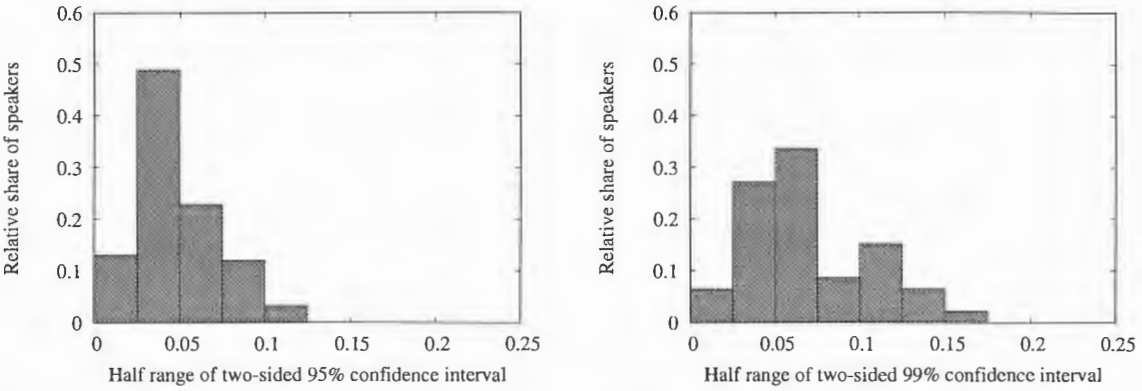


Figure 5.2: *Histogram distribution of the half range of the two-sided 95% (left) and 99% (right) confidence intervals for the 96 speaker level mean ratings (ATR SLT data).*

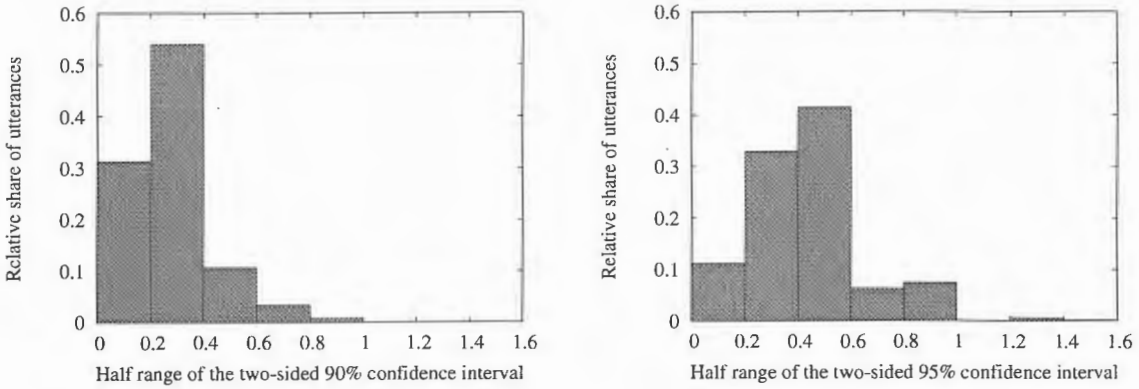


Figure 5.3: *Histogram distribution of the half range of the two-sided 90% (left) and 95% (right) confidence intervals for the 96 utterance level jackknife mean ratings (ATR SLT data).*

the jackknife variance estimate is used for confidence interval computation. The statistic of the corresponding estimates of the confidence interval is shown in Figure 5.3. From the figure it is clear, that the confidence intervals are much smaller for the jackknife mean than the traditional mean. The average half range of the 90% confidence intervals is 0.28, which is nearly half than for the traditional mean ratings. Consequently, resampling is used to increase the confidence of ratings. In following chapters, the jackknife mean ratings will be used for the utterance level and the traditional mean ratings on the speaker level.

5.3 Word Mispronunciation Model

The relative frequencies of mispronunciation markings for any speaker by any evaluator was determined in Section 5.1. These relative frequencies can be regarded as the mispronunciation probabilities for each word $\boldsymbol{w}$, which will be denoted as $P^{(mis)}(\boldsymbol{w})$. Starting from these probabilities a word mispronunciation model can be devised. This model should reflect the true circumstances of mispronunciation events and their detection by the human evaluator.

Let $Q = (q_1, \ldots, q_M)$ be the English phoneme set. Consider the event $E^{(mis)}_{p_i=q_j}$ that a phoneme p_i is mispronounced. This event will obviously have a relation to the event $E^{(mark)}_{\boldsymbol{w}}$ that word $\boldsymbol{w}$, which consists of a sequence of N phonemes, i.e. $\boldsymbol{w} = (p_1, p_2, \ldots, p_N)$, will be marked by an evaluator. In order to find out, which kind of relationship between these two events exists, two probabilistic mispronunciation models are proposed deductively and they are verified by comparing re-estimated with the already known word mispronunciation probabilities $P^{(mis)}(\boldsymbol{w})$.

Model A. A simple idea is to assume, that the probability $P^{(mis)}(\boldsymbol{w})$ and the detection of this event is related to the arithmetic or geometric mean of the probabilities $P^{(mis)}(p_i) = P(E^{(mis)}_{p_i=q_j})$, that phonemes p_i in word $\boldsymbol{w}$ are mispronounced:

$$P_a^{(mis)}(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} P^{(mis)}(p_i) \tag{5.1}$$

$$P_g^{(mis)}(\boldsymbol{w}) = \left[\prod_{i=1}^{N} P^{(mis)}(p_i)\right]^{\frac{1}{N}} \tag{5.2}$$

By taking the logarithm on both sides of Equation 5.2, the product is transformed into a sum, thus obtaining a linear relationship as in Equation 5.1. Consequently, in the following the solution will only be explained for the case of the arithmetic mean.

The phoneme mispronunciation probabilities $P^{(mis)}(p_i)$ are unknown. Since the probabilities $P^{(mis)}(\boldsymbol{w})$ can be estimated from a training sample, Equation 5.1 can be set up for each word $\boldsymbol{w}$. By taking all these single equations together an overdetermined system of linear equations for the probabilities $P^{(mis)}(q_i)$ is obtained.

Let the system of linear equations be denoted as $\boldsymbol{Ax} = \boldsymbol{y}$. Each row of matrix $\boldsymbol{A}$ contains the relative phoneme frequencies of the words $\boldsymbol{w}$, the phoneme mispronunciation probabilities $P^{(mis)}(q_i)$ are the components of vector $\boldsymbol{x}$ and the word mispronunciation probabilities $P^{(mis)}(\boldsymbol{w})$ are the components of vector $\boldsymbol{y}$. Applying the matrix $\boldsymbol{A}^T$ to both sides leads to the system of normal Equations (5.3). This system can be solved with Cholesky's method as in the case of linear regression (cf. Section 2.3.5).

$$\boldsymbol{A}^T\boldsymbol{Ax} = \boldsymbol{A}^T\boldsymbol{y} \tag{5.3}$$

In order to see whether model A applies to real data, the phoneme mispronunciation probabilities obtained were used to reconstruct the word mispronunciation via Equation 5.1 or 5.2, respectively, in order to compare them to the already known values. This procedure is carried out for each of the five major non-native speaker groups of the non-native database. The correlations between reconstructed and originally determined probabilities are shown in Table 5.7.

Table 5.7: *Correlations $C_a^{(word)}$ and $C_g^{(word)}$ between probabilities reconstructed with model A and original word mispronunciation probabilities, and correlations $C_a^{(phon)}$ and $C_g^{(phon)}$ between phoneme recognition accuracy and estimated phoneme mispronunciation probabilities (ATR SLT data).*

Speaker Group	Non-Native	German	French	Indon.	Chin.	Japan.
Arithmetic $C_a^{(word)}$	0.46	0.40	0.45	0.48	0.54	0.48
Arithmetic $C_a^{(phon)}$	-0.29	-0.16	-0.31	-0.41	-0.24	-0.28
Geometric $C_g^{(word)}$	0.40	0.34	0.34	0.40	0.44	0.36
Geometric $C_g^{(phon)}$	-0.31	-0.14	-0.28	-0.42	-0.31	-0.31

The correlation between original and reconstructed word mispronunciation probabilities as well as correlation between estimated phoneme mispronunciation and phoneme recognition accuracy is lower than 0.50 except one value, which suggests, that Model A does not match with real data.

Model B. A reasonable alternative is to think of the process of phoneme mispronunciation and its detection by a human evaluator as a Markov chain. To keep the model simple, we assume as a first approximation that the detection of mispronounced phonemes is perfect, i.e. if there is one mispronounced phoneme in a word, it will always be marked by the evaluator. Figure 5.4 depicts the random process.

p_i^c and p_i^f are the probabilities that phoneme p_i is pronounced correctly or mispronounced respectively. Their sum $p_i^c + p_i^f$ must be 1.0 in order to comply with the laws of probability theory. Based on this Markov chain, the mispronunciation probability of a word $P^{(mis)}(\boldsymbol{w})$ can be defined as:

$$P^{(mis)}(\boldsymbol{w}) = 1.0 - p_1^c \cdots p_N^c = p_1^f + p_1^c p_2^f + p_1^c p_2^c p_2^f + \cdots + p_1^c \cdots p_{N-1}^c p_N^f \tag{5.4}$$

Equation (5.4) can be rewritten as Equation (5.5) by adding $p_1^c \cdots p_N^c$, subtracting $P^{(mis)}(\boldsymbol{w})$ and taking the logarithm on both sides. As before a system of linear equations can be set up for each word with mispronunciation probability not equal to 1.0.

$$\log\left[1.0 - P^{(mis)}(\boldsymbol{w})\right] = \sum_{i=1}^{N} \log p_i^c = \sum_{j=1}^{M} n_j \log q_j^c = \sum_{j=1}^{M} n_j x_j \tag{5.5}$$

In Equation (5.5), n_j denotes the absolute number of occurrences of the j-th phoneme $q_j \in Q$ in word $\boldsymbol{w}$ and $p_i^c = q_j^c$ the probability that phoneme q_j is pronounced correctly. M is the cardinality of the phoneme set Q. The equations can be combined so that matrix $\boldsymbol{A}$ contains the occurrence frequencies n_j, the components x_j of vector $\boldsymbol{x}$ the probabilities $\log q_j^c$ and vector $\boldsymbol{y}$ the probabilities of the left sides of Equation (5.5). Matrix $\boldsymbol{A}$ has as many rows as there are words and as many columns as there are phonemes. The result is a system of linear equations as in Equation (5.3). After solving the system, the phoneme mispronunciation probabilities are calculated via $q_j^f = 1.0 - \exp(x_j)$.

The higher correlation values in Table 5.8, ranging from 0.54 to 0.69 for the re-estimated word mispronunciation probabilities and -0.32 and -0.59 for the estimated phoneme mispronunciation probabilities, show that model B is more suitable than model A to describe mispronunciation and detection events. Nevertheless it must be mentioned, that the value of the correlation $C^{(phon)}$ has a rather low confidence, since it is based on only $M = 41$ values: approximately ± 0.25. The correlation $C^{(word)}$ is more reliable with a confidence interval of about ± 0.07. The error level was set to 5%.

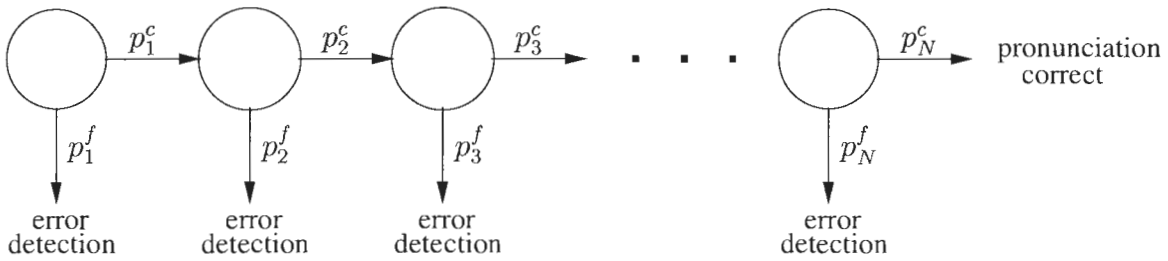


Figure 5.4: *Mispronunciation and detection process based on a Markov chain.*

Table 5.8: *Correlation $C^{(word)}$ between probabilities reconstructed with model B and original word mispronunciation probabilities, and correlation $C^{(phon)}$ between English phoneme recognition accuracy and re-estimated phoneme mispronunciation probabilities (ATR SLT data).*

Speaker Group	Non-Native	German	French	Indon.	Chin.	Japan.
Correlation $C^{(word)}$	0.69	0.54	0.64	0.67	0.61	0.68
Correlation $C^{(phon)}$	-0.59	-0.32	-0.58	-0.56	-0.46	-0.49

Further refinements of model B are possible in theory. For example, in model B it is assumed, that the evaluators mark any word with only one phoneme mispronounced, i.e. the detection rate is 100%. Unless the evaluator is very strict and diligent, perfect detection is unlikely to be true in reality.

Table 5.9 shows the list of phonemes with the highest mispronunciation probability for each non-native accent group. The probabilities are estimated with model B. The phonemes /th/ and /er/ are among the phonemes with the highest mispronunciation probability of all accent groups. Among the worst ten candidates of four groups are the phonemes /dh/, /sh/, /aw/ and /ax/.

Table 5.9: *List of phoneme mispronunciation probabilities (ATR SLT data).*

French		German		Indonesian		Japanese		Chinese	
er	0.46	**th**	0.29	sh	0.53	**th**	0.47	**th**	0.50
th	0.36	y	0.25	**er**	0.45	**er**	0.44	sh	0.45
ax	0.29	**er**	0.25	**th**	0.36	sh	0.35	dh	0.36
dh	0.27	sh	0.25	ax	0.31	axr	0.29	**er**	0.32
aw	0.24	aw	0.20	ch	0.30	r	0.26	l	0.29
ah	0.24	g	0.17	y	0.28	l	0.26	aw	0.28
jh	0.22	ax	0.13	uw	0.25	ax	0.25	w	0.25
ao	0.22	jh	0.13	dh	0.25	dh	0.22	v	0.25
ow	0.20	uw	0.13	jh	0.25	g	0.22	r	0.24
uh	0.20	ow	0.10	g	0.24	aw	0.20	ng	0.24

Chapter 6

Features for Pronunciation Scoring

In this chapter features describing the pronunciation quality of utterances and words are defined. Besides employing features already known from literature (cf. Chapter 3), a large number of additional features are considered. The usefulness of utterance and speaker level features is investigated by analyzing the correlation coefficient between human ratings and scores for the ATR SLT data. Actual evaluation and combination of utterance and word level features is carried out in Chapter 8 for both the ATR SLT and FAU LME data.

6.1 Experimental Setup for Feature Extraction

This section describes the setup of experiments for investigation of correlations of human ratings with automatically extractable pronunciation scores. Figure 6.1 gives an overview of data flow (black arrows), models (round boxes with white arrows) and processing steps (rectangle boxes). The setup includes a speech recognizer for unconstrained word and unconstrained phoneme recognition. The recognizer is also necessary for obtaining the forced-alignment of phoneme level acoustic models for an utterance with the word level utterance transcription and a pronunciation dictionary. Furthermore, a native phoneme duration statistic for extraction of duration-related features and a native phoneme language model (LM) for extraction of phoneme sequence probabilities are required. The software toolkit HTK V3.2 is employed for all recognition and alignment experiments and for phoneme sequence probability computation.

Native acoustic model (AM). The acoustic model is built with native English speech data from the WSJ corpus (cf. Section 4.3). 39 MFCC features are extracted every 10 ms: 12 cepstral coefficients and normalized log-energy with first and second derivations. The derivations are calculated as the regression line over five frames. The frame duration is 20 ms. Cepstral mean subtraction (CMS) is applied. A monophone AM consisting of 44 3-state phoneme HMMs and one combined 3-state silence (sil) / 1-state short pause (sp) HMM is trained from scratch: the parameters of single-mixture Gaussian distributions with diagonal covariance matrix for each HMM state are initialized with the mean and variance of the MFCC feature vector sequence of acoustic segments of equal length. There is one segment for each HMM state of the state sequence which corresponds to the concatenation of each word's phoneme HMM sequence of the

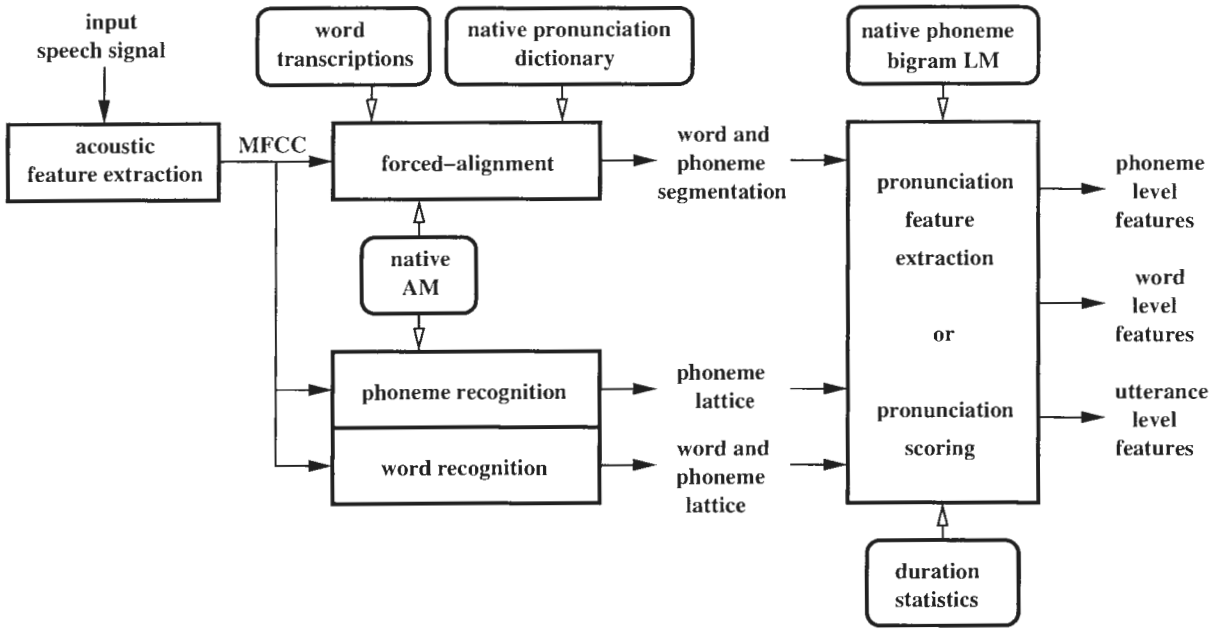


Figure 6.1: *Experimental setup for pronunciation feature extraction at different levels.*

utterance. The number of mixtures was increased successively during training one by one until 16 mixtures were reached. Models were retrained for four iterations after each mixture increment. The performance of the acoustic model just described was evaluated for native English speech with the 5000 word Hub2 evaluation test set from the WSJ corpus. The recognition accuracy with a bigram language model was 80.8%.

Native pronunciation dictionary. The pronunciation dictionary for experiments with the ATR SLT non-native database contains 429 phoneme transcriptions for 290 words including native pronunciation variants. The pronunciation dictionary is used for computation of the forced-alignment of the TIMIT SX sentences of the database and for doing unconstrained word recognition, i.e. decoding without a statistic language model.

Native phoneme language model (LM). In order to compute probabilities of phoneme sequences, a bigram phoneme language model (LM) is employed. It serves the purpose of calculating the probability of phoneme sequences, which are obtained by unconstrained phoneme recognition. The LM is estimated from all SX, SI and SA sentences in the TIMIT corpus.

Native phoneme duration statistic. In order to calculate the expected duration of words and phoneme duration scores, the distribution of phoneme durations has to be modeled. To estimate a phoneme duration histogram or the parameters of an analytic probability density function, a large number of samples is required. The phoneme duration statistics employed in the following experiments are estimated from the SA, SI and SX sentences of the TIMIT database, since the sentences contain relatively many samples also of rare phonemes like /zh/. Phoneme durations are extracted from the utterance's forced-alignment obtained with the native AM. More accurate phoneme durations could have been extracted from the phonetic transcriptions of the TIMIT corpus which are made by phoneticians. However using them leads to a mismatch condition.

Manually derived phoneme durations are different from those based on the forced-alignment with an HMM-based speech recognizer and for pronunciation scoring phoneme durations must inevitably be determined automatically in real time.

Word/phoneme recognition. For speech recognition a statistical language model, e.g. n-gram with $n \geq 2$, is usually employed to exclude unlikely word sequences and to reduce the search space. However in pronunciation scoring the application is different in the sense, that the recognition performance should reflect the quality of a non-native speaker's pronunciation. A statistical language model with strong constraints would introduce the undesired effect, that the recognition is still rather high, despite the pronunciation skill of a non-native speaker being rather low. Consequently, a word-loop and a phoneme-loop recognition network are employed for phoneme and word recognition, i.e. any phoneme (word) may follow any phoneme (word).

6.2 Features and Scores

Section 6.2.1 introduces a set of base features. Some features of this set can be used without modifications as phoneme, word or sentence scores. The features and scores are intended to measure segmental and temporal qualities of non-native speech. Additionally word and utterance level scores, which are combinations of certain base features, are defined in Sections 6.2.2 and 6.2.3. The performance of utterance and speaker level features is evaluated by analyzing the correlation between human ratings and scores.

The base features are extracted with the setup described in Section 6.1. For every utterance unconstrained phoneme and word recognition is carried out. Furthermore, the forced-alignment is computed. From the recognition output and the segmentation into phonemes and words the features describing the pronunciation quality of phonemes, words and sentences can be calculated.

6.2.1 Base Features

Most of the base features are based on information about phoneme segments. There are two kinds of **acoustic scores**: **likelihood** and **posterior** scores. The higher the acoustic score, the better the match of the acoustic observation with the acoustic model. The higher the acoustic match, the more similar is the non-native speaker's pronunciation to native speech, i.e. the acoustic score is a possible metric of segmental aspects of pronunciation.

Two more features are based on phoneme durations. Besides the **actual duration** t of phonemes in the current word $\boldsymbol{w}$ or sentence $\boldsymbol{u}$, their **expected duration** $t^{(exp)}$ is also considered. The expected duration of a phoneme segment is estimated as the mean duration of the corresponding phoneme in a native speech corpus. Since the duration of a certain phoneme is distributed log-normal, the **duration probability** can also be used as a feature. These duration-related features can measure temporal aspects of pronunciation. Further temporal aspects can be captured by the **number and duration of pauses** between words in an utterance.

Let $\boldsymbol{q} = (q_1, \ldots, q_N)$ be the phoneme sequence obtained by phoneme or word recognition. Given a language model (LM) trained on native phoneme sequences, i.e. phoneme sequences

which correspond to standard English words, sentences and texts, a **phoneme sequence probability** $P(\boldsymbol{q}|\mathrm{LM})$ can be calculated. If the pronunciation characteristics of a non-native speaker are near to native speech, phoneme recognition will work well. Consequently the phoneme sequence probability will be rather high. Opposite effects will occur for speech of low proficiency speakers. The same argument applies to the **recognition performance**, which will also be employed as base feature.

6.2.2 Utterance Features

Different kinds of utterance scores are investigated. Only the ATR SLT data is employed for the investigations, since utterance level ratings are not available for the FAU LME data. Each feature defined has an identifier consisting of a capital letter followed by a number. There is a different capital letter for each feature kind, which are related to likelihood (L), expected word duration (E), likelihood ratio (K), rate of speech (R), between-word pauses (P), duration (D), recognition performance (X), phoneme sequence likelihood (M) and others (Y).

Likelihood. There are many possibilities to define a likelihood score for a whole utterance. Frame level scores can be combined to phoneme level scores, phoneme level scores to word level scores, and word level scores to utterance scores. Here the direct combination of phoneme level scores to sentence scores is considered. The score $L(\boldsymbol{x}_i)$ for a phoneme segment $\boldsymbol{x}_i$ is calculated as the logarithmic probability of the best path through the HMM of the segment's phoneme p_i.

Since the likelihood scores a speech recognizer calculates for each unit are logarithmic probabilities, the simplest way to obtain an utterance score is to sum up the scores of all phoneme segments. However this score value does not only depend on the acoustic match, but also on the utterance duration and the number of phoneme segments. Consequently, the score has to be normalized. There are several possibilities for normalization. Dividing the sum of phoneme likelihoods by the utterance duration or number of phoneme segments is most obvious. However, this way of score normalization is not optimal in the sense, that the score's correlation with utterance level ratings becomes maximum. In the following, experimental results will reveal a good combination of score normalization factors.

Table 6.1 and Figure 6.2 illustrate the various elements of an utterance, which are involved in utterance score calculation. An utterance $\boldsymbol{u}$ consists of phoneme segments $\boldsymbol{x}_i$, short pauses (sp) and silence segments (sil) segments. The duration of inter-word short pauses (sp) may be zero. Successive phoneme segments $\boldsymbol{y}_i^j$ without any pauses in between make up word segments $\boldsymbol{w}_j$. The total sentence duration $t_{\boldsymbol{u}}$ is defined as the sum of the durations t_i of all phoneme segments plus the sum of interleaving sp/sil segments. sp/sil segments before the first and after the last phoneme are removed. The *rate of speech* in terms of phonemes $R^{(phon)}$ is calculated as the number of phonemes n divided by the total utterance duration $t_{\boldsymbol{u}}$. The calculation of $R^{(word)}$ is similar (cf. Table 6.1). The unit to measure durations may be chosen arbitrarily, although it has to be consistent for all score definitions. A preferable unit is the number of frames.

Table 6.2 gives the definition of several likelihood scores. The correlation coefficients for these scores and the discrete human ratings are shown in Table 6.3.

Table 6.1: *Definition of variables and symbols.*

Name of entity	Symbol	Definition
Utterance (as phoneme sequence)	$\boldsymbol{u}$	$\boldsymbol{u} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
Utterance (as word sequence)		$\boldsymbol{u} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)$
Word (as phoneme sequence)	$\boldsymbol{w}$	$\boldsymbol{w}_j = (\boldsymbol{y}_1^j, \ldots, \boldsymbol{y}_{n_j}^j)$
Phoneme segment (in sentence)	$\boldsymbol{x}$	$\boldsymbol{x}_i = (x_1^i, \ldots, x_{t_i}^i)$
Phoneme segment (in word)	$\boldsymbol{y}$	$\boldsymbol{y}_i = (y_1^i, \ldots, y_{t_i}^i)$
Segment frame	x, y	Acoustic observation
Duration	t_i	Duration (of segment $\boldsymbol{x}_i$)
	d_j	Duration (of word $\boldsymbol{w}_j$)
# Segments	n	# Segments (in utterance $\boldsymbol{u}$)
	n_j	# Segments (in word $\boldsymbol{w}_j$)
# Words	m	# Words (Sentence)
Sentence duration	$t_{\boldsymbol{u}}$	Sentence duration (leading and trailing sp/sil removed)
Rate of speech	$R^{(phon)}$	# Phonemes (n) / Sentence duration ($t_{\boldsymbol{u}}$)
	$R^{(word)}$	# Words (m) / Sentence duration ($t_{\boldsymbol{u}}$)

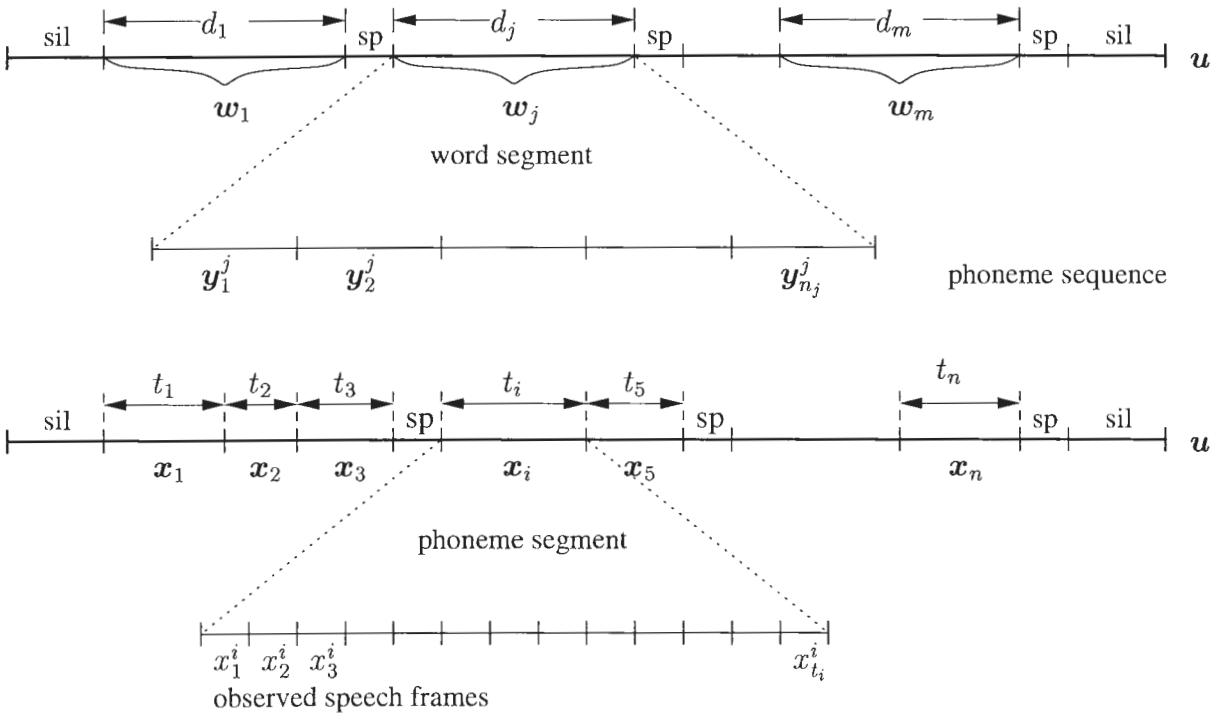


Figure 6.2: *Illustration of various utterance elements. The duration of inter-word short pauses (sp) may be zero, or there may be silence segments (sil), i.e. long pauses between words.*

Table 6.2: *Definition of likelihood scores.*

Score-ID	Name	Symbol	Definition
-	Segment-likelihood	$L(\boldsymbol{x})$	$\log P(\boldsymbol{x}\|\text{HMM}\lambda)$
-	Word-likelihood	$L^{(word)}(\boldsymbol{w}_j)$	$\sum_{i=1}^{n_j} L(\boldsymbol{y}_i^j)$
L3	-	$L^{(sent)}(\boldsymbol{u})$	$\sum_{i=1}^{n} L(\boldsymbol{x}_i)$
L4	Global sentence-lh.	$S'^{(sent)}_{global}$	$\frac{\sum_{i=1}^{n} L(\boldsymbol{x}_i)}{\sum_{i=1}^{n} t_i * R^{(phon)}}$
L5	-	-	$\frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{x}_i)$
L6	-	-	$\frac{1}{m}\sum_{i=1}^{n} L(\boldsymbol{x}_i)$
L7	Local sentence-lh.	$S'^{(sent)}_{local}$	$\frac{1}{n}\sum_{i=1}^{n} \frac{L(\boldsymbol{x}_i)}{t_i * R^{(phon)}}$
L8	-	-	$\frac{1}{m}\sum_{j=1}^{m} \frac{L^{(word)}(\boldsymbol{w}_j)}{d_j * R^{(word)}}$
L9	-	-	$\frac{1}{m}\sum_{j=1}^{m} \frac{L^{(word)}(\boldsymbol{w}_j)}{n_j}$

Table 6.3: *Correlation between sentence level human ratings and likelihood scores.*

Score-ID	L3	L4	L5	L6	L7	L8	L9
Correlation	-0.24	-0.41	-0.34	-0.28	**-0.42**	-0.37	-0.35

The highest correlation is present for the scores $S'^{(sent)}_{global}$ (L4) and $S'^{(sent)}_{local}$ (L7) with 0.41 and 0.42. These scores are not only normalized by the duration or the number of segments, but also by the rate of speech (ROS), which is defined in Table 6.1. The calculation of the rate of speech was carried out separately for each utterance. In the research of Neumeyer et al. (cf. Table 3.3, Chapter 3) this kind of normalization was not considered. The correlation of the original scores $S^{(sent)}_{global}$ and $S^{(sent)}_{local}$ with human ratings was only 0.18 and 0.29, respectively.

The correlation for the scores {L5,L8,L9} is higher than 0.33. The remaining scores {L3,L6} have a correlation lower than 0.28 and may not be of much use. Which of the five scores with correlation higher than 0.33 {L4,L5,L7,L8,L9} should finally be selected as features for utterance classification does not only depend on the degree of correlation but also on the inter-score correlation. Since the scores L3-L9 are all based on similar information, the inter-score correlation may be high. From Table 6.4 we see, that the scores of pair (L4,L7) have a correlation of 1.0, i.e. these two scores are identical measures. Using both of these two scores for classification will probably not lead to better results. Furthermore, the score-pairs {(L4,L5); (L4,L9); (L5,L7); (L5,L9); (L7,L9)} are highly correlated. Since the inter-score correlation of

Table 6.4: *Correlation between likelihood scores (ATR SLT data).*

Correlation	L5	L7	L8	L9
L4	+0.86	+1.00	+0.54	+0.84
L5		+0.85	+0.32	+0.95
L7			+0.54	+0.83
L8				+0.36

the pairs {(L4,L8); (L8,L9); (L7,L8); (L5,L8)} are relatively low, one of these two pairs may be chosen as features for pronunciation scoring. In Appendix C.3 detailed information about inter-score correlation is given.

There is an alternative and more convenient way to get rid of correlated scores automatically rather than doing manual analysis: Application of principal component analysis (PCA). This approach will be employed in Chapter 7 when score combination is considered.

Expected word duration. Given the likelihood scores $L(\boldsymbol{x}_i)$, the true t_i and expected durations $t_i^{(exp)}$ of each utterance's phoneme segments, additional features describing the pronunciation quality of an utterance can be defined. The expected duration of a phoneme (segment) is approximated by the mean duration of the phoneme in the utterances of native speakers from the TIMIT DB. In order to obtain accurate estimates for the expected duration of words, a large number of samples of each word is necessary. For the special case of the SX sentences of the ATR SLT non-native database, the estimates could be obtained from the SX sentences of the TIMIT DB. However this kind of word duration statistic may not be available in general. Consequently, the expected duration of a word $\boldsymbol{w}_j$ with the phoneme sequence $(p_1, \ldots, p_{n_j})$ is approximated by the sum of mean durations of the phonemes p_i.

Table 6.5: *Definition of features involving the expected duration of phonemes and words.*

Score-ID	Symbol	Definition	Meaning
-	$d_j^{(exp)}$	-	Expected duration of word $\boldsymbol{w}_j$
-	$t_i^{(exp)}$	-	Expected duration of segment $\boldsymbol{x}_i$
-	Δt_i	$t_i - t_i^{(exp)}$	Duration deviation
E1	$\overline{\Delta t}$	$\frac{1}{n}\sum_{i=1}^{n} \Delta t_i$	Mean duration deviation
E2	-	$\frac{1}{n}\sum_{i=1}^{n}(\Delta t_i - \overline{\Delta t})^2$	Duration deviation scatter
E3	-	$\frac{1}{m}\sum_{j=1}^{m} \frac{L^{(word)}(\boldsymbol{w}_j)}{d_j^{(exp)} * R^{(word)}}$	Word-based sentence score
E4	-	$\frac{1}{n}\sum_{i=1}^{n} \frac{L(\boldsymbol{x}_i)}{t_i^{(exp)} * R^{(phon)}}$	Phoneme-based sentence score

The mean duration of a phoneme is derived from the annotations of phonemes in the TIMIT

DB. Table 6.5 gives the definitions of the six features E1-E4. Cf. Tables 6.1 and 6.2 for already introduced definitions. The results in Table 6.6 show, that there is almost no improvement in correlation, e.g. compare E3 with L8 and E4 with L7.

Table 6.6: *Correlation between sentence level human ratings and features based on the expected duration of phoneme segments (ATR SLT data).*

Score-ID	E1	E2	E3	E4
Correlation	+0.30	+0.28	**-0.43**	-0.41

Likelihood ratio. The concept of the likelihood ratio was already introduced in Section 3.4.1. It is the logarithm of the quotient $\log \frac{P}{P'}$ of two probabilities or the difference $L - L'$ of two log-likelihoods. If the logarithm of the posterior probability $\log P(p|x_i)$ is approximated as

$$\log P(p|x_i) = \log \frac{P(x_i|p)P(p)}{\sum\limits_{q \in Q} P(x_i|q)P(q)} \approx \log \frac{P(x_i|p)}{\max\limits_{q \in Q} P(x_i|q)} \tag{6.1}$$

which is in the line of the GOP score calculation in Equation 3.4, it has the form of a likelihood ratio.

$$L^{(ratio)}(x_i|p, q^*) = \log P(x_i|p) - \log P(x_i|q^*) \qquad q^* = \operatorname*{argmax}_{q \in Q} P(x_i|q) \tag{6.2}$$

The phoneme set of the target language is denoted as Q. The variable x_i symbolizes the i-th frame of the currently considered phoneme segment $\boldsymbol{x} = (x_1, \ldots, x_t)$. The log-likelihood $\log P(x_i|p)$ can be approximated as $\frac{1}{t} \log P(\boldsymbol{x}|\lambda_p)$, where $P(\boldsymbol{x}|\lambda_p)$ is the probability of the best path through the HMM λ_p for phoneme p. This probability is obtained from the forced-alignment. The segment likelihood $\log P(\boldsymbol{x}|\lambda_{q^*})$ to approximate $\log P(x_i|q^*)$ is determined by unconstrained phoneme recognition. Since the recognized phonemes and their segment boundaries may not match the phoneme segments of the forced-alignment, the likelihood ratio has first to be calculated separately for each frame. Figure 6.3 illustrates this calculation.

The likelihood ratio score $L^{(ratio)}(\boldsymbol{x})$ for the whole phoneme segment $\boldsymbol{x}$ is calculated as the sum of the likelihood ratios $L^{(ratio)}(x_i|p, q^*)$ of all frames t of $\boldsymbol{x}$. The utterance level scores K1-K3 based on the likelihood ratio are defined in Table 6.7.

Table 6.7: *Correlation between sentence level human ratings and likelihood ratio.*

Score-ID	K1 (cf. L5)	K2 (cf. L4)	K3 (cf. L4)
Correlation	-0.48	-0.50	**-0.52**
Equation	$\frac{1}{n} \sum\limits_{i=1}^{n} L^{(ratio)}(\boldsymbol{x}_i)$	$\frac{\sum\limits_{i=1}^{n} L^{(ratio)}(\boldsymbol{x}_i)}{\sum\limits_{i=1}^{n} t_i * R^{(phon)}}$	$\frac{\sum\limits_{i=1}^{n} L^{(ratio)}(\boldsymbol{x}_i)}{\sum\limits_{i=1}^{n} t_i^{(exp)} * R^{(phon)}}$

The results of the correlation analysis show, that a sentence score based on the frame-wise likelihood ratio is a better pronunciation score than the scores L3-L9. Maximum correlation of 0.52 is present for score K3.

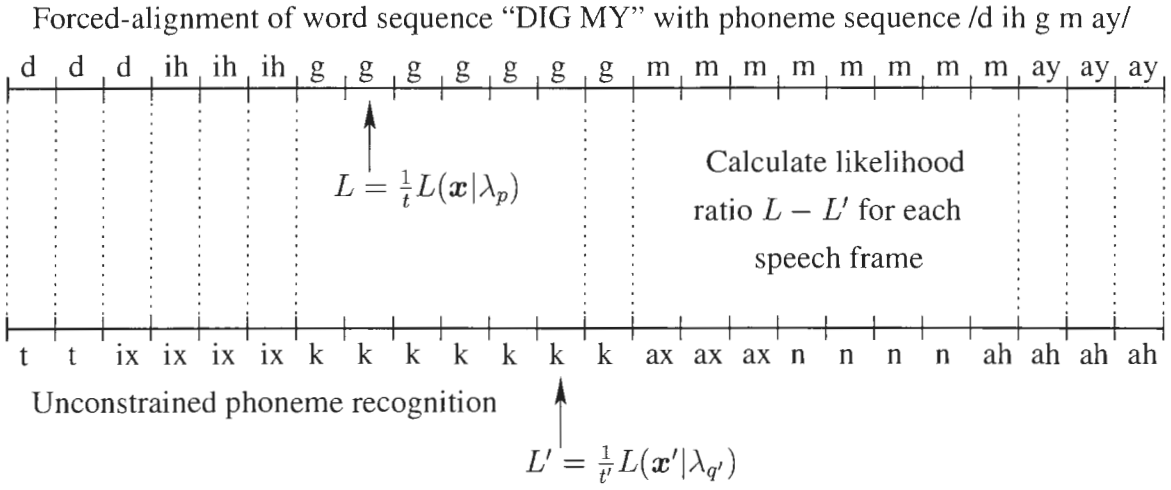


Figure 6.3: *Illustration of the likelihood ratio score.*

Speaking rate and pauses. Since evaluators were asked to consider the aspect fluency for evaluation, the features word-based (R1) and phoneme-based (R2) rate of speech as well as the total duration of between-word pauses may contain useful information. The rate of speech is defined in Table 6.1. Additionally the reciprocal features, i.e. the average word (R3) and phoneme duration (R4) and the phonation-time ratio (R5) are considered here. The total duration of between-word pauses $D^{(pause)}(\boldsymbol{u})$ is calculated as sum of all pause segment durations excluding the segments before the first word and after the last word of a sentence. Results are summarized in Table 6.8. There is a correlation of $+0.37$ for the average phoneme duration and the reciprocal rate of speech features perform better than the traditional ones. The total duration of between-word pauses (P1) is correlated with sentence level ratings by 0.33. An interesting observation is, that there is a correlation of $+0.14$ for the first between-word pause. As more of the following pause segments durations are taken into account, the correlation increases until a maximum of $+0.33$ is reached. Normalization of the pause duration by the number of pauses did not lead to an increase in correlation. The number of between-word pauses longer than 0.2 seconds (P2) itself is correlated to human ratings with 0.32.

Table 6.8: *Correlation of sentence level ratings to rate of speech and pause-related features. The bottom row gives information about the mean value of each feature for non-native speech.*

Score-ID	R1	R2	R3	R4	R5	P1	P2
Symbol	$R^{(word)}$	$R^{(phon)}$	$\frac{1}{R^{(word)}}$	$\frac{1}{R^{(phon)}}$	$\frac{\sum t_i}{t_{\boldsymbol{u}}}$	$D^{(pause)}(\boldsymbol{u})$	-
Correlation	-0.34	-0.37	+0.37	**+0.39**	-0.32	**+0.33**	+0.32
Mean Value	2.74 [Hz]	9.90 [Hz]	390 [ms]	105 [ms]	0.95	0.38 [s]	1.34

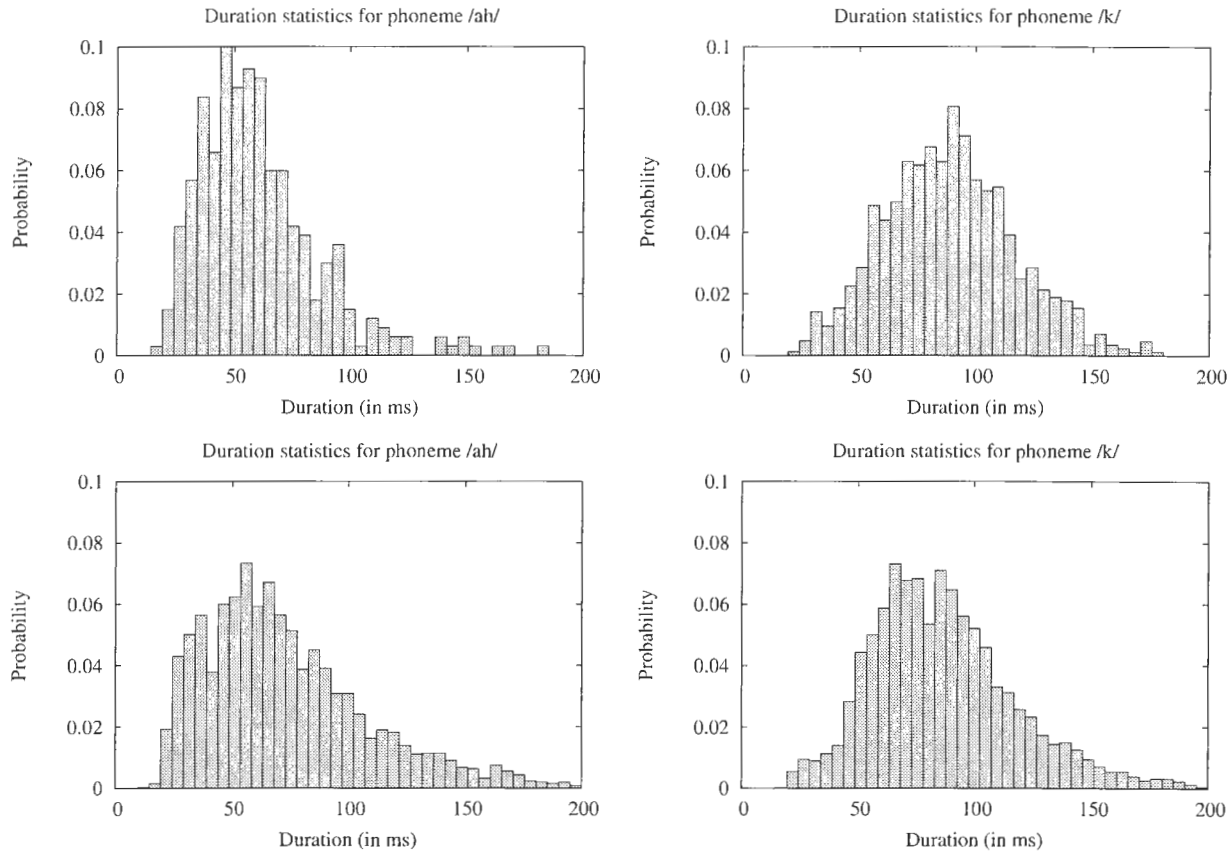


Figure 6.4: *Duration statistics for phonemes /ax/ and /k/ uttered by* **native (upper histograms)** *and* **non-native (lower histograms)** *speakers. The histograms were estimated from the forced-alignment of utterances in the ATR SLT non-native database, which contains also data of some native speakers uttering the same sentences as the non-native speakers.*

Duration scores. In every language, the durations of phonemes may be different. The typical duration of phonemes of a non-native speaker's first language may influence the duration of second-language phonemes. Additionally severe deviations can occur if there are mispronunciations or non-uniform speaking behavior. Phoneme durations are obtained from the forced-alignment. Consequently, there can be extraordinarily short or long phoneme segments, since the recognizer is forced to match reference phonemes regardless whether they are present in the speech signal or not.

As already mentioned in previous sections, phoneme durations depend on the speaking rate. Consequently, normalization (multiplication) of phoneme durations with the rate of speech is imperative. A phoneme duration statistic for native phonemes can be approximated by a histogram or a log-normal distribution. Figure 6.4 shows the difference of duration histograms of the vowel /ax/ and the plosive /k/ for native and non-native speakers.

For each phoneme segment $\boldsymbol{x}$ with phoneme label p and duration t a **phoneme duration score** $S_{dur}^{(phon)} = \log P_{dur}^{(phon)}$ can be calculated. $P_{dur}^{(phon)}$ is defined as:

Table 6.9: *Correlation of utterance level ratings to duration likelihoods (D1-D3), phoneme (X1) and word (X2) recognition performance, and phoneme sequence likelihood (M1-M3).*

Score-ID	D1	D2	X1	X2	M1	M2	M3
Correlation	-0.45	**-0.46**	**-0.45**	-0.38	-0.22	-0.28	**-0.40**

$$P_{dur}^{(phon)}(T|p,\boldsymbol{x}) = \frac{1}{T\sqrt{2\pi\sigma_p}} \exp\left[-\frac{(\log T - \nu_p)^2}{2\sigma_p}\right] \qquad T = t * R^{(phon)} \tag{6.3}$$

The parameters σ_p and ν_p of the log-normal duration probability density function can be obtained by ML estimation from a sample of durations of phoneme p. The normalization by multiplication with the rate of speech is applied as suggested in literature (cf. Section 3.4.3).

The duration scores $S_{dur}^{(phon)}$ of all phoneme segments in an utterance can be summed up to obtain the sentence level duration score (D1) as already proposed in literature (cf. Equation 3.11). A normalized duration score (D2) is obtained by dividing D1 with the total number of phoneme segments n. As Table 6.9 shows, there is a high correlation between duration scores and human ratings.

Recognition performance. For practical reasons, the recognition performance is calculated as the minimum-edit-distance between the recognized and the reference token sequence divided by the number of tokens in the longer of the two sequences (cf. Eq. 2.9). This feature has values between 0.0 and 1.0. Its extreme values are 1.0, if not even one token is correct, and 0.0, if the two sequences are identical. Table 6.9 shows the correlation of ratings with phoneme (X1) and word (X2) recognition performance. The computation of (X2) is based on the phoneme sequence corresponding to the recognized word sequence. The correlation for score (X1) is higher than for score (X2). This result can be explained easily by the fact, that there are many times more phonemes than words in a sentence on which the calculation of the recognition accuracy is based. Consequently, the phoneme-based accuracy is a more reliable measure than the word-based accuracy. The correlation for (X1) is slightly lower (-0.43), when its calculation is based on the phoneme sequence obtained by unconstrained phoneme recognition.

Phoneme sequence probability. Every natural language has a certain phonotactic structure. There are phoneme sequences which have a high probability and others which have a low probability due to a language's phoneme set, its vocabulary, its grammar and word usage. Since a non-native speaker does not pronounce all phonemes and words of a sentence correctly, the resulting phonotactic structure of his utterances will differ from native speech. Consequently, it is worth considering the logarithmic probability $\log P(\boldsymbol{q}|\mathrm{LM})$ of recognized phoneme sequences $\boldsymbol{q}$ as a feature for pronunciation quality. As phoneme language model (LM) a phoneme bigram trained on the TIMIT corpus (cf. Section 6.1) is employed here. Table 6.9 shows that the phoneme sequence likelihood (M1) divided by the number of phonemes in the current sentence (M2) has a correlation of -0.28 with the utterance level ratings. The correlation increases to -0.40 if M1 is divided by the rate of speech (M3).

Second order features. The likelihood scores L5-L9 are in fact a first order statistic of phoneme or word likelihoods. As additional features, the scatter of word likelihoods (Y3), the

Table 6.10: *Correlation of utterance ratings to minimum (Y1) and maximum word likelihood (Y2), scatter of word (Y3) and phoneme likelihoods (Y4), and minimum of the local ROS (Y5).*

Features	Y1	Y2	Y3	Y4	Y5
Correlation	**-0.26**	-0.21	+0.17	+0.20	+0.24

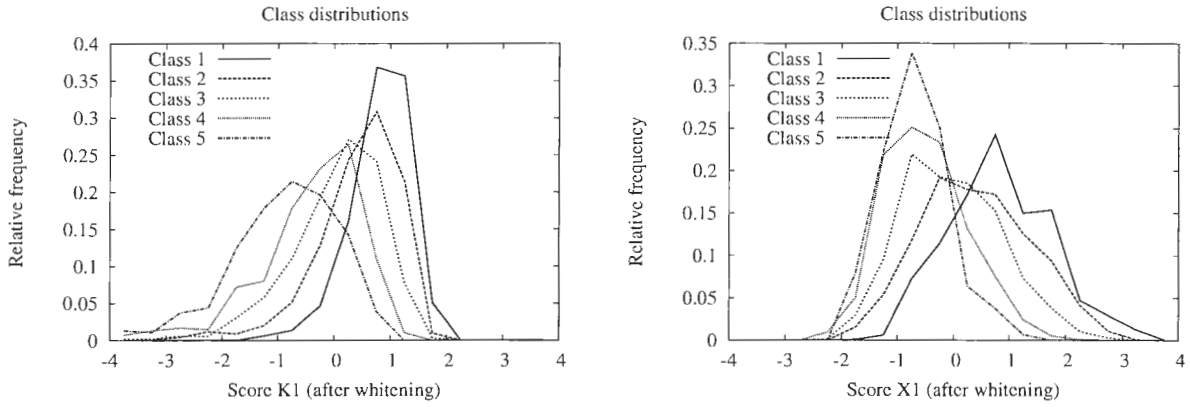


Figure 6.5: *Distribution of scores K1 (left) and X1 (right) w.r.t. each rating class (ATR SLT data).*

scatter of phoneme likelihoods (Y4) and the minimum of the local rate of speech (Y5) are used. The local rate of speech is defined as the number of phoneme segments in a word, divided by its duration. Furthermore, the minimum (Y1) and maximum likelihood (Y2) of all words in a sentence are considered, since an evaluator may especially be influenced by the pronunciation of the best or worst word segment. As Table 6.10 shows, the correlation to the human ratings for all these features is lower than 0.30. Further investigation was also done for the maximum and scatter of the local rate of speech and the minimum and maximum phoneme likelihoods. However the correlation of these features was even less than 0.20. Nevertheless, features Y1-Y5 may be useful, since they have a rather low correlation with other utterance level pronunciation scores.

Score distributions. Figure 6.5 visualizes the distribution of the two pronunciations scores, K1 and X1, which have one of the highest correlations with the human ratings among the investigated scores. The distribution for each discrete human rating class is shown. The average human ratings for each utterance were rounded in order to obtain five classes of pronunciation quality. From the figure it can be seen that the distributions of neighbored classes overlap and there is less overlap between non-neighbored classes. Furthermore it is apparent that the scores build a continuum. This means, that scores for utterances with the highest (1) pronunciation quality rating make up the left-most and with the lowest (5) rating the right-most distribution. Additionally scores for the remaining pronunciation levels lie between the two extremes in a reasonable order. Score distributions of other features can be found in the score gallery in the Appendix C.

Subjectiveness of pronunciation evaluation. In Section 5.1 the hypothesis was claimed, that the importance of the various aspects of pronunciation may weighted different by each

evaluator. To test this hypothesis the rating to score correlation for utterances was calculated separately for each evaluator (cf. Table B.3 in Appendix B). Figure 6.6 visualizes the differences graphically. The line L7 is the value of the correlation coefficient of feature L7, the line K3 the sum of features L7 and K3, the line R4 the sum of features L7, K3 and R4, a.s.o. The left graph shows, that if there is a relatively high correlation between a score and the ratings of one evaluator, there is also a relatively high correlation for the other scores. From the right graph it is apparent, that there are some differences in the relative degree of correlation depending on the evaluator. However the differences are not remarkable enough to be able to conclude that there is a strong inter-rater discordance. The difference between raters lies in the degree of correlation in general, rather than in the difference of correlations for each score. The low rating to score correlation of some evaluators suggests that for these evaluators, despite clear instructions, other aspects (e.g. prosodic) than the investigated aspects (segmental, temporal) may have been more important or that they were unsure about how and what to evaluate.

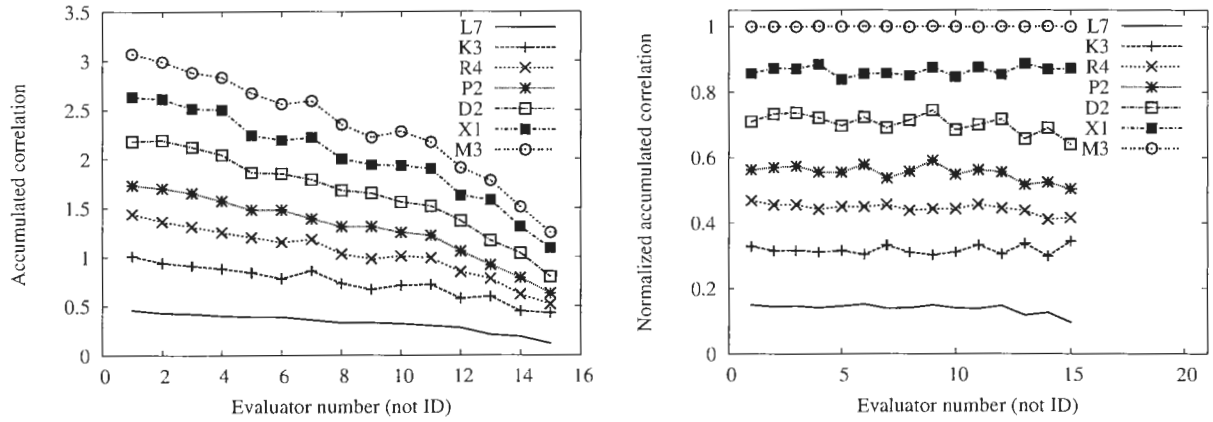


Figure 6.6: **Left:** *Accumulation of the correlation coefficients of selected features from Table B.3. Line L7 is the correlation for L7, line K3 is the sum of correlations for L7 and K3, line R4 the sum of correlations L7, K3 and R4, a.s.o.* **Right:** *Same as left graph, but normalized by dividing with the sum of correlations for all considered features (ATR SLT data).*

6.2.3 Word Features

For words there are only binary annotations: $\mathcal{X}$ for mispronounced and $\mathcal{O}$ for correctly pronounced words. Consequently, correlation analysis is not the optimal way to examine the usefulness of word level pronunciation scores. In the following word level features are only defined. The quality of the features is evaluated in Chapter 7 for both the ATR SLT and the FAU LME database.

Most utterance level features from Section 6.2.2 can be adopted for the word level. The word level scores are analogue to the scores of one-word sentences. Except for the word-based rate of speech and pause-related features, all feature kinds can also be applied to words. However, in a preliminary investigation it was found out, that a normalization of word likelihood scores by

dividing with the number of phoneme segments n or the expected word duration $d^{(exp)}$ decreases the quality of the feature. This is apparent when comparing the distributions, e.g. of features W01 (analogue to L3) and W02 (analogue to L5) in Appendix C. Consequently, the definition of features W03 and W04 deviates from other utterance level likelihood features. Table 6.11 gives the complete overview to other word level features (W08-W19), which were also defined for the utterance level.

Additionally, the expected word duration (W05), the actual word duration (W06), the number of the word's phoneme segments (W07), the relative word likelihood (W20), the duration ratio (W21), features based on the fluctuation of the rate of speech and the duration ratio (W22-W27), features based on the phoneme confusion matrix for correctly pronounced and mispronounced words (C01-C05) and word confidence measures (C06-C08) based on N-best lists of word recognition are employed to score words. The following passages explain some of these features in more detail and give the motivation for their employment.

Relative word likelihood. The distribution of acoustic likelihoods generally depends on the individual speaker. Furthermore, some evaluators may only have marked those words with a relatively bad pronunciation, if comparing it with other words in the sentence. Feature W20 is an attempt to account for these two issues. It is defined as the difference between the likelihood of the current word $\boldsymbol{w}$ and the likelihood of the current utterance $\boldsymbol{u}$, which are both normalized by the corresponding durations d and $t_{\boldsymbol{u}}$.

Duration ratio. The rate of speech is an absolute measure of speaking rate. The advantage of the duration ratio is, that it can measure the relative lengthening or shortening of the duration of a phoneme or word. It is defined as the quotient of the expected duration and the actual duration. For calculation of the expected duration confer the corresponding passage in Section 6.2.2. The duration ratio is also employed as a base feature for measuring fluctuations of the speaking rate.

Fluctuations. A non-native speaker's rate of speech can vary while reading a sentence. This can be due to unfamiliarity with certain words and the difficulty to pronounce certain speech sounds. Consequently, features which are able to measure fluctuations of the local rate of speech and the duration ratio may be a useful indicator of a speaker's fluency.

The objective in word scoring is to identify mispronounced words. The features W22-W27 are intended to detect such words based on fluctuations of the speaking rate: The value of feature W22 becomes large if the current word's rate of speech r_j is lower than the neighbored words' rate of speech r_{j-1} and r_{j+1}, i.e. the event that a non-native speaker gets stuck at an unfamiliar word can be captured. Feature W24 does only consider the left context of the current work, so that its value is smaller than 1.0 for increasing ROS and larger than 1.0 for decreasing ROS, but it serves the same purpose. Features W23 and W25 are analogue to W22 and W24. The only difference is, that the duration ratio $\frac{d^{(exp)}}{d}$ instead of the local rate of speech r_j is employed. Feature W26 is similar to W25, but even considering the duration ratio of the previous two words.

For the features W23 and W25-W27, the duration ratio $d^{(ratio)}$ may not be defined for the first or last words of an utterance. In that case, a default value of 1.0 is used for the duration ratio. Feature W22 is defined as $\frac{r_j}{r_{j+1}}$ for the first word and $\frac{r_j}{r_{j-1}}$ for the last word. Feature W24 is defined as 1.0 for the first word. If there are even too few words in an utterance to calculate any of the fluctuation features W22-W27, their value defaults to 1.0.

Phoneme confusion matrix. In Section 2.2.1 a confidence measure based on the frame level phoneme confusion matrix for both correctly recognized and misrecognized words was introduced. Here the measure definition is transferred to correctly pronounced ($\mathcal{O}$) and mispronounced words ($\mathcal{X}$). Figure 6.3 shows an example of a frame level alignment at the frame level of the reference phoneme sequence $\boldsymbol{p} = (p_1, \ldots, p_t)$ of an utterance fragment to the recognized phoneme sequence $\boldsymbol{q} = (q_1, \ldots, q_t)$. The segmentation of the reference sequence is obtained by forced-alignment. The phoneme symbols of two corresponding frames are different in case of recognition errors, i.e. there are phoneme confusions. A confusion matrix can be estimated for both classes $\mathcal{O}$ and $\mathcal{X}$. These matrices contain the probabilities $P(q_i|p_i)$, that the i-th frame is recognized as phoneme q_i if belonging to phoneme p_i in the reference sequence. Given these two matrices, the ratio of the confusion probabilities of $\mathcal{X}$ and $\mathcal{O}$ can be calculated for each speech frame. Any speech frame which belongs to a pause (sp) or silence (sil) segment in the alignment or the recognition result is discarded in advance. The mean (C01), maximum (C02), minimum (C03), scatter (C04) and median (C05) of these ratios of all frames of the current word are finally used as features (e.g. cf. Eq. 6.4 for C01, C02).

$$C01 = \frac{1}{t}\sum_{i}^{t} \log \frac{P(q_i|p_i, \mathcal{X})}{P(q_i|p_i, \mathcal{O})} \qquad C02 = \max_{i \in \{1,\ldots,t\}} \log \frac{P(q_i|p_i, \mathcal{X})}{P(q_i|p_i, \mathcal{O})} \tag{6.4}$$

Word posterior. For speech recognition the word posterior probability is employed to measure the confidence of each word hypothesis from the recognizer. Here it is used to measure the probability of words in the reference sequence, given a non-native speaker's utterance. The assumption is: the better the pronunciation of a particular word is, the higher is its posterior probability. The calculation of the word posterior probability based on N-best lists is described in Section 2.2.1. For feature calculation the language model (LM) probabilities $P(\boldsymbol{w})$ in Equation (2.12) are set uniform, because at most a unigram LM was employed during recognition. The reason for not using higher order LMs was already explained in Section 6.1. Feature C06 is the relative share of N-best hypothesis, which contains the word w_j at an overlapping interval:

$$C06 = \frac{1}{N}\sum_{\boldsymbol{v}} f([w_j]|[v_i]) \tag{6.5}$$

Function $f([w_j]|[v_i])$ returns 1.0, if the overlap condition for word w_j in the forced-alignment with word v_i in the recognition hypothesis is met. Let o_i be the acoustic observation belonging to word v_i. Furthermore let $\boldsymbol{v}$ denote any of the N-best hypothesis. The posterior probability of word w_j based only on sentence or word likelihoods of N-best hypothesis are used as features C07 and C08:

$$C07 = \frac{\sum_{\boldsymbol{v}} P(\boldsymbol{o}|\boldsymbol{v}) f([w_j]|[v_i])}{\sum_{\boldsymbol{v}} P(\boldsymbol{o}|\boldsymbol{v})} \qquad C08 = \frac{\sum_{\boldsymbol{v}} P(o_i|v_i) f([w_j]|[v_i])}{\sum_{\boldsymbol{v}} P(o_i|v_i)} \tag{6.6}$$

Since we are interested in a posterior probability which is only based on acoustic scores, the probabilities $P(\boldsymbol{v})$ and $P(v_i)$ of word sequences and words are not taken into account.

Table 6.11: *Definition of word level pronunciation quality features.*

Score-ID	Name	Symbol	Definition
-	Phoneme segment lh.	$L(\boldsymbol{y})$	$\log P(\boldsymbol{y}\|\text{HMM}\lambda)$
W01	Word likelihood	$L^{(word)}(\boldsymbol{w})$	$\sum_{i=1}^{n} L(\boldsymbol{y}_i)$
W02	Mean segment lh.	$\overline{L(\boldsymbol{y})}$	$\frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{y}_i)$
W03	-	-	$\frac{1}{R^{(phon)}}\sum_{i=1}^{n} L(\boldsymbol{y}_i)$
W04	-	-	$\sum_{i=1}^{n} \frac{L(\boldsymbol{y}_i)}{t_i}$
W05	Expect. word duration	$d^{(exp)}$	$\sum_{i=1}^{n} t_i^{(exp)}$
W06	Total word duration	d	$\sum_{i=1}^{n} t_i$
W07	# phoneme segments	$n = \|\boldsymbol{p}\|$	# phonemes in word $\boldsymbol{w}$
W08	Rate of speech (ROS)	$r = \frac{n}{d}$	# phonemes / word duration
W09	Mean phoneme duration	$\frac{1}{r} = \frac{d}{n}$	word duration / # phonemes
W10	Word duration score	$S_{dur}^{(word)}(\boldsymbol{w})$	$\sum_{i=1}^{n} S_{dur}^{(phon)}(t_i * R^{(phon)}\|p_i, \boldsymbol{y}_i)$
W11	-	-	$\frac{1}{n} S_{dur}^{(word)}(\boldsymbol{w})$
W12	-	-	$\frac{1}{R^{(phon)}} S_{dur}^{(word)}(\boldsymbol{w})$
W13	Recognition performance	-	$\frac{\text{mineditdist}(\boldsymbol{q},\boldsymbol{p})}{\max\{\|\boldsymbol{q}\|,\|\boldsymbol{p}\|\}}$
W14	Phoneme sequence lh.	$L(\boldsymbol{q}\|\text{LM})$	Phoneme sequence likelihood
W15	-	-	$\frac{1}{n} L(\boldsymbol{q}\|\text{LM})$
W16	-	-	$\frac{1}{R^{(phon)}} L(\boldsymbol{q}\|\text{LM})$
W17	Minimum phoneme lh.	-	$\min_{i=1\ldots n} L(\boldsymbol{y}_i)$
W18	Maximum phoneme lh.	-	$\max_{i=1\ldots n} L(\boldsymbol{y}_i)$
W19	Scatter of phoneme lh.	-	$\frac{1}{n}\sum_{i=1}^{n}[L(\boldsymbol{y}_i) - \overline{L(\boldsymbol{y})}]^2$
W20	Relative word lh.	-	$\frac{L^{(word)}(\boldsymbol{w})}{d} - \frac{L^{(sent)}(\boldsymbol{u})}{t_{\boldsymbol{u}}}$
W21	Duration ratio (DR)	$d^{(ratio)}$	$\frac{d^{(exp)}}{d}$
W22	Context fluct. (ROS)	-	$\frac{2r_j}{r_{j-1}+r_{j+1}}$
W23	Context fluct. (DR)	-	$\frac{2d_j^{(ratio)}}{d_{j-1}^{(ratio)}+d_{j+1}^{(ratio)}}$
W24	ROS fluctuation	-	$\frac{r_{j-1}}{r_j}$
W25	DR fluctuation	-	$\frac{d_{j-1}^{(ratio)}}{d_j^{(ratio)}}$
W26	DR fluctuation	-	$\frac{d_{j-2}^{(ratio)}+d_{j-1}^{(ratio)}}{2d_j^{(ratio)}}$
W27	Context fluct. (DR)	-	$\frac{4d_j^{(ratio)}}{d_{j-2}^{(ratio)}+d_{j-1}^{(ratio)}+d_{j+1}^{(ratio)}+d_{j+2}^{(ratio)}}$

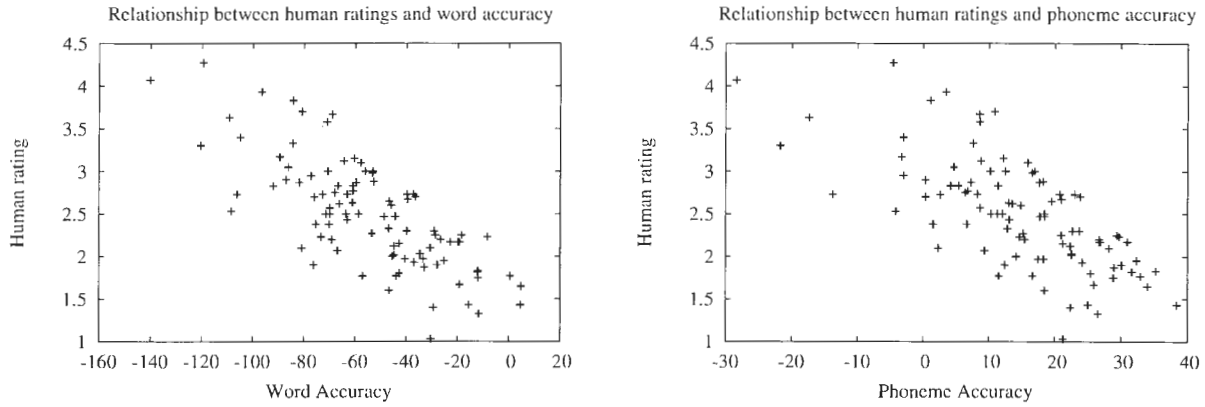


Figure 6.7: *Relationship between human ratings and phoneme and word accuracy for the TIMIT sentences. The values of the word accuracy are extremely low, because no statistical language model was employed during recognition and it is non-native speech (ATR SLT data).*

6.2.4 Speaker Scores

These average speaker level human ratings (HumRating), the relative number of words each speaker mispronounced (MisRatio) and the recognition performance (PhonAcc, WordAcc) in an unconstrained phoneme and word recognition task without a statistical language model are summarized in Appendix A.1.1. The recognition accuracies are calculated from the recognition result of the 48 TIMIT sentences with a monophone acoustic model trained on native English speech data. The relationship between the four variables {HumRating, MisRatio, PhonAcc, WordAcc} can be examined by looking at the inter-variable correlations (cf. Table 6.12).

Table 6.12: *Correlation between speaker level human ratings (HumRating), relative number of mispronounced words (MisRatio), phoneme accuracy (PhonAcc) and word accuracy (WordAcc).*

Correlation	MisRatio	WordAcc	PhonAcc
HumRating	+0.87	-0.75	-0.70
MisRatio		-0.58	-0.52

It is not surprising that speakers' ratings and the number of words they mispronounce are highly correlated (0.87). This suggests, that the evaluators were strongly influenced by mispronounced speech sounds in their rating decision. The correlation of human ratings with phoneme (0.70) and word accuracy (0.75) is lower, but the values still suggest a strong relationship between recognition errors and pronunciation quality. The relationship is also clear from the phoneme and word accuracy versus rating plots in Figure 6.7. The correlation between mispronounced words and recognition accuracy, either word (0.58) or phoneme (0.52), is much lower. This may be due to the fact, that for calculation of the recognition accuracy not only substitutions, which ideally correspond to mispronounced words, but also insertions, deletions and substitutions of tokens are taken into account and to the phenomenon that there are many recognition errors.

Table 6.13: *Correlation between averaged utterance level ratings and scores.*

Correlation	L7	K1	R4	P2	D2	X1	M3	E1
Utter. level	-0.42	+0.48	+0.39	+0.32	-0.46	-0.45	-0.40	+0.30
Speaker level	-0.58	+0.80	+0.55	+0.55	-0.72	-0.76	-0.60	+0.61

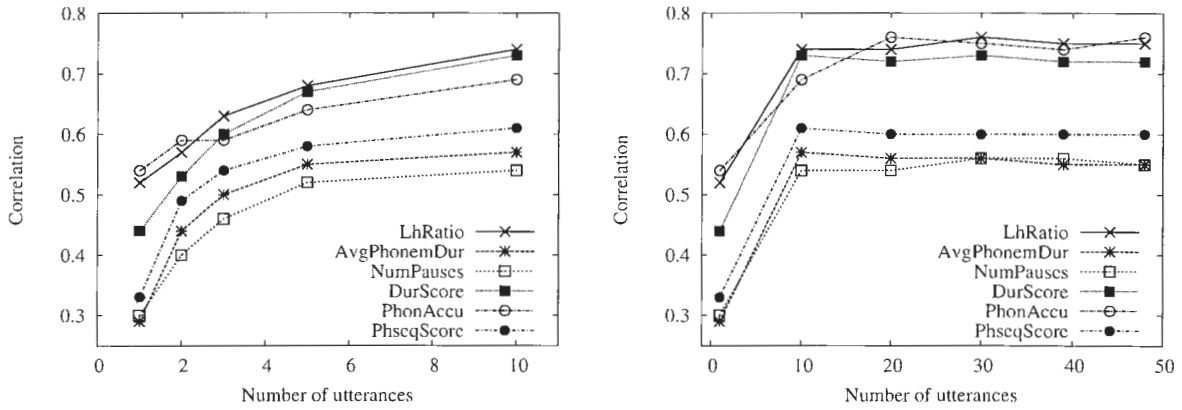


Figure 6.8: *Relationship between the rating to score correlation and the number of utterances for averaging in order to obtain speaker level ratings and scores (ATR SLT data). The scores are LhRatio (K1), AvgPhonemDur (R4), NumPauses (P2), DurScore (D2), PhonAccu (X1) and PhseqScore (M3).*

For the speakers of the ATR SLT non-native database no speaker level annotations for pronunciation quality are available. However a speaker level pronunciation rating is easily obtained by averaging all the utterance ratings available for each speaker (cf. Section 5.2).

In Section 6.2.2 the correlation between several pronunciation-related features and human ratings was examined on the utterance level. That investigation can be extended to the speaker level. By averaging the ratings and scores of two or more utterances of the same speaker, speaker level ratings and scores are obtained. The benefit of averaging is, that scores and ratings become more accurate and reliable. Table 6.13 compares the speaker level correlation of the most successful scores from Section 6.2.2. For all scores there is a remarkable increase in score to rating correlation.

Figure 6.8 shows the relationship between the correlation and the number of utterances employed for averaging. Even when averaging only the first two utterances, there is an increase in correlation for all scores. The correlation reaches almost its maximum after about 10 utterances. There is a rating to score correlation higher than 0.70 for the scores K1 and D1. Consequently, segmental aspects of pronunciation seem to be as important to the human evaluators as temporal aspects.

The high correlation for score X1 can easily be explained by the fact, that the recognition performance will be low, if the duration and the spectral characteristics of phonemes uttered by non-natives are different from native phonemes.

Chapter 7

Experimental setup

In this chapter the features defined in Chapter 6 are employed for the classification of words and the scoring of single utterances. While the target of word classification (Section 7.1) is the detection of mispronounced words, the aim of utterance scoring (Section 7.2) is to obtain an automatic assessment of the overall pronunciation quality of an utterance. If more than one utterance from the same speaker is available, an assessment of a speaker's pronunciation proficiency can be made (Section 7.3).

7.1 Word Classification

The task is to build a classifier which can discriminate correctly pronounced and mispronounced words. Each word is represented as one feature vector $\boldsymbol{c}$. 35 word level pronunciation features are defined in Section 6.2.3. However, since the amount of available data is relatively small (37,920 samples = 96 speakers times 395 words) and the discrimination ability of each feature is unknown, the number of feature components has to be reduced. To achieve this, methods for feature selection and feature space transformation are applied (see Figure 7.1).

Feature selection. In order to determine the single best feature and heuristically a suboptimal set of good features, the floating search (FS) algorithm (cf. Section 2.4.2) is employed. For that algorithm an optimization criterion, which describes the quality of a feature set has to be defined. Here, the criterion is a gain function $f(\boldsymbol{M}, \boldsymbol{P})$ which is defined as the sum of the point-wise multiplication of the confusion matrix $\boldsymbol{P} = (p_{\kappa\lambda})$ of the classifier with a gain matrix $\boldsymbol{M} = (m_{\kappa\lambda})$:

$$f(\boldsymbol{M}, \boldsymbol{P}) = \frac{1}{K} \sum_{\kappa=1}^{K} \sum_{\lambda=1}^{K} m_{\kappa\lambda} p_{\kappa\lambda} \tag{7.1}$$

Entry $p_{\kappa\lambda}$ of matrix $\boldsymbol{P}$ is the probability that the classifier confuses class λ with class κ and entry $m_{\kappa\lambda}$ of matrix $\boldsymbol{M}$ defines the corresponding gain. With this definition the search algorithm will find a feature set which maximizes the classification gain as much as possible. The notion of classification gain is similar to the classification risk (cf. Section 2.4.1) with the difference, that the risk has to minimized and the gain to be maximized.

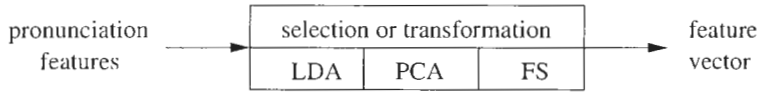


Figure 7.1: *Feature preprocessing to reduce feature dimension.*

In principle any classifier can be used to obtain the confusion matrix. Nevertheless, the choice should be made in favor of a classifier which is able to model the class distributions as accurately as possible. However, as a large number of feature subsets has to be examined by the algorithm, the evaluation speed needs to be fast. To meet this requirement, the Gaussian classifier with single densities is employed, since the estimation of its parameters as well as the classification of test samples can be carried out quickly.

Feature transformation. An alternative to feature selection is the transformations of the original features space. Two standard transformations are LDA and PCA. LDA yields feature components which are ordered by their class discrimination ability, PCA components are ordered after their variance. Which transformation to chose depends on the sample distribution.

As mentioned in Chapter 6, some of the pronunciation features are highly correlated. The higher the correlation between two features, the less beneficial will be the usage of both of them for classifier construction. An advantage of PCA is, that it yields uncorrelated feature components. By applying PCA to the original features space, features components with higher correlations are removed automatically if the final dimension of the feature space is set small enough.

Classification. There are several possibilities for the partitioning of words into classes. Details will be discussed in Chapter 8 when reporting about experimental results. In the following it is assumed that there are at least two classes: Class $\mathcal{O}$ for correctly pronounced words and class $\mathcal{X}$ for mispronounced words. Furthermore there may be a rejection class $\mathcal{R}$ to avoid unreliable classification decisions.

The Gaussian classifier with either a single Gaussian density or multiple Gaussian densities is employed. In the single density case, one multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_\kappa$ and full covariance matrix $\boldsymbol{\Sigma}_\kappa$ is obtained by ML estimation from a set of samples $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots\}$ for each class κ. A model with multiple Gaussian densities is called Gaussian mixture models (GMM) and was introduced in Chapter 2. For classification, the Gaussian classifier yields one score for each class: the conditional probability $P(\boldsymbol{c}|\omega)$. The standard way is to use the Bayes rule to decide whether a word's pronunciation is correct or wrong (cf. Equation 7.2, Figure 7.2). For this simple setup, the prior probabilities $P(\omega)$ should be uniform, since it is not reasonable to assume a relatively higher probability for any each class if nothing in known about the target speaker's pronunciation proficiency.

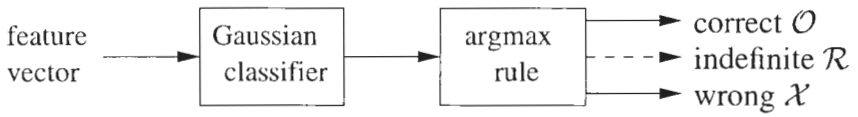


Figure 7.2: *Setup for the detection of mispronounced words.*

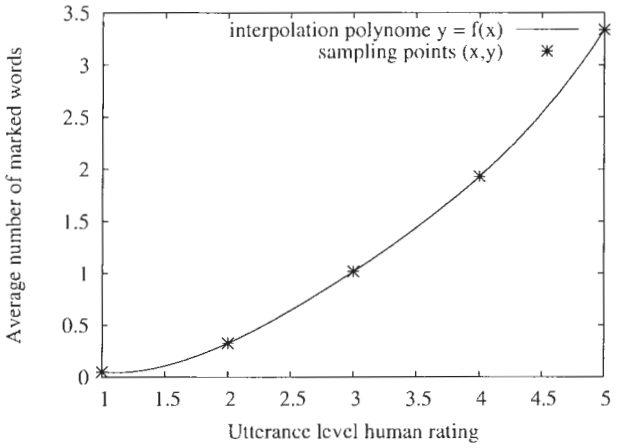


Figure 7.3: *Relationship between the utterance rating and the average number of marked words (ATR SLT data).*

$$\omega^* = \underset{\omega \in \{\mathcal{O},\mathcal{X},\mathcal{R}\}}{\operatorname{argmax}} P(\boldsymbol{c}|\omega)P(\omega) \tag{7.2}$$

Besides the word level features additional information may be considered for classification: the sentence score, prior probabilities and marking strictness. The sentence score can be derived automatically from the current utterance. Figure 7.3 and Table 7.1 show, that the worse the pronunciation score of an utterance, the more words are marked as mispronounced. The graph in Figure 7.3 is obtained by estimating the interpolation polynome through five supporting points (h, m_h), where $h \in \{1, \ldots, 5\}$ and m_h is the average number of words marked as mispronounced in utterances with rating h.

Table 7.1: *Distribution of the number of marked words for each utterance rating (ATR SLT data).*

# words marked	none	one	two	three	four	five
Rating 1	0.96	0.04	0.00	0.00	0.00	0.00
Rating 2	0.70	0.28	0.02	0.00	0.00	0.00
Rating 3	0.30	0.44	0.21	0.04	0.01	0.00
Rating 4	0.09	0.27	0.37	0.19	0.07	0.01
Rating 5	0.00	0.06	0.17	0.33	0.35	0.10

Prior probabilities and strictness have to be set manually. The parameter strictness is intended to control the number of words to be marked out of a set of probable mispronunciation candidates. By modification of the priors, the probability for marking words as mispronounced can be increased or decreased. Especially it may be useful to set a higher probability for class $\mathcal{O}$ and a lower for class $\mathcal{X}$ to not confuse the student by false rejections.

Figure 7.4 shows the extended setup for word level classification. The output of the Gaussian classifier and the additional information is combined in a post processing unit. It remains to

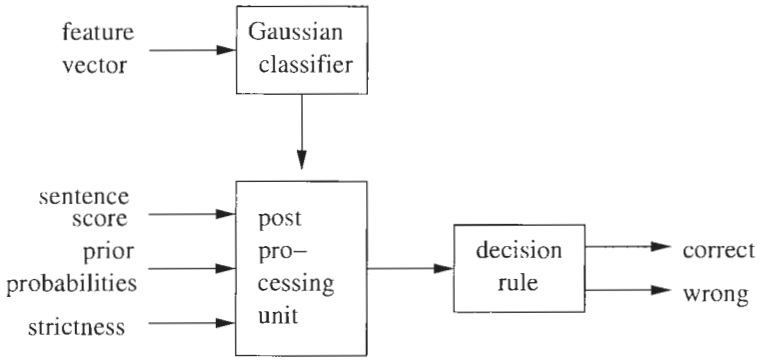


Figure 7.4: *Extended setup for the detection of mispronounced words.*

design the post-processing procedure and the final decision rule. One possibility is to use prior probabilities, the likelihood ratio $\log P(\boldsymbol{c}|\mathcal{X}) - \log P(\boldsymbol{c}|\mathcal{O})$ of the classifier output, the mean value of these likelihood ratios of all words in the current utterance and the sentence score as new features to train a classifier, e.g. decision tree. The procedure for automatic detection of mispronounced words would then consist of two steps: In the first stage scoring of the utterance and classification of all its words with a Gaussian classifier. The final classification decision w.r.t. mispronounced words is obtained with the decision tree in a second stage.

For evaluation in Chapter 8, only a variation of the class prior probabilities $P(\mathcal{O})$ and $P(\mathcal{X})$ to shift the decision boundary is carried out in order to obtain performance curves based on precision and recall. Furthermore a recombination of the classification result with the three classes $\mathcal{O}$, $\mathcal{X}$ and $\mathcal{R}$ is considered.

7.2 Utterance Scoring

As for word classification, each utterance $\boldsymbol{u}$ is represented as one feature vector $\boldsymbol{c} = (c_1, \ldots, c_d)$. To assess the overall pronunciation quality of a single utterances, there are two possible viewpoints:

- **Hard scoring:**
  assignment of a discrete value $h \in \{1, 2, 3, 4, 5\}$
- **Soft scoring:**
  assignment of a continuous value $s \in [1.0, 5.0]$

From these two viewpoints, the approaches outlined in Figure 7.5 can be applied for realizing an utterance scoring scheme. For hard scoring it is necessary to make a "hard" decision, i.e. the outcome must be an integer value like 2 or 3 but nothing in between like 2.5. Since the human reference ratings may also be continuous, e.g. mean rating of two or more evaluations, they have to be discretized first. Rounding is the simplest way to achieve that. This yields five rating classes. Finally, pattern recognition methods for the discrimination of multiple classes can be applied.

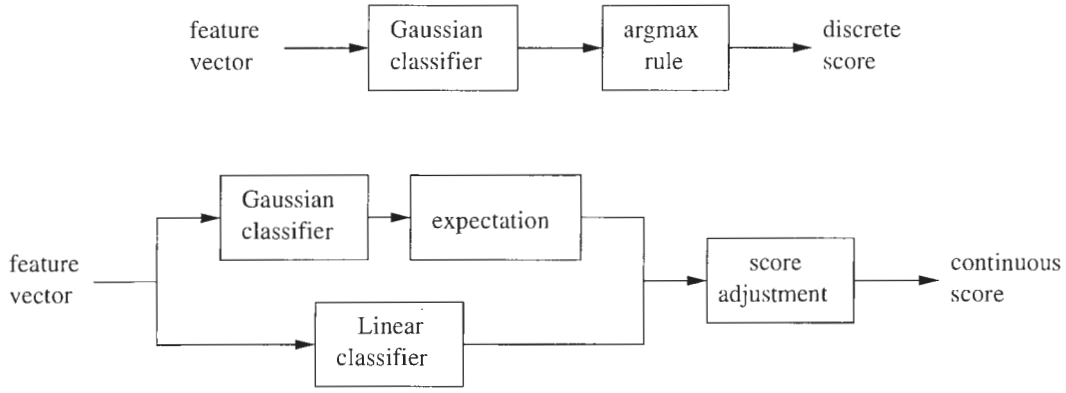


Figure 7.5: *Scoring of single utterances.*

There is only relatively little training data, but relatively many features (33) were defined. Consequently, the floating search (FS) technique is also applied for utterance classification. The difference to word classification is, that there are five classes to recognize and a different gain matrix has to be chosen. For each rating class one Gaussian density is estimated. Mixture model densities are not employed, because the available number of samples is too small.

The Gaussian classifier yields one probability $P(\boldsymbol{c}|h)$ for each class h. The hard classification decision is made with the argmax-rule. Class prior probabilities $P(h)$ are assumed uniform.

Hard scoring has the disadvantage, that the differences between the reference rating and the estimated score can easily become greater than 1.0. This is due to rounding and the hard classification decision. Soft scoring alleviates this problem, since scores may assume any value within the interval $[1.0, 5.0]$.

The soft scoring scheme is illustrated in the bottom part of Figure 7.5. Two methods for soft scoring are used: Gaussian classification and Linear classification, both with appropriate post processing. In the Gaussian approach the classifier output, i.e. the class likelihoods $P(\boldsymbol{c}|h)$, are used to calculate the expected score $E[h|\boldsymbol{c}]$:

$$s = E[h|\boldsymbol{c}] = \sum_h h * P(h|\boldsymbol{c}) \qquad P(h|\boldsymbol{c}) = \frac{P(\boldsymbol{c}|h)P(h)}{\sum_g P(\boldsymbol{c}|g)P(g)} \tag{7.3}$$

This approach was already employed in literature (cf. Section 3.5). The linear classifier yields a continuous value, which is a linear combination of one or more pronunciation features:

$$s = r_0 + \sum_{i=1}^{d} r_i * c_i \tag{7.4}$$

The coefficients r_i of the linear combination can be estimated by linear regression for a set of training samples $\{(\boldsymbol{c}, h)\}$, where $\boldsymbol{c}$ is the vector of utterance level features and h the human reference rating. The same relationship (Eq. 7.4) is then used to predict the utterance score s with the utterance features $\boldsymbol{c}$. To get robust regression coefficients, the training data for estimating the coefficients must be balanced w.r.t. the reference ratings. However, the data available for this

thesis do not fulfill this requirement. Resampling can be used to make the distributions uniform: from the available training samples for each class as many samples as in the set of the class with highest cardinality are selected randomly with replacement.

Experimental results in Chapter 8 will show, that especially the linear classifier has the tendency to assign too bad scores even for the training sample if reference ratings are close to 1, and too good scores if reference ratings are close to 5. To alleviate this problem, the score values s are adjusted to s' with a combination of two transformations:

1. linear transformation:
   $x = g(s, \boldsymbol{a}) = a_0 + a_1 s$

2. multiplicative polynomial transformation:
   $s' = x * f(x, \boldsymbol{b}) = b_0 x + b_1 x^2 + b_2 x^3 + \ldots + b_k x^{k+1}$

The purpose of the linear transformation is to adjust the mean of the automatically assigned scores and the slope of the regression function (Eq. 7.4). The multiplicative polynomial transformation can cope with certain non-linear distortions of the scoring output. From Figure 8.1 the necessity to apply both transformations is apparent.

The parameters $\boldsymbol{a} = (a_0, a_1)$ of $g(s, \boldsymbol{a})$ can be obtained by linear regression based on the actual scoring output for the training data and the corresponding reference labels. However, instead of estimating $\boldsymbol{a}$ for the mapping of scores s to the reference ratings x, the variables s and x are interchanged, i.e. the parameters $\boldsymbol{a}'$ of $s = g(x, \boldsymbol{a}')$ are calculated. The coefficients $\boldsymbol{a}$ of the desired linear mapping are then taken from the reciprocal function of $g(x, \boldsymbol{a}')$, i.e. $a_0 = -\frac{a'_0}{a'_1}$ and $a_1 = \frac{1}{a'_1}$.

Finally, the coefficients $\boldsymbol{b}$ of the polynome $s' = f(x, \boldsymbol{b})$ must be determined. For training samples which belong to the discrete rating class h (by rounding), the mean μ_h of the estimated scores $x = g(s, \boldsymbol{a})$ is calculated. To bring a score for a sample of class h closer to its reference value, it is multiplied by $\frac{h}{\mu_h}$. A more smooth transformation is the multiplication x with a polynome $f(x, \boldsymbol{b})$, which fits through the coordinates $(h, \frac{h}{\mu_h})$. The polynome coefficient $\boldsymbol{b}$ are obtained by interpolation with Newton's method [BS97].

Such a score adjustment has also the disadvantage, that a scoring result of e.g. $t = s + d_1$ for a reference value of $u = s - d_2$ with $d_1, d_2 > 0$ gets worse, if $t < t * f(g(t, \boldsymbol{a}), \boldsymbol{b})$ and $u > u * f(g(u, \boldsymbol{a}), \boldsymbol{b})$ and vice versa.

7.3 Speaker Scoring

The overall pronunciation proficiency of a non-native speaker can be estimated on a set of that speaker's utterances. The sentence set should ideally be phonetically balanced or at least cover a wide range of phonetic contexts. Otherwise the proficiency assessment is likely to be biased. For example a speaker's pronunciation skill would be overestimated, if speech sounds he pronounces correctly occur more frequently in the sentences than those he is likely to mispronounce.

There are possibilities to estimate a speaker level rating directly avoiding to go the way round utterance scoring (cf. Section 3.6). However such a scoring method is rather computationally intensive and requires a certain amount of speech data with examples for all phonemes. Utterance-based estimation is much faster and straightforward. Figure 7.6 shows the experimental setup for speaker level scoring. The Figure describes two approaches which differ in the order initial features and final scores are processed. The method in the upper part of the figure first extracts the feature vectors $\boldsymbol{c}_i$ from all utterances, averages the feature vectors and finally performs scoring identical to utterance scoring based on a single feature vector. The method in the bottom part first scores each utterance separately with its corresponding feature vector and finally takes the average of all utterance scores.

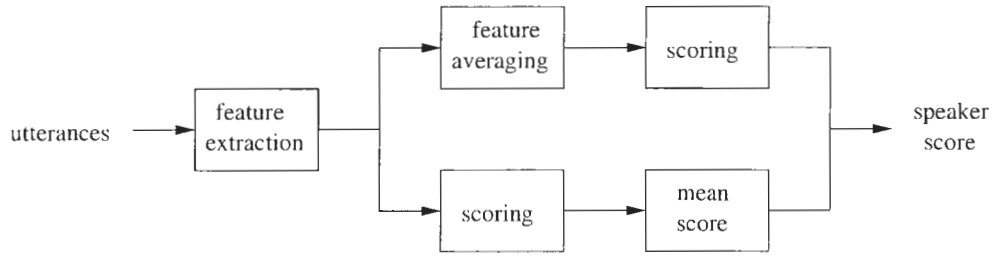


Figure 7.6: *Scoring of multiple utterance to obtain a speaker level score.*

Chapter 8

Results

This chapter reports the results for experiments described in Chapter 7. Experiments are carried out for two corpora: ATR SLT non-native English database and FAU LME non-native English children speech corpus. For validation purposes some native speech is employed additionally: English speech from seven adult native speakers uttering the same sentences as the non-native speakers in the ATR SLT database, and British English children speech from the PF_STAR BE corpus.

The utterance scoring accuracy is indicated by three measures: the correlation (COR) for soft scoring and the class-wise average recognition rate (CL) as well as the class-wise average recognition rate which tolerates confusions of neighbored classes (CL-A) for hard scoring. For example, if the reference rating is 3 and the classification result is 4 or 5 it is still considered as correct with CL-A.

Discrimination of correctly pronounced and mispronounced words is a 2-class classification problem. Besides the total recognition rate (RR), i.e. the relative share of correctly classified tokens, and the class-wise average recognition rate (CL), the accuracy can also be indicated by the measures recall (REC) and precision (PRC). They can be calculated from the confusion matrix $\boldsymbol{P}$ of the classifier:

$$\boldsymbol{P} = \left[\begin{array}{c|c} c_1 & f_1 \\ \hline f_2 & c_2 \end{array}\right] = \left[\begin{array}{c|c} \text{\# correctly assigned to } \mathcal{O} & \text{\# tokens misclassified as } \mathcal{X} \\ \hline \text{\# tokens misclassified as } \mathcal{O} & \text{\# correctly assigned to } \mathcal{X} \end{array}\right] \tag{8.1}$$

Table 8.1 summarizes the definitions of all measures of classification accuracy. Recall ω is the relative share of the tokens with reference label ω, which are classified as ω. Precision ω is the relative share of correct classifications among the tokens classified as ω. By shifting the decision boundary a recall vs. precision for one class or a recall vs. recall curve for both classes can be obtained. The decision boundary can be shifted by setting the class prior probabilities $P(\mathcal{O}) = 1 - P(\mathcal{X})$ to values sampled from the interval $[0.0; 1.0]$.

The number samples of each class for word classification and each rating class for utterance scoring is unbalanced. Resampling is carried out to obtain a balanced training and test set for each class. For each class as many examples as in the sample of the class with highest cardinality are redrawn randomly with replacement.

Table 8.1: *Accuracy measures for word classification experiments.*

Abbrev.	Name	Definition
CL	class-wise average recognition rate	$\frac{1}{2}(\frac{c_1}{c_1+f_1} + \frac{c_2}{c_2+f_2})$
RR	total recognition rate	$\frac{c_1+c_2}{c_1+c_2+f_1+f_2}$
REC 1	recall of class 1	$\frac{c_1}{c_1+f_1}$
PRC 1	precision of class 1	$\frac{c_1}{c_1+f_2}$
REC 2	recall of class 2	$\frac{c_2}{c_2+f_2}$
PRC 2	precision of class 2	$\frac{c_2}{c_2+f_1}$

Table 8.2: *Number of utterances w.r.t. human ratings in the ATR SLT database.*

Data	Training set					Test set				
Class	1	2	3	4	5	1	2	3	4	5
Part. $\mathcal{A}$	333	1,298	1,308	462	55	49	493	436	161	13
Part. $\mathcal{B}$	335	1,356	1,289	430	46	47	435	455	193	22
Part. $\mathcal{C}$	160	1,322	1,441	492	41	222	469	303	131	27
Part. $\mathcal{D}$	318	1,397	1,194	485	62	64	394	550	138	6

8.1 ATR SLT Data

Tables 8.2 and 8.3 show the amount of data available for utterance scoring and word classification experiments. There are four groups of annotators and the members of each group evaluated the data of 24 non-native speakers. The data was divided into a training set consisting of the data of three groups (72 speakers) and a test set consisting of the data of one group (24 speakers). Hence, there are four possibilities ($\mathcal{A}, \mathcal{B}, \mathcal{C}$ and $\mathcal{D}$) for data partitioning, if training and test set are kept disjoint w.r.t. speakers and evaluators. Initial experiments are carried out with $\mathcal{C}$. Final experiments are done with 4-fold cross-validation (CV), i.e. evaluation is carried out for all four possibilities taking the average result.

The jackknife mean estimate of the human evaluation is employed as reference rating for each utterance. For hard utterance scoring these continuous ratings are rounded to obtain five discrete rating classes. For classification experiments on the word level, words are categorized according

Table 8.3: *Number of words in each class (ATR SLT data).*

Data	Training			Test		
Class	$\mathcal{O}$	$\mathcal{R}$	$\mathcal{X}$	$\mathcal{O}$	$\mathcal{R}$	$\mathcal{X}$
Part. $\mathcal{A}$	20,532	5,173	2,735	6,533	2,154	793
Part. $\mathcal{B}$	20,639	4,910	2,891	6,426	2,417	637
Part. $\mathcal{C}$	19,445	6,307	2,688	7,620	1,020	840
Part. $\mathcal{D}$	20,579	5,591	2,270	6,486	1,736	1,258

to the number of evaluators who marked it. A word which was not marked at all belongs to class $\mathcal{O}$ and is considered as correctly pronounced. The words with two or more markings are treated as mispronounced and are part of class $\mathcal{X}$. For words with only one marking it is difficult to decide whether they should be assigned to $\mathcal{O}$ or $\mathcal{X}$. They are categorized into class $\mathcal{R}$.

The speech recognizer and its modules are set up in the same way as described in Section 6.1. The phoneme confusion matrices for calculation of the features C01-C05 are estimated in advance with the training data and the associated word markings.

8.1.1 Utterance Scoring

First the results for *soft scoring* with linear combination of one, two or more features is reported. Only those features in each group (L,K,D,R,P,M,E) for which already a relatively high correlation to the human ratings was observed in Section 6.2.2 are considered. The IDs of these feature are: L7,K1,K3,X1,M3,P1,P2,D2,R4,E2,E3. Table 8.4 shows the performance of utterance soft scoring by linear feature combination. For single features only the bias term and the slope of the straight line, which describes the relationship between the feature value and the human reference rating, are estimated. The best single feature is the likelihood ratio (K3). The feature combination with the highest rating to score correlation is {K3,X1,M3}. If the X (recognition performance), M (phoneme sequence likelihood) and K (likelihood ratio) features, which are based on the recognition result, are not available, the same performance cannot be achieved even if several other features {D2,L7,R4,P1,E3} are employed.

Table 8.4: *Experimental results for utterance soft scoring with non-native ATR SLT data. Performance is measured by the correlation coefficient.*

#F	Features	Part. $\mathcal{C}$	Cross-Vali
4	{K3,X1,M3,D2} (**)	**0.64**	**0.59**
3	{K3,X1,M3}	0.64	0.59
2	{K3,M3}	0.63	0.56
2	{K3,X1}	0.59	0.55
5	{D2,L7,R4,P1,E3}	**0.57**	**0.51**
3	{D2,L7,E3}	0.56	0.51
2	{D2,E3}	0.56	0.51
2	{D2,L7}	0.52	0.47
1	Likelihood ratio (K3)	**0.57**	**0.52**
1	Likelihood score (E3)	0.49	0.44
1	Recognition performance (X1)	0.48	0.46
1	Duration score (D2)	0.47	0.45
1	Likelihood score (L7)	0.47	0.42
1	Phoneme sequence lh. (M3)	0.48	0.40

The highest rating to score correlation is 0.59 with cross-validation. In Section 5.1 an average inter-rater open correlation of 0.60 was reported. Consequently, the automatic scoring

by linear combination of features may be considered almost as reliable as the human evaluation. However, a closer look at the scoring result reveals, that the output of the regression function is distorted. Utterances with a good pronunciation quality are scored too bad and vice versa. With a linear and a polynomial transformation as described in Section 7.2 this undesired effect can be alleviated. If ratings and scores are discretized by rounding, the classification gain $f(\boldsymbol{M}_2, \boldsymbol{P})$ can be calculated. For example, it descreases from -0.59 without adjustment to -0.24 after adjustment when scoring is based on the feature set (**). The correlation between human ratings and scores does not change significantly with and without score adjustment (cf. Figure 8.1).

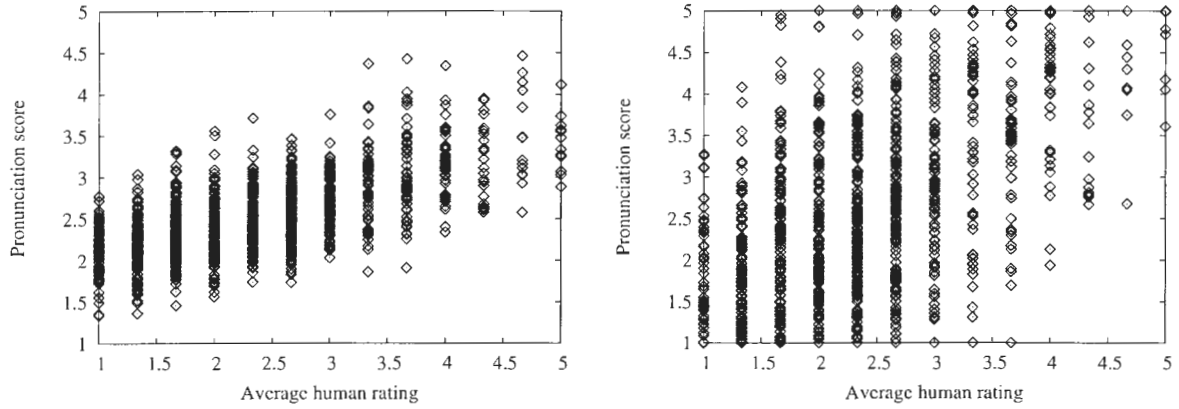


Figure 8.1: *Soft scoring result for Part. C with linear regression and feature combination (**). Plot shows ratings vs. scores before (left) and after (right) adjustment with a linear and a polynomial transformation (ATR SLT data).*

Nevertheless, linear feature combination does not work as good for pronunciation scoring as the reader might have been expected. The reason is an unsatisfactory scoring accuracy for native speech as shown in Table 8.5. The values in each column describe the recall w.r.t. each rating class h for each scoring approach. The soft scoring result was discretized for easier comparison with the approach based on the Gaussian classifier. Without score adjustment only 7.4% of the natives' utterances are classified correctly. If adjustment is applied, the scoring accuracy increases to 85.4%.

Experimental results for both *hard and soft scoring with a Gaussian classifier* are shown in Tables 8.6 and 8.7. Besides manually selected feature combinations, which proved to be successful for soft scoring with linear regression, automatic feature selection with the floating

Table 8.5: *Scoring of utterances of native speakers with the Gaussian classifier and linear regression (LR) with and without score adjustment (ATR SLT data).*

Method	1	2	3	4	5
LR (**)	7.4	88.7	3.9	0.0	0.0
LR [adjust]	85.4	10.7	3.6	0.3	0.0
Gaussian (*)	**90.2**	5.7	3.6	0.6	0.0

Table 8.6: *Experimental results for utterance hard and soft scoring with the Gaussian classifier. Results are shown for evaluation with Part. C (ATR SLT data).*

#F	Features		COR	CL (in %)	CL-A (in %)
5	{K3,M3,X2,Y1,L3}		0.58	**46.4**	86.8
4	{K3,M3,X2,Y1}	FS,M_1	0.58	45.4	86.7
3	{K3,M3,X2}		**0.60**	42.2	**87.3**
2	{K3,M3}		**0.60**	42.3	86.4
5	{E3,K2,X1,Y1,M1}		0.58	**48.0**	85.2
4	{E3,K2,X1,Y1}	FS,M_2	0.58	47.4	85.1
3	{E3,K2,X1}		**0.60**	46.1	**85.5**
2	{E3,K2}		0.59	44.1	84.5
5	{K3,M3,E1,P2,X1}		**0.60**	**46.1**	**87.2**
4	{K3,M3,E1,P2}	FS,COR	0.59	45.7	86.1
3	{K3,M3,E1}		0.58	43.9	86.7
2	{K3,M3}		**0.60**	42.3	86.4
3	{K3,X1,M3}		**0.60**	**40.9**	**85.9**
5	{D2,L7,R4,P1,E3}		0.51	39.0	82.2
3	{D2,L7,E3}	manual	0.52	38.1	81.1
2	{D2,E3}		0.53	37.6	80.4
2	{K3,X1}		0.56	34.5	79.9
1	Likelihood ratio (K3)		**0.56**	36.0	**79.8**
1	Duration-normalized lh. (E3)		0.48	**39.9**	76.4
1	Phoneme sequence lh. (M3)	single	0.47	37.7	75.6
1	Phoneme recognition (X1)		0.46	35.5	71.5
1	Phoneme duration score (D2)		0.45	32.8	73.6

search (FS) algorithm was carried out. Different types of optimization criterions are examined: the gain function (Eq. 7.1) with two different gain matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ as shown in Equation 8.2 and the correlation coefficient. The matrices are designed so that the more severe the confusion is, the higher the loss becomes, e.g. the worst case is a confusion between class 1 and class 5. The only difference between $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ is, that the gain function based on $\boldsymbol{M}_1$ ignores confusions of neighboring classes, while with $\boldsymbol{M}_2$ they have a negative effect.

$$\boldsymbol{M}_1 = \begin{bmatrix} +1 & 0 & -2 & -4 & -8 \\ 0 & +1 & 0 & -2 & -4 \\ -2 & 0 & +1 & 0 & -2 \\ -4 & -2 & 0 & +1 & 0 \\ -8 & -4 & -2 & 0 & +1 \end{bmatrix} \qquad \boldsymbol{M}_2 = \begin{bmatrix} +1 & -1 & -2 & -4 & -8 \\ -1 & +1 & -1 & -2 & -4 \\ -2 & -1 & +1 & -1 & -2 \\ -4 & -2 & -1 & +1 & -1 \\ -8 & -4 & -2 & -1 & +1 \end{bmatrix} \tag{8.2}$$

The likelihood ratio (K3) turns out to be the best single feature. The scoring accuracy for native utterances is also highest (94.4%) with feature K3. If other features like E3, M3, D2 or X1

Table 8.7: *Results for scoring utterances of non-natives and natives with the Gaussian classifier (ATR SLT data).*

#F	Features	Non-Native (Cross-Vali) COR	CL	CL-A	$f(\boldsymbol{M}_2, \boldsymbol{P})$	Native RR
3	{E3,K2,X1} (*)	**0.54**	**40.1**	**79.9**	-0.25	90.2
3	{K3,M3,X2}	0.54	38.7	78.6	-0.27	91.1
2	{K3,M3}	0.54	39.6	78.0	-0.27	92.7
5	{K3,M3,E1,P2,X1}	0.54	38.5	78.2	-0.29	91.1
5	{D2,L7,R4,P1,E3}	**0.47**	34.0	74.8	-0.39	**85.7**
3	{D2,L7,E3}	0.46	34.4	75.5	-0.40	82.7
1	Likelihood ratio (K3)	**0.51**	**36.7**	**75.1**	-0.32	**94.4**
1	Phoneme recognition (X1)	0.44	35.3	71.5	-0.43	86.6
1	Phoneme duration score (D2)	0.42	32.3	70.2	-0.43	87.8
1	Duration-normalized lh. (E3)	0.40	34.5	69.1	-0.49	69.9
1	Phoneme sequence lh. (M3)	0.38	31.5	66.7	-0.59	78.3

are used individually, the performance decreases especially w.r.t. COR and CL-A for non-native data and RR for native data (cf. Table 8.7).

Depending on the optimization criterion, different feature combinations are found by the floating search (FS) algorithm. The search is carried out with all features. The best combinations with up to five features are shown in Table 8.6. For the combinations it is characteristic that at most one feature from each feature group (L,K,E,M,X,P,Y,R) is present. Furthermore, the features of the groups K,M,X,E seem to carry most of the relevant information.

Cross-validation was carried out for feature combinations from Table 8.6. If only the utterance segmentations are available, the highest accuracy is achieved with the feature combination {D2,L7,R4,P1,E3}. If also any of the X, K and M features are available, which depend on the utterance recognition result, the correlation increases to 0.54 for the feature set {E3,K2,X1}. At the same time the scoring accuracy of native utterances is 90.2%.

Conclusion. In terms of the correlation, the performance of *soft scoring* with a *linear regression* function of the pronunciation features (0.59) is almost as high as the average inter-rater open correlation between the human evaluators (0.60). Nevertheless, the output of the regression function is distorted and the scoring accuracy is only 7.4 % for native speech. By adjusting scores with a linear and a multiplicative polynomial transformation, the performance increases to 85.4% (cf. Table 8.5).

Regarding correlation, the *Gaussian classifier* performs worse (0.54) than linear regression. However, the scoring accuracy can be higher. This is evident from Table 8.5, which shows a recognition rate of 90.2% vs. 85.4% for native utterances. This result suggests, that the correlation coefficient alone is not a reliable indicator of utterance scoring performance. The classification gain $f(\boldsymbol{M}_2, \boldsymbol{P})$ may be used additionally.

The lower performance with linear regression than with the Gaussian classifier may be due to the following reasons:

Table 8.8: *Hard scoring based on the Gaussian classifier and the feature combination (*).*

Confusion matrix for evaluation with Part. C (in %)						
↓ Ref.	1	2	3	4	5	$h \pm 1$
1	**69.7**	*13.0*	12.4	4.9	0.0	**82.7**
2	*46.9*	**25.2**	*20.7*	6.0	1.3	**92.7**
3	24.7	*21.8*	**38.0**	*14.3*	1.3	**74.0**
4	4.5	4.3	*30.3*	**43.3**	*17.7*	**91.2**
5	0.0	0.0	13.2	*32.2*	**54.6**	**86.8**

1. The training data is not balanced w.r.t. the human ratings. Most utterances are labeled with rating 2 or 3 and only very few utterances with rating 1 or 5 exist.

2. A linear relationship between pronunciation features and ratings was assumed. However, this assumption is not true for many scores as the plots in Appendix C show. Hence, the Gaussian classifier is employed in pending experiments for speaker scoring.

8.1.2 Speaker Scoring

Table 8.9 shows the experimental results for speaker level scoring. Evaluation is carried out with 4-fold cross-validation. The calculation of the correlation (COR) and its 95% confidence interval is based on 1000 bootstrap samples. The final speaker level score is obtained by taking all available 48 utterances per speaker into account.

There is no significant difference in performance regarding to whether first score utterances and then average the scores and vice versa. Feature set {E3,K2,X1}, which was most effective for utterance scoring, seems to be superior to {K3,M3}, but the difference is not significant. In Section 5.1, the average inter-rater open correlation at the speaker level was found to be 0.94. Since the performance of the automatic procedure is only 0.84, the human evaluators have to be considered as more reliable.

Figure 8.2 shows the plot of reference ratings versus automatic scores to get an impression of scoring accuracy. Furthermore the relationship between the number of utterances employed to calculate speaker level scores and each speaker's ratings is shown. There is a steady increase

Table 8.9: *Speaker level scoring based on 48 utterances per speaker (ATR SLT data).*

Features	Method	Mean COR	95% Conf-Int
{K3,M3}	first score, then average	0.79	[0.72;0.85]
	first average, then score	0.80	[0.73;0.86]
{E3,K2,X1}	first score, then average	0.83	[0.77;0.88]
	first average, then score	0.84	[0.78;0.88]

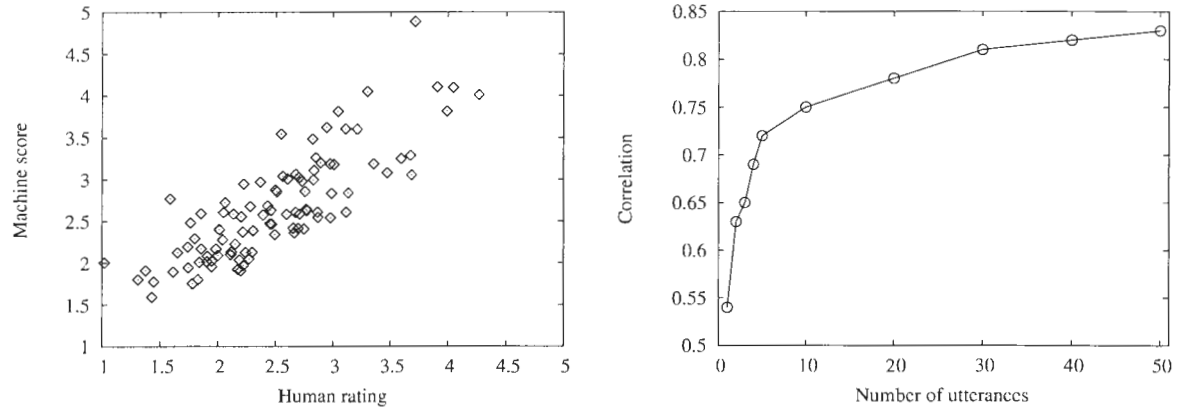


Figure 8.2: **Left:** *Speaker level scoring based on the Gaussian classifier and the "first score, then average" method with the feature set {E3,K2,X1}* - **Right:** *Relationship between the number of utterances employed for scoring and the rating to score correlation (ATR SLT data).*

in correlation with the number of utterances. Most is gained for the first ten utterances and the curve's slope indicates that further increase can be expected for using more than 50 utterances.

8.1.3 Word Classification

Three different classification approaches are examined: Floating search (FS) with a single density Gaussian classifier, decision tree building with CART and training of GMMs in a feature space reduced by LDA or PCA. The gain matrix $\boldsymbol{M}_3$ employed for FS is defined as

$$\boldsymbol{M}_3 = \begin{bmatrix} +1 & -3 \\ -1 & +1 \end{bmatrix} \tag{8.3}$$

With this matrix, the negative effect of wrong classifications of tokens belonging to class $\mathcal{O}$ as $\mathcal{X}$ is three times higher than classifying mispronounced tokens as correct. The reason for this definition is, that misclassifications of the former kind are much more disadvantageous for the second language learner than misclassifications of the latter kind (cf. Section 1.3).

Decision trees are trained stepwise by greedy selection of features. The performance of three pre-selected feature sets is examined for stepwise training: (a) all 35 features, (b) 33 features, i.e. without W05 and W07, which are based on prior information, (c) 21 features, i.e. without {W05,W07} and {W13-W16,C01-C08} which are based on the word recognition output.

GMMs are trained for each feature dimension $d \in \{1, 2, 3, 4, 5, 10, 15\}$. A maximum number of $d * 2$ densities was allowed for training with the FJ algorithm.

For evaluation two kinds of reference labels for mispronounced words are examined:

1. Words from $\mathcal{R}$ belong to $\mathcal{X}$, i.e. a word is considered as mispronounced, if it was marked by one or more evaluators

2. Words from $\mathcal{R}$ belong to $\mathcal{O}$, i.e. a word is treated as mispronounced if it is marked by at least two evaluators.

Table 8.10: *Word classification with class $\mathcal{R}$ belonging to $\mathcal{X}$, i.e. reference word is considered as mispronounced if it was marked by one or more evaluators. Part C (ATR SLT data).*

Classifier	#Fts	Feature selection	CL / $f(\boldsymbol{M}_3, \boldsymbol{P})$
Single Gaussian (FS,CL)	1	{C02}	65.4
	2	{C02,C08}	66.4
	3	{C02,C08,W08}	66.9
	4	{C02,C08,W10,C07}	67.5
	5	{C02,C08,W10,C07,W06}	67.9
	6	{C02,C08,W10,C07,W06,C03}	**68.4**
Single Gaussian (FS,$\boldsymbol{M}_3$)	1	{W12}	60.1 / 0.098
	2	{W12,C01}	63.6 / 0.133
	3	{W12,C01,C07}	64.9 / 0.146
	4	{W12,C01,C07,W17}	66.2 / 0.152
	5	{W12,C01,C07,W10,W01}	65.3 / 0.170
	6	{W12,C01,C07,W10,W01,W16}	65.9 / **0.176**
CART[STEP]	(a) 35 $\rightarrow$ 3	{W05,C07,C06}	**67.8**
	(b) 33 $\rightarrow$ 5	{C08,W03,C04,W22,C06}	66.5
	(c) 21 $\rightarrow$ 3	{W03,W12,W18}	63.6

The results for both categorizations are shown in Tables 8.10 and 8.11, respectively. The maximum accuracy observed for (2) is 73.5 %, which is higher than the maximum performance of 68.4 % achieved for categorization (1). This result backs the hypothesis, that the boundary between correctly pronounced and mispronounced words is unclear and the decision of the human evaluators to mark words is highly subjective. If two or more evaluators agree about the minor pronunciation quality of a word token, the determination of mispronounced words becomes more consistent. Consequently, the higher classification accuracy for (2) is reasonable, since classifiers are trained on data with more reliable and more valid reference labels.

The single density Gaussian classifier based on a feature set selected by the floating search algorithm yields the highest performance. There is no significant difference in accuracy to GMMs and reduction of the feature space dimension with PCA. The performance is significantly lower, if LDA is used for feature reduction. In Table 8.11 only the result of the GMM/LDA and the GMM/PCA approach for feature space dimension d with highest accuracy are shown. Since the computational complexity of the GMM/PCA approach is higher, the single Gaussian approach should be favored. The classification accuracy with a decision tree is best, if it is constructed stepwise from the full feature set (a). Performance decreases when using the feature subset (b) or (c).

Important features, which are selected by the floating search algorithm or in the stepwise CART training procedure are the confidence features based on the phoneme confusion matrix {C01,C02}, the confidence measures {C07,C08} based on the word posterior probability, duration features W10-W12, word likelihood features W01-W03 and the recognition performance W13. Features which do not occur in features sets with six or less features

Table 8.11: *Word classification with class $\mathcal{R}$ belonging to $\mathcal{O}$, i.e. reference word is considered as mispronounced, if it was marked by two or more evaluators. Part C (ATR SLT data).*

Classifier	#Fts	Feature selection / transformation	CL / $f(\boldsymbol{M}_3, \boldsymbol{P})$
Single Gaussian (FS,CL)	1	{C01}	67.6
	2	{W01,C08}	70.9
	3	{W01,C08,W11}	72.1
	4	{W01,C08,W11,C01}	72.8
	5	{W01,C08,W11,C01,C05}	**73.5**
Single Gaussian (FS,$\boldsymbol{M}_3$)	1	{W10}	63.7 / 0.156
	2	{W10,C02}	68.1 / 0.202
	3	{W10,W21,C01}	70.0 / 0.229
	4	{W10,W21,W20,C01}	70.4 / 0.233
	5	{W10,C01,W11,W02,W13}	71.0 / 0.233
	6	{W10,C01,W11,W02,W13,W21}	71.4 / **0.236**
CART[STEP]	(a) 35 $\rightarrow$ 2	{W05,C08}	**73.0**
	(b) 33 $\rightarrow$ 5	{W13,W06,C08,C01,W12}	71.2
	(c) 21 $\rightarrow$ 4	{W03,W12,W19,W08}	67.3
GMM	10	Linear Discriminant Analysis (LDA)	68.9
	4	Principal Component Analysis (PCA)	**72.6**

selected by the floating search algorithm nor selected during stepwise decision tree learning are {W09,W14,W15,W23-W27}. An interesting observation is the selection of the features W05 and W06, the expected and the actual word duration. From the mispronunciation index in Appendix A and the distribution plots of W05 and W06 in Appendix C it is clear, that the longer the duration of a word or the larger the number of phones in a word is, the more likely is its mispronunciation.

As for utterance scoring, the word classifier has to be validated with native speech data. All words of the native data are assumed to be pronounced correctly. The performance for native speech is evaluated for all single features and all feature combinations found by the floating search algorithm. The best twenty feature combinations determined for native speech are then evaluated again for non-native speech with cross-validation. Table 8.12 shows a selection of the best results for both native and non-native speech. Since it is still difficult to determine the best feature combination among those listed in the table, final comparison is made by looking at the trade-off between precision and recall.

Figure 8.3 shows the recall $\mathcal{O}$ vs. recall $\mathcal{X}$ and the recall $\mathcal{X}$ vs. precision $\mathcal{X}$ curves for part $\mathcal{C}$. For better visibility, the curves are only plotted for four feature combinations. Performance for native speech is highest with feature W12. However, the precision for the detection of mispronounced words for non-native speech is too low. The feature set {W01,C08,W11,C01,C05} seems to give the highest performance for non-native speech, although RR for native speech decreases to 90.3 %.

Table 8.12: *Results for cross-validation with non-native data and results for validation with native data (words in the TIMIT SX sentences of 7 native speakers). $\mathcal{R}$ belongs to $\mathcal{O}$.*

Classifier	Features	Native RR	Non-Native (Cross-Vali) CL	REC $\mathcal{O}$	REC $\mathcal{X}$
Gaussian	W12	98.8	61.2	92.1	30.3
	W10	96.2	64.0	88.0	40.1
	W12,C01,C07,W10,W01	95.9	69.0	86.5	51.5
	W10,C01,W11,W02,W13	93.1	69.5	81.4	57.5
	W01,C08,W11,C01,C05	90.3	72.2	73.7	70.8

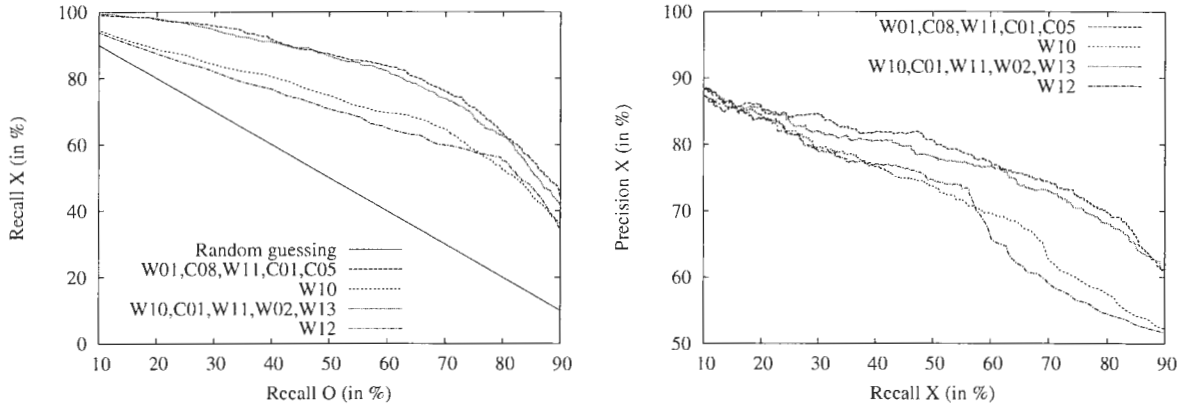


Figure 8.3: *Performance curves for word classification experiments with Part. C (ATR SLT data).*

It remains to analyze, whether the accuracy of automatic word classification can be considered as "acceptable" in comparison to the human evaluators. Their performance is shown in Table 8.13. The confusion matrix "pairwise average" is obtained with the following procedure: select two evaluators; use the 1st evaluator's labels as reference and the 2nd evaluator's labels as classification result and obtain the confusion matrix; repeat calculation for all evaluator pairs and take the average result. The confusion matrix "open average" is obtained in the following way: determine a reference from the labels of all but one evaluator: a token belongs to class $\mathcal{X}$ if it is marked twice or more, otherwise it belongs to class $\mathcal{O}$; take labels of the remaining evaluator as classification result and obtain the confusion matrix; repeat this procedure for each possible combination and take the average result.

Disagreement among evaluators is apparent, since 8.1 % of the presumably correct tokens

Table 8.13: *Performance of human evaluators.* **Left:** *pairwise average* **Right:** *open average.*

Humans	$\mathcal{O}$	$\mathcal{X}$
Class $\mathcal{O}$	88.8	11.2
Class $\mathcal{X}$	42.3	57.7

Humans	$\mathcal{O}$	$\mathcal{X}$
Class $\mathcal{O}$	91.9	8.1
Class $\mathcal{X}$	42.2	57.8

are marked as mispronounced w.r.t. the open average. When employing the feature combination {W10,C01,W11,W02,W13} the classification error w.r.t. class $\mathcal{O}$ is 18.6 %, which is more than two times higher. At the same time, the detection accuracy of mispronounced words is equal to the human evaluators. The classification error w.r.t. class $\mathcal{O}$ can be reduced by shifting the decision boundary. This inevitably leads to a lower recall, but also a higher precision for class $\mathcal{X}$. Though the automatic method cannot achieve the same accuracy as the human evaluators, performance is promising.

The result for three-way classification, i.e. discrimination of the classes $\mathcal{O}$, $\mathcal{X}$ and $\mathcal{R}$, is shown in Table 8.14. The feature combination {C02,C07,W12,W10,W03} is found by the floating search algorithm. As for two-way classification the same important features are selected. The gain matrix $\boldsymbol{M}_4$ employed for FS is:

$$\boldsymbol{M}_4 = \begin{bmatrix} +4 & +1 & -4 \\ 0 & +1 & 0 \\ -2 & -1 & +2 \end{bmatrix} \tag{8.4}$$

For its definition, the same considerations as for matrix $\boldsymbol{M}_3$ apply. Furthermore it does not matter if class $\mathcal{R}$ is classified as $\mathcal{O}$ or $\mathcal{X}$, but it should also have a positive effect, if it is identified correctly (2nd row). Furthermore tokens belonging to class $\mathcal{X}$ should not be classified as $\mathcal{R}$, but for tokens from class $\mathcal{O}$ it should have a positive effect (2nd column).

With the usage of the rejection class $\mathcal{R}$, two possibilities of how to process the classification result arise:

1. Tokens classified as $\mathcal{O}$ may be highlighted in green (pronunciation correct), $\mathcal{R}$ in yellow (indefinite) and $\mathcal{X}$ in red (mispronounced). Such a color coding scheme serves as feedback to the language learner like it is employed in existing CAPT systems (cf. Section 1.3).

2. Treat tokens classified to $\mathcal{R}$ as correctly pronounced. This leads to an increase of the recall of class $\mathcal{O}$. Furthermore, consider reference tokens of class $\mathcal{R}$ as mispronounced. This is an acceptable assumption, since such tokens have been marked as mispronounced by at least one human evaluator.

With a recombination of classification results after (2), the confusion matrix in the right half of Table 8.14 is obtained. The misclassification rate of tokens from class $\mathcal{O}$ is only 6.9%,

Table 8.14: **Left:** *Result of word classification for all three classes $\mathcal{O}, \mathcal{R}, \mathcal{X}$ with a Gaussian classifier and the features set {C02,C07,W12,W10,W03}. Evaluation is carried out with cross-validation.* **Right:** *Recombination of results for two classes: treat reference labels of $\mathcal{R}$ as $\mathcal{X}$ and any classification outcome of $\mathcal{R}$ as $\mathcal{O}$ (ATR SLT data).*

Machine	Class $\mathcal{O}$	Class $\mathcal{R}$	Class $\mathcal{X}$
Class $\mathcal{O}$	85.3	7.8	6.9
Class $\mathcal{R}$	65.7	14.6	19.7
Class $\mathcal{X}$	41.0	16.2	42.8

Machine	$\mathcal{O}$	$\mathcal{X}$
Class $\mathcal{O}$	93.1	6.9
Class $\mathcal{X}$	68.7	31.3

lower than the human evaluators. At the same time 31.3% of the mispronounced words are still detected. The classification accuracy of mispronounced tokens belonging to class $\mathcal{X}$ is even 42.8% (left table).

8.2 FAU LME Data

The corpus contains non-native English speech of German children. Children speech has different acoustic characteristics than adult speech, since the vocal tract of children up to an age of about 12 years is shorter than adults on average. Furthermore, German school teachers usually tend to speak British rather than American English. Since the acoustic model is the most crucial part of the pronunciation scoring system, preliminary recognition experiments are carried out to cope with the two possible mismatch conditions AE/BE and child/adult. A unigram language model covering 970 words estimated on the transcriptions of the corpus is employed for word recognition. Phoneme recognition was conducted without constraints.

Table 8.15 shows the recognition results for three different acoustic models. The AE model is trained with American English adult speech from the WSJ0/WSJ1 corpus, the BE model with British English adult speech from the WSJCAM0 corpus and the BE_PF_STAR model with British English children speech from the PF_STAR corpus. Additionally, normalization of acoustic features by vocal tract length normalization (VTLN) [LR98] is carried out for recognition with the AE and BE models. Performance is examined for six different warping factors $\alpha \in \{0.80, 0.85, 0.90, 0.95, 1.00, 1.05\}$. The factor $\alpha = 1.0$ corresponds to recognition without VTLN.

Performance is worst for the AE and BE models without VTLN. The word accuracy is almost equal for recognition with the AE model plus VTLN and the BE_PF_STAR model. Word accuracy with the AE model becomes highest, if the optimal VTLN normalization factor (α^*) for each speaker is employed. Consequently, for pronunciation feature extraction the AE model and acoustic feature normalization with the optimal α^* is used.

For comparison with previous results for SLT data, the correlation between the pronunciation features and the human ratings is analyzed. Since only speaker level ratings are available, correlation can only be examined at the speaker level. Speaker level features are obtained by

Table 8.15: *Mean recognition performance for all non-native children speech data in the FAU LME database with one acoustic model each for American (AE) and British English (BE) plus optional VTLN [α], and with one acoustic model BE children speech (BE_PF_STAR).*

Acoustic model	WordAcc	Acoustic Model	WordAcc	PhonAcc
AE/BE + VTLN [0.80]	20.2 / 18.3	AE + NO VTLN	22.1	-16.3
AE/BE + VTLN [0.85]	25.0 / 22.7	AE + VTLN [α^*]	**27.4**	-11.3
AE/BE + VTLN [0.90]	25.9 / 22.8	BE + NO VTLN	18.6	-16.3
AE/BE + VTLN [0.95]	25.3 / 21.5	BE + VTLN [α^*]	24.3	-9.5
AE/BE + VTLN [1.05]	17.8 / 15.2	BE_PF_STAR	25.0	**1.8**

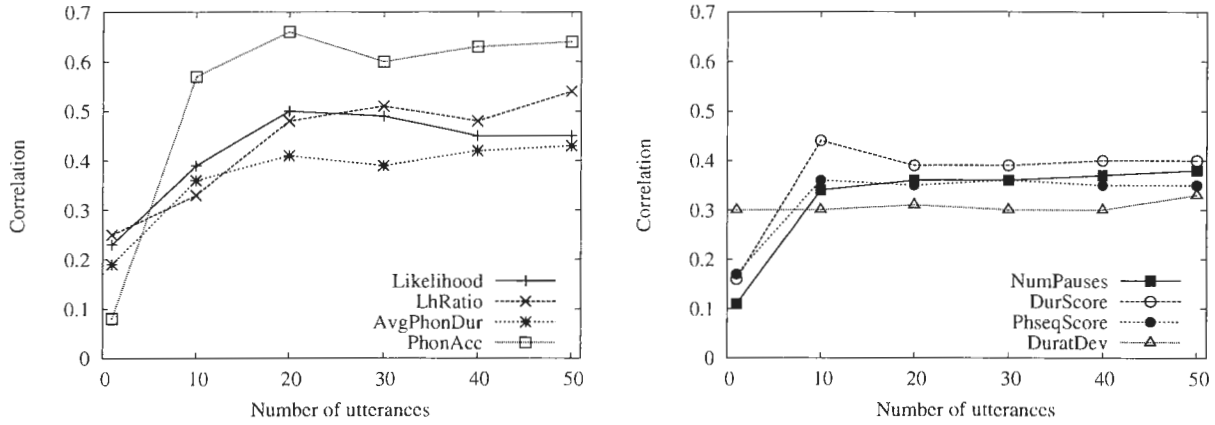


Figure 8.4: *Relationship between the rating to score correlation and the number of utterances used for averaging in order to obtain speaker-level ratings and scores (FAU LME data). The scores considered are: Likelihood (L7), LhRatio (K1), AvgPhonDur (R4), PhonAccu (X1), DurScore (D2), NumPauses (P2), PhseqScore (M3) and DuratDev (E1).*

Table 8.16: *Correlation between speaker level ratings and scores.*

Correlation	L7	K1	R4	P2	D2	X1	M3	E1
FAU LME data	-0.45	-0.56	+0.43	+0.38	-0.41	-0.66	-0.35	+0.32
ATR SLT data	-0.58	+0.80	+0.55	+0.55	-0.72	-0.76	-0.60	+0.61

averaging the feature values of several utterances. Figure 8.4 shows the relationship between the number of utterances used for averaging and the speaker level correlation. There is an increase in correlation with the number of utterances. While the same tendency can be observed for LME and SLT data (cf. also Figure 6.8), the maximum correlation reached is lower for LME data than for SLT data (cf. Table 8.16). One reason may be the fact, that the speaker level ratings of the LME data are only based on one assessment of one evaluator, and the ratings are only of the from a or a-b, where $a, b \in \{1, 2, 3, 4, 5\}$. The speaker level ratings for the SLT data are more nuanced because they are based on multiple assessments by several evaluators for 48 phonetically rich sentences. Feature X1 seems to be most reliable for both corpora. The features K1, L7, R4 and D2 have a correlation greater than 0.4.

The complete information about each non-native speaker including the number of utterances and words, share of mispronounced words, recognition performance, optimal α^* a.s.o. is given in Appendix A. Table 8.17 shows the absolute number and relative share of words considered as correctly pronounced or as mispronounced for all speakers in the database.

In the following experiments, LME data are only used as test data. The pronunciation scoring system is trained with non-native speech from the SLT database and native speech data from the WSJ and TIMIT corpora.

Table 8.17: *Number of words considered as correctly pronounced and as mispronounced in the human annotations (FAU LME data).*

Classes	# words	share (%)
Correct $\mathcal{O}$	17605	91.4
Mispronounced $\mathcal{X}$	1657	8.6

8.2.1 Speaker Scoring

Since utterance level ratings are not available for LME data, only experiments for speaker level scoring are carried out. Besides the feature combinations that were effective for SLT data, additional combinations of the five features X1, K1, L7, R4 and D2 with a relatively high correlation to human ratings as a shown in Table 8.16 are considered. Classification with a single Gaussian per rating class as well as linear combination of features is examined for scoring single utterances. The final speaker level score is obtained by averaging two or more utterances.

Table 8.18 shows the scoring performance for the feature combinations among the best results. All utterances available were taken into account to calculate the score for each speaker. The overall performance of all examined feature sets for LME data is significantly lower than for SLT data. The set {E3,K2,X1}, which is most suitable for SLT data with a correlation of 0.84, has only a correlation of 0.52 for LME data. Scoring becomes more reliable with different feature combinations. Maximum correlation is present when employing the feature set {X1,K1,L7}. There is no significant difference in scoring reliability if employing the Gaussian classifier and linear combination for this feature set.

The speaker level correlation for linear combination of features {X1,K1,L7} is 0.68. Score adjustment was applied. The parameters of the regression function and the score transformation are only estimated on SLT data. Figure 8.5 illustrates the relationship between human reference ratings and the automatic scoring result. As for the SLT data, speaker level correlation increases, if more utterances are taken into account for scoring. Since many utterances are single word utterances, they were reordered randomly before determining the right graph of Figure 8.5. The

Table 8.18: *Speaker level scoring with the Gaussian classifier and linear feature combination considering all utterances available for each speaker (FAU LME data).*

Features	Method	Mean COR	95% Conf-Int
{D2,L7,R4,P1,E3}		0.45	[0.19;0.65]
{E3,K2,X1}	Single	0.52	[0.29;0.69]
{X1}	Gaussian	0.65	[0.47;0.78]
{X1,K1,L7}		0.66	[0.52;0.78]
{D2,L7}		0.62	[0.48;0.73]
{K3,X1,M3}	Linear	0.66	[0.52;0.78]
{X1}	Combination	0.64	[0.46;0.78]
{X1,K1,L7}		0.68	[0.53;0.80]

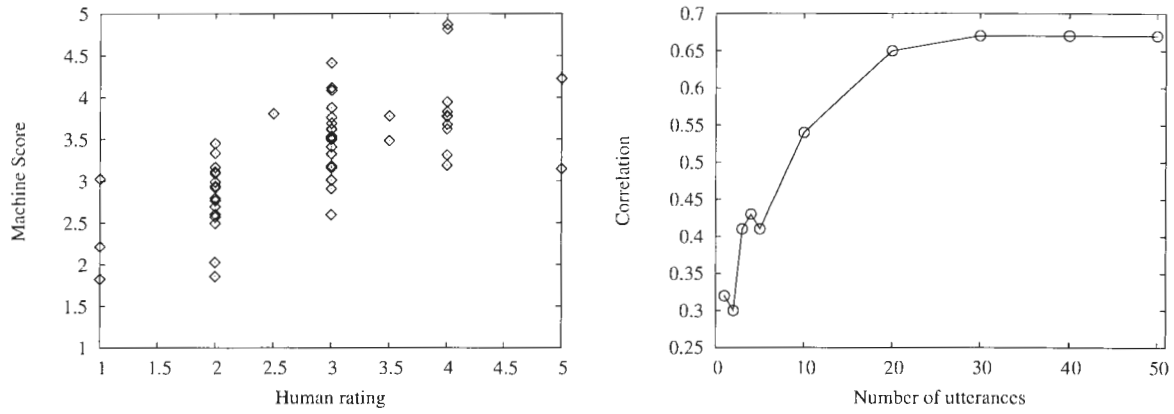


Figure 8.5: **Left:** *Speaker scoring with the feature set {X1,K1,L7} and a linear regression function.* - **Right:** *Relationship between the number of utterances employed for speaker scoring and the correlation coefficient between human ratings and scores (FAU LME data).*

correlation reaches its maximum after 30 utterances. Nevertheless, the highest increase can already be observed for the first 10-20 utterances.

Despite the evaluation of the SLT and LME data are lacking a common basis, speaker scoring performance is promising. The speaker level correlation is lower for LME than for SLT data, but this may also be to the fact, that the speaker ratings for LME data are less nuanced. Furthermore, the LME data does not consist of phonetically balanced sentences, which would allow a more accurate estimate of each speaker's pronunciation skill.

8.2.2 Word Classification

Word classification for LME data is carried out with the feature combinations and classifiers with best performance for SLT data. Classifiers are only trained with SLT data. Additionally, the floating search algorithm is employed for selection of features, which may be especially suitable for the LME corpus. Class $\mathcal{X}$ consists of all words which were marked as incorrectly pronounced, substitutions and garbage words, i.e. non-English words and abortions, class $\mathcal{O}$ of all other words.

Two out of three combinations of five features approved for SLT data also seem to have an equally high discrimination ability for LME data (cf. Table 8.19). The floating search algorithm finds the features {W12,C08,W06,C03,W13} as best combination of five features by optimization with the LME data, but performance is only slightly higher than for the approved feature sets. While good results are achieved with a single density Gaussian classifier, the approaches based on GMM/PCA or CART lead to worse results.

The performance curves (precision and recall) in Figure 8.6 show, that the best result (71.6%) is obtained with feature combination {W12,C01,C07,W10,W01}. The accuracy was 70.4% for the feature set {W10,C01,W11,W02,W13} which was also effective for SLT data. It can be concluded, that the duration-related features {W10,W11,W12}, the confidence measure {C01,C07}, the word likelihood features {W01,W02} and the recognition performance feature

W13 are a good choice for discrimination of correctly pronounced and mispronounced words.

Table 8.19: *Word classification with LME data. The results of feature combinations found effective for SLT data and those selected by the floating search algorithm are shown.*

Classifier	#Fts	Features	CL / $f(\boldsymbol{M}_3, \boldsymbol{P})$
Gaussian	5	{W01,C08,W11,C01,C05}	68.4 / 0.210
	5	{W12,C01,C07,W10,W01}	71.6 / 0.292
	5	{W10,C01,W11,W02,W13}	70.4 / 0.314
Single Gaussian (FS,$\boldsymbol{M}_3$)	1	{W12}	66.5 / 0.237
	2	{W12,C08}	69.1 / 0.280
	3	{W12,C08,W06}	70.4 / 0.303
	4	{W12,C08,W06,C03}	70.9 / 0.316
	5	{W12,C08,W06,C03,W13}	71.7 / 0.322
CART	2	{W05,C08}	61.4
	5	{W13,W06,C08,C01,W12}	66.3
	4	{W03,W12,W19,W08}	63.9
GMM	2	33 → PCA → 2	67.8

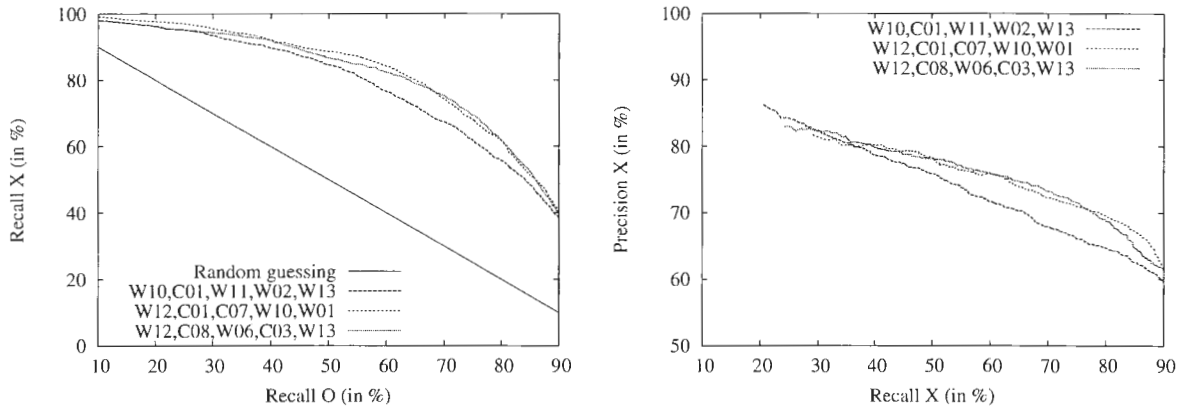


Figure 8.6: *Performance curves for word classification experiments (FAU LME data).*

8.3 PF_STAR BE: Utterance Scoring

Utterance scoring worked quite good for both non-native and native speech of adults (cf. Section 8.1.1). Whether this also applies to speech of native children is the investigation target of this section. The native data is taken from the PF_STAR BE corpus: 11527 utterances from 92 native children speakers. Utterance scoring is carried out with the Gaussian classifier trained on SLT data. The AE acoustic model, acoustic feature normalization by VTLN with optimal α^*

Table 8.20: *Results for scoring utterances of native children in the PF_STAR BE corpus with the Gaussian classifier. A reference rating of 1 is assumed for each utterance.*

Features	1	2	3	4	5
{X1}	56.6	6.7	2.6	4.0	30.1
{K1}	54.6	5.8	6.0	11.0	22.6
{X1,K1}	48.8	6.8	4.6	24.7	15.1
{K3,M3,X2}	52.8	10.7	8.1	4.0	24.4
{E3,K2,X1}	28.7	14.1	12.9	17.4	27.0

Table 8.21: *Result for scoring utterances of native children in the PF_STAR BE corpus with the Gaussian classifier based on feature X1. The younger the child, the more probable is the worst scoring result of "5".*

Age	#spk	#utr	1	2	3	4	5
13,14	6	536	76.3	5.0	1.7	1.5	15.5
11,12	10	1,114	62.7	6.0	1.3	4.4	25.8
9,10	41	5,137	62.0	5.6	2.6	4.0	25.9
7,8	28	3,816	51.5	7.6	3.0	3.5	34.4
4,5,6	7	924	29.2	10.4	2.8	7.5	50.1

and a bigram language model trained on the transcriptions of the test data were employed for recognition.

Table 8.20 shows the utterance scoring result. The single features {K1,M3,D2,L7,R4,X1} and feature combinations effective in previous experiments were examined. A reference rating of 1 is assumed for each utterance. The best results of 56.6% and 54.6% are achieved with the single features X1 (phoneme recognition rate) and K1 (likelihood ratio), respectively. All other considered feature combinations perform worse. Remarkably many utterances are scored as "5", i.e. they seem have a very low pronunciation quality.

To get more insight into possible reasons, utterance scoring is carried out separately for five age groups. Table 8.21 shows the corresponding result. It is apparent, that the younger the speaker, the more likely is the worst case scoring result of "5". For example, while only 29.2% of the children's utterances of four to six years of age are scored as "1" and 50.1% as "5", only 15.5% of the childrens' utterances of thirteen and fourteen years of age are scored as "5", and 76.3% are classified as "1". Furthermore, most utterances are either scored as "1" or "5" and rather very few utterances as "2", "3" or "4". A possible interpretation may be, that the younger children do not utter the utterance prompts smoothly enough. However, it cannot be assumed that the children make severe mistakes w.r.t. the English pronunciation of words, since they are native speakers.

Chapter 9

Outlook and Future Work

This chapter presents several ideas of the author, how pronunciation features, scoring and classification methods can be optimized further. Moreover, two possible applications of the pronunciation features and the results of this thesis are mentioned.

Features. For the word and phoneme recognition performance features it was shown, that their correlation with human ratings is high for the speakers in the SLT corpus. The result was less good for the FAU LME data. An obvious drawback of the performance features is, that their values do not only vary with the pronunciation skill of the speaker, but also with the difficulty of the recognition task and mismatched conditions, e.g. child/adult speech. Furthermore, the feature is dependent on whether a zerogram or higher-order n-gram language model is employed. To avoid too much recognition errors, the usage of a unigram or bigram language model may be recommendable, but it has also the negative effect of distorting the relationship of the recognition performance to the pronunciation skill. Consequently, the performance features have to be normalized, e.g. by adding a constant, so that a non-native speaker with high pronunciation skill is scored as high as a native speaker. Normalization is especially important, if the scoring system is trained on data different from the application.

The necessity of normalization also applies to the likelihood based features. Likelihoods are dependent on acoustic conditions and speaker characteristics. A possible approach to likelihood normalization is the subtraction of the likelihood $P(x|\text{GMM})$ for each speech frame x given a GMM which is trained on non-pause and non-silence native speech data. This approach is similar to the acoustic normalization as employed by [SSN+02] for confidence measures.

The fluctuation features W23-W27 turned out to have almost no discriminative ability as the feature distribution plots in Appendix C show. Let $d^{(pau)}$ be the duration of the silence segment following a word. Extended duration ratios may then be defined as

$$W28 = \frac{d_{j-1} + d_{j-1}^{(pau)} + d_j + d_j^{(pau)}}{d_{j-1}^{(exp)} + d_j^{(exp)}} \qquad W29 = \frac{d_{j-1} + d_j}{d_{j-1}^{(exp)} + d_j^{(exp)}} \tag{9.1}$$

From the feature distribution plots in Appendix C it is apparent, that features W28 and W29 have better discriminative abilities than W22-W27. The consideration of an even wider context may improve the quality of these new features.

Classification. An alternative to direct classification of words with one feature vector is the application of the GOP algorithm (cf. Section 3.4.1), which was designed for the detection of mispronounced phonemes, to words. If the mispronunciation event of a word is defined as, e.g. $H\%$ of its phonemes are mispronounced, a new approach for detection of mispronounced words evolves: Classify each phoneme of each word with a thresholding method as the GOP algorithm. If the relative number of mispronounced phonemes is above $H\%$, consider the word as mispronounced. For a practical algorithm, phoneme dependent mispronunciation thresholds and the threshold $H\%$ have to be determined. Besides the thresholds proposed by Witt et al. [WY00], the following global threshold $\tau^{(c)}$ and individual thresholds $\tau_j^{(c)}$ may be worth examination:

$$\tau_j^{(c)} = \log \frac{p_j^f}{\sum_{i=1}^{M} p_i^f} \qquad \tau^{(c)} = \frac{1}{M} \sum_{j=1}^{M} \tau_j^{(c)} \tag{9.2}$$

p_j^f are the mispronunciation probabilities of phonemes which can be estimated with the word mispronunciation model B proposed in Section 5.3.

There is a polynomial relationship between the utterance rating and the number of marked words in the same utterance. A post-processing of word classifier scores considering the utterance score may improve accuracy. Alternatively, the number of words marked automatically could be used as feature for utterance scoring. However, a proper investigation remains.

Scoring. For utterance soft scoring (cf. Section 8.1.1) based on linear regression the scoring result was skewed. This was also due to the uneven distribution of samples for each pronunciation rating class. For the robust training of a pronunciation scoring system the availability of non-native speech data of speakers with a very high and a very low pronunciation skill is important. A further reason for the distorted scoring result is the assumption of a linear relationship between pronunciation scores and ratings. The feature plots in Appendix C show, that this is not the case for many pronunciation scores. The relationship could be linearized by applying certain transformations to these features.

Applications. The pronunciation features employed in this thesis are intended to measure the degree of non-nativeness. Consequently, they may also suitable for the discrimination of native and non-native speakers in general.

From the analysis of rating to score correlation carried out in this thesis and previous research, the following differences between native and non-native read speech are apparent: There are different segmental characteristics. The GOP score, which is an approximation of the phoneme posterior probability, can be regarded as a measure of the distance of segmental characteristics between native and non-native pronunciation. Also for the SLT data analyzed in this thesis, the dependency was highest for a GOP based utterance score. Furthermore, the duration characteristics of phonemes differ, because there is a correlation between the duration score and the utterance rating and there are deviations between the actual and expected duration of phonemes and words. Last but not least it must be mentioned, that there may be longer inter-word pauses (hesitations) in non-native speech, which may be due to a speaker's unfamiliarity with certain words. Methods, which are intended to improve the recognition performance for non-native speech will have to cope with these non-native effects.

Chapter 10

Summary

Pronunciation scoring is the automatic assessment of the pronunciation quality of phonemes, words or utterances especially for non-native speakers. A possible application are systems for computer assisted pronunciation training (CAPT) to support the student of a foreign language to acquire correct pronunciation. Such a system is intended to provide also detailed feedback like the localization of mispronounced phonemes to pinpoint mistakes. The necessity for improvement of pronunciation is due to the circumstances, that a non-native pronunciation can hinder inter-human communication and that non-native speech is in general more difficult to recognize than native speech.

There are segmental, temporal and other prosodic aspects of pronunciation. Multiple aspects of quality considered at the same time are referred to as overall pronunciation quality. In order to design a pronunciation scoring scheme, reference data with examples of different degrees of pronunciation quality are required. Such reference data can be obtained by a human evaluation of non-native speech data (cf. Chapter 1).

Most CAPT systems employ speech recognition technology for pre-processing the speech of non-natives. Furthermore, concepts and techniques from pattern recognition, including feature extraction, feature selection or transformation and classification are required. Moreover, to analyze a human evaluation and experimental results, methods from statistics, especially for the estimation of reliability and confidence have to be employed. The necessary fundamentals for building a pronunciation scoring system are introduced (cf. Chapter 2).

In literature examples for human evaluation of pronunciation are available. Studies for assessing the quality of single phonemes and whole utterances are most extensive: Phonemes with wrong pronunciation are marked and the quality of utterances is measured on a discrete scale by human experts. Since such an evaluation is subjective, the same material was evaluated by several experts. It became apparent, that the higher the level of evaluation (e.g. phoneme vs. utterance), the higher the reliability of assigned markings or ratings is. In a study for the evaluation of utterance quality of phonetically rich sentences, the individual aspects fluency, speaking rate and segmental quality were rated separately from overall pronunciation quality. It became clear, that the correlation is highest between the overall quality and the segmental quality (0.9) and lowest between speaking rate and segmental quality (0.6). Since the segmental aspect of pronunciation has the highest correlation to the overall rating, one overall measure for

utterance pronunciation quality may suffice.

Several methods for scoring the pronunciation quality of phonemes and whole utterances have already been proposed in literature. The employment of speech recognition technology is prevalent. The goodness of pronunciation (GOP) algorithm is a thresholding method, which declares single phonemes as mispronounced if their acoustic posterior score is below a global or phoneme-dependent threshold. For automatic assessment of utterances, several kinds of scores were examined. An acoustic sentence posterior score, a phoneme duration score, a syllable timing score and the recognition accuracy were found to be good indicators of pronunciation quality. Unnormalized likelihood scores were less effective. Additionally it was shown, that the combination of multiple scores improves scoring reliability. The detection of mispronounced phonemes as well as utterance scoring performance was promising. The reliability of the automatic method was comparable to the human evaluation (cf. Chapter 3).

For this thesis two databases of non-native speech were available: The ATR SLT non-native database, which consists of English speech from 96 adult speakers of multiple accent groups, mainly German, French, Chinese, Japanese and Indonesian, and the FAU LME non-native children speech corpus, which comprises English speech of 57 German children. A part of the SLT data, phonetically rich sentences, were annotated by 15 natives with teaching experience at word and utterance level, the LME data by one German student of Anglistics at word and speaker level. For training of the acoustic model of a speech recognizer and models of a pronunciation scoring system, corpora with native speech were required. The Wall Street Journal corpora with American English and British English adult speech and the PF_STAR BE corpus with British English children speech serve for acoustic model training. Phoneme duration statistics were derived from the TIMIT corpus (cf. Chapter 4).

Since the same material of each speaker in the SLT data was evaluated by three to four human evaluators, the reliability of the annotations was analyzed. The inter-rater reliability at utterance and speaker level measured by the average open correlation is 0.6 and 0.9, respectively. These values are comparable to those reported in other studies. Moreover, the confidence of speaker and utterance level reference ratings was examined. The confidence interval for the jackknife mean utterance ratings was ± 0.28 at the 10%, for the mean speaker level ratings ± 0.05 at the 1% error level. In further investigations, it was found out, that there is a high correlation between the number of words marked as mispronounced and the utterance rating. From the relative marking frequency of each word, which can be treated as mispronunciation probabilities of each word, the mispronunciation probabilities of single phonemes were estimated with a newly proposed word mispronunciation model (cf. Chapter 5).

To describe the pronunciation quality of words and utterances, 35 and 33 features are defined, respectively. While employing those utterance level features, which were already shown to be effective in literature, additionally a score based on the probability of recognized phoneme sequences, features based on expected phoneme durations and second order features based on phoneme segment likelihoods and the local rate of speech are examined. Furthermore, the normalization of durations by multiplication with the rate of speech (ROS), which was effective in previous research for the calculation of duration scores, was applied for normalization of sentence scores. The correlation to the human ratings became higher than without ROS normalization. The correlation between a phoneme sequence probability score and human ratings

was almost as high as for likelihood-based scores. The highest correlation (0.5) was present for a sentence score based on the likelihood ratio similar to the sentence posterior score and the GOP score proposed in literature. Scores based on prosodic features were not examined, since in previous research they had only a comparably low correlation of about 0.3 with human ratings. Moreover, the correlation did not increase even after combining prosodic with non-prosodic scores. Finally, in an analysis of correlation between utterance ratings and several pronunciation scores carried out separately for each evaluator no remarkable differences in the rating behavior w.r.t. the individual aspects of pronunciation considered could be observed.

No literature was available for comparison w.r.t. word level features. Besides applying utterance level scores to words, especially word confidence measures were considered: the word posterior probability based on acoustic probabilities, confidence measures based on the phoneme confusion matrix for correctly pronounced and mispronounced words and features measuring fluctuations of the relative and absolute rate of speech are defined. Feature distribution plots for correctly pronounced and mispronounced words revealed, that features based on confidence measures, acoustic score, phoneme sequence probability and phoneme duration probability have a good discrimination ability. The fluctuation features turned out to be ineffective (cf. Chapter 6).

Experiments were carried out for the discrimination of correctly pronounced and mispronounced words, hard and soft scoring of utterances and combination of the scoring result for several utterances of one speaker to obtain an assessment of the overall pronunciation skill. For classification at word level three approaches were examined: The floating search algorithm in connection with a single density Gaussian classifier, decision trees trained with the CART framework, and training of Gaussian mixture models (GMMs) after feature space reduction by PCA or LDA. Since there is a polynomial relationship between utterance ratings and the number of marked words in the same sentence, a post-processing step considering classifier scores and utterance scores may improve the detection accuracy of mispronounced words.

Soft scoring of utterances is carried out for a linear combination of features. The weighting coefficients for each feature are obtained by linear regression. Furthermore a Gaussian classifier with one density per human rating is employed to calculate the expectation of the utterance score. The same Gaussian classifier was also employed for hard scoring, i.e. assignment of a discrete score rather than a continuous value. Since the result of utterance scoring may be skewed, score adjustment with a combination of a linear and a multiplicative polynomial transformation is proposed.

The estimation of the overall pronunciation skill of a non-native speaker is based on utterance scoring. The approaches of averaging the feature vector for multiple utterances followed by a single classification step as well as separate scoring of each utterance and finally taking the average utterance score are examined (cf. Chapter 7).

Experiments were evaluated with cross-validation for the SLT data. Training and data sets are disjoint w.r.t. to the speakers and the human evaluators. Additionally, validation experiments with native speech data were conducted. With linear feature combination the rating to score correlation (0.59) was almost as high as the average inter-rater open correlation (0.60). Despite the automatic scoring being almost as reliable as a human evaluation w.r.t. the correlation, it is lacking accuracy. By adjusting the scoring output, the scoring accuracy could be improved for native (85%) and non-native speech. Even better results for utterances of natives (90%) could

be achieved with hard and soft scoring based on a Gaussian classifier. As alternative measure of utterance scoring accuracy, a measure of classification gain (analogue to classification risk) is recommendable. If only the segmentation and not the recognition result of the utterances to be scored is available, rating to score correlation is 0.53.

An assessment of pronunciation skill at the speaker level is obtained by averaging the pronunciation scores of two or more utterances. The maximum rating to score correlation was 0.84 based on 48 phonetically rich sentences. The range of the 95% confidence interval was at most ± 0.06. Most of correlation (0.75) was gained for the first ten utterances.

The performance of the discrimination of correctly pronounced and mispronounced words was 72% with a single-density Gaussian classifier w.r.t. the class-wise average recognition rate (CL). At the same time 90% of the words uttered by native speakers were classified correctly, assuming a correct pronunciation for all words. Performance was not higher when employing a decision tree, or GMMs and feature space transformation.

Evaluation of the LME data is conducted with a pronunciation scoring system trained with SLT data. For scoring on the speaker level, the correlation was 0.68. A different feature combination than that most effective for SLT data had to be employed to achieve this result. A possible explanation for the lower reliability may be the fact, that the speaker level ratings of the LME data are less nuanced than those of the SLT data, which are obtained by averaging multiple utterance ratings. The word classification accuracy was 70% w.r.t. CL for the same feature set as for the SLT data. The accuracy improved to 72% with a feature set selected by the floating search algorithm by optimization with LME data.

The results for word classification are promising, since a cross-validation analysis of the word markings of the SLT data revealed, that on average 8% of the correctly pronounced words are marked as mispronounced and the detection accuracy of mispronounced word by the human evaluators is 58%.

When scoring utterances of native children in the PF_STAR BE corpus, most utterances are scored with a high pronunciation quality. An interesting result is, that the younger the children speaker, the higher is the share of utterances with a low pronunciation quality (cf. Chapter 8).

Some of the pronunciation features may not only depend on a speaker's pronunciation skill but also on various other factors. Further normalization especially of those features based on acoustic scores and those measuring recognition performance may be necessary. The detection of mispronounced words may improve when employing the GOP algorithm, which is intended to detect mispronounced phonemes. For example, a word may then be considered as mispronounced if a certain number of its phonemes are classified as mispronounced (cf. Chapter 9).

In this thesis several methods for pronunciation scoring on word, utterance and speaker level were examined. Good results are achieved on all levels. The class-wise average recognition rate for discrimination of correctly pronounced and mispronounced is 72% for native adult and children speech. For native adult speech, the recognition rate is even 90%. The reliability of utterance scoring for non-native adult speech is almost as high as the reliability of the human evaluators. At the same time, 90% of the utterances of native adult speakers are scored correctly. For speaker level scoring the reliability is highest with a correlation of 0.84 for adult speech and 0.69 for children speech.

Appendix A

Databases and Corpora

A.1 ATR SLT Non-Native Database

A.1.1 Speaker Information

Information about the average utterance level human rating (HumRating), number of mispronounced words (Miscnt), the ratio of mispronounced words (MisRatio), phoneme recognition accuracy (PA), word recognition accuracy (WA), age and first language (NatLang) of each non-native speaker in the ATR SLT database.

SpkID	HumRating	Miscnt	MisRatio	PA	WA	Age	NatLang
F018	1.03	7	0.02	21.27	-30.38	-	Japanese
M076	1.30	29	0.07	26.48	-11.65	43	German
M036	1.40	21	0.05	22.40	-29.37	30	German
M052	1.43	17	0.04	38.35	4.56	36	German
M078	1.43	18	0.05	25.00	-15.70	35	German
F022	1.60	27	0.07	18.43	-46.58	45	Japanese
M001	1.65	33	0.08	33.99	4.81	39	German
M055	1.67	66	0.17	25.88	-19.24	52	Chinese
M054	1.75	54	0.14	28.88	-11.90	28	German
M040	1.77	41	0.10	11.56	-56.96	21	Chinese
M051	1.77	62	0.16	32.90	0.51	26	German
M071	1.77	25	0.06	16.67	-43.80	39	German
F026	1.80	56	0.14	25.39	-42.78	28	Chinese
M042	1.82	52	0.13	31.59	-12.15	33	Hungarian
M033	1.83	47	0.12	35.26	-11.90	30	Indonesian
M056	1.83	59	0.15	28.95	-32.91	35	German
M014	1.87	58	0.15	12.47	-76.20	35	Japanese
F021	1.90	66	0.17	30.08	-27.85	40	Korean

SpkID	HumRating	Miscnt	MisRatio	PA	WA	Age	NatLang
M044	1.90	71	0.18	17.50	-40.51	25	French
M066	1.93	72	0.18	24.20	-36.96	30	French
M026	1.95	71	0.18	32.33	-25.32	42	French
M072	1.97	62	0.16	18.27	-33.42	26	French
M032	2.00	49	0.12	14.16	-45.32	32	Spanish
F009	2.02	77	0.19	22.52	-44.81	-	Japanese
F025	2.03	41	0.10	22.55	-34.68	35	Portuguese
F012	2.07	72	0.18	2.30	-80.76	-	Japanese
M039	2.07	37	0.09	9.37	-66.84	33	Sinhalese
M061	2.10	58	0.15	26.87	-22.78	25	Indonesian
M073	2.10	37	0.09	28.25	-30.63	27	French
F023	2.12	64	0.16	22.34	-44.81	26	Japanese
F024	2.15	65	0.16	21.20	-42.78	25	French
M045	2.17	64	0.16	30.95	-19.24	23	French
M050	2.17	45	0.11	26.67	-19.75	29	Indonesian
M010	2.20	85	0.22	22.68	-39.75	25	German
M024	2.20	47	0.12	15.45	-69.11	21	French
M034	2.20	86	0.22	26.67	-26.58	23	German
F014	2.23	78	0.20	14.66	-73.16	-	Japanese
M006	2.23	64	0.16	29.73	-8.35	26	German
M043	2.25	74	0.19	20.97	-28.61	24	French
M085	2.25	68	0.17	29.40	-18.48	28	French
F010	2.27	63	0.16	15.21	-53.42	-	Japanese
M021	2.30	72	0.18	23.69	-29.11	26	German
M080	2.30	87	0.22	12.87	-46.84	24	French
M022	2.38	90	0.23	1.50	-75.19	42	Bulgarian
M059	2.38	62	0.16	6.66	-70.13	43	Indonesian
M037	2.43	71	0.18	13.17	-63.04	25	French
F019	2.47	62	0.16	18.45	-48.61	26	Indonesian
M023	2.47	109	0.28	10.35	-69.87	24	French
M030	2.47	82	0.21	12.14	-71.39	35	Chinese
M077	2.47	103	0.26	11.32	-63.54	30	French
M092	2.47	47	0.12	17.76	-44.30	26	German
F008	2.50	83	0.21	18.42	-58.48	-	Japanese
M016	2.53	60	0.15	-4.13	-108.35	37	Japanese
F013	2.57	65	0.16	8.67	-69.87	-	Japanese
M089	2.60	101	0.26	14.88	-45.82	25	Indonesian
M060	2.62	97	0.25	13.59	-66.08	24	Japanese
M035	2.65	121	0.31	19.51	-46.58	23	French

SpkID	HumRating	Miscnt	MisRatio	PA	WA	Age	NatLang
M027	2.67	101	0.26	13.08	-61.01	31	Hindi
M075	2.67	85	0.22	20.99	-39.49	29	Indonesian
M068	2.70	95	0.24	0.41	-75.95	32	Indonesian
M082	2.70	121	0.31	23.84	-36.46	22	Japanese
M031	2.73	96	0.24	8.28	-63.04	38	Indonesian
M058	2.73	88	0.22	2.57	-72.66	39	Japanese
M063	2.73	93	0.24	23.04	-39.75	36	Indonesian
M065	2.73	92	0.23	20.75	-37.22	22	Japanese
M084	2.73	56	0.14	-13.75	-106.08	25	Chinese
M074	2.75	107	0.27	6.45	-67.85	25	Chinese
M011	2.77	78	0.20	6.78	-60.76	41	Japanese
M012	2.83	100	0.25	5.42	-91.90	29	Japanese
M067	2.83	96	0.24	4.20	-66.58	25	Chinese
M087	2.83	88	0.22	11.48	-60.76	24	Indonesian
M046	2.87	74	0.19	17.68	-59.49	31	Indonesian
M053	2.87	79	0.20	7.33	-81.77	40	Chinese
M029	2.88	121	0.31	18.24	-52.66	22	French
M069	2.90	108	0.27	0.34	-86.84	24	Japanese
M070	2.95	125	0.32	-2.98	-77.22	22	Japanese
M041	2.98	124	0.31	16.62	-53.16	30	Indonesian
F011	3.00	132	0.33	12.59	-70.63	-	Japanese
M028	3.00	109	0.28	16.98	-55.95	33	Chinese
M090	3.00	110	0.28	10.26	-52.91	30	Indonesian
M083	3.05	131	0.33	4.68	-86.08	24	Japanese
M038	3.10	108	0.27	15.92	-57.72	30	Indonesian
M015	3.12	100	0.25	8.82	-64.30	29	Japanese
M057	3.15	120	0.30	12.27	-60.25	26	Chinese
M064	3.23	117	0.30	-3.25	-89.37	21	Japanese
M013	3.30	76	0.19	-21.71	-120.51	22	Japanese
F020	3.33	166	0.42	7.67	-84.30	30	Chinese
M049	3.47	132	0.33	-2.99	-104.81	33	Chinese
M086	3.58	124	0.31	8.63	-70.89	28	Indonesian
M062	3.67	148	0.37	8.60	-68.86	31	Chinese
M047	3.70	128	0.32	10.93	-80.76	37	Chinese
M093	3.73	167	0.42	-17.39	-109.37	31	Japanese
M088	3.93	181	0.46	3.45	-96.46	27	Chinese
M025	3.97	216	0.55	1.09	-84.05	37	Chinese
M081	4.07	118	0.30	-28.39	-140.25	28	Chinese
M091	4.27	137	0.35	-4.61	-119.49	31	Chinese

A.1.2 Mispronunciation Index

The following tables show the mispronunciation frequency for all words of the phonetically rich sentences of the TIMIT set. How often any of the annotators marked a word is shown in the third columns. This number divided by the occurency of the word in the sentence prompts and the number of evaluators is shown in the second column.

Word	Normalized	Absolute
EXTRA	1.00	96
EXPOSURE	1.00	96
EXAM	1.00	96
BOX	1.00	96
MIRAGE	0.91	87
CENTRIFUGE	0.85	82
BUGLE	0.85	82
FRANTICALLY	0.84	81
OASIS	0.77	74
PURCHASE	0.75	72
CONTAGIOUS	0.74	71
AMBULANCE	0.73	70
PIZZERIAS	0.72	69
FORMULA	0.72	69
RARE	0.71	68
DEVELOPMENT	0.69	66
CHABLIS	0.69	66
GUARD	0.68	65
GARBAGE	0.67	64
COLORED	0.67	64
COLESLAW	0.67	64
THURSDAYS	0.66	63
MERGERS	0.66	63
AMBLED	0.66	63
OVERCHARGED	0.65	62
ALLOW	0.65	62
WELFARE	0.62	60
SYNAGOGUE	0.61	59
PEWTER	0.61	59
CLOTH	0.61	59

Word	Normalized	Absolute
THURSDAY'S	0.60	58
HALLOWEEN	0.60	58
FORTUNE	0.60	58
DISEASES	0.60	58
IDLY	0.59	57
GENEROUS	0.59	57
CRAYONS	0.59	57
THEY'RE	0.58	56
THESE	0.58	56
SLIPPED	0.58	56
OIL	0.58	111
MONTHS	0.58	56
GLOVES	0.58	56
ARGUED	0.58	56
WORN	0.57	55
THING	0.57	55
ELEGANT	0.57	55
CONVENIENT	0.57	55
GALLON	0.56	54
THOUGHT	0.55	53
SOLVE	0.55	53
BEG	0.55	53
SERVE	0.54	52
DECORATE	0.54	52
INGREDIENTS	0.53	51
BOWL	0.53	51
BEAUTIFUL	0.53	51
AUDITION	0.52	50
WHILE	0.51	49
TWELFTH	0.51	49

Word	Normalized	Absolute
COMPLETELY	0.51	49
SURFACE	0.50	48
BEDROOM	0.50	48
ANNOYING	0.50	48
SPRAINED	0.49	47
RUN	0.49	47
PROBLEMS	0.49	47
LEARN	0.48	46
HOLIDAYS	0.48	46
CLEANERS	0.48	46
WERE	0.47	179
POPULAR	0.47	45
LUNCH	0.47	45
HARD	0.47	45
ANTELOPE	0.47	45
WORKING	0.46	44
TWILIGHT	0.46	44
REMOTE	0.46	44
POTATOES	0.46	44
GOAT	0.46	44
FROST	0.46	44
ARRANGE	0.46	44
SHOULDER	0.45	43
RATHER	0.45	43
PROJECT	0.45	43
CLASSROOMS	0.45	43
AVOID	0.45	43
WITHIN	0.44	42
REVIEW	0.44	42
JANUARY	0.44	42
DIRTY	0.44	42
CHARGE	0.44	42
OWN	0.43	41
EARN	0.43	41
CHANGE	0.43	41
ADD	0.43	41
THROUGH	0.42	40
SHOES	0.42	40

Word	Normalized	Absolute
SCARED	0.42	40
PROBLEM	0.42	40
OBJECTS	0.42	40
LORI'S	0.42	40
KINDERGARTEN	0.42	40
UPON	0.41	39
NEAREST	0.41	39
MIAMI	0.41	39
LOWER	0.41	39
BLACK	0.41	39
AWAY	0.40	38
ASSISTANCE	0.40	38
ANSWER	0.40	38
YARD	0.39	37
THIS	0.39	37
LESSONS	0.39	37
LARGE	0.39	37
GRADES	0.39	37
EVERY	0.39	37
WHERE	0.38	36
DIAGRAM	0.38	36
ANKLE	0.38	36
SHAVING	0.36	35
LIE	0.36	35
THEIR	0.35	34
STEEP	0.35	34
RIDE	0.35	34
QUICK	0.35	34
NANCY'S	0.35	34
LAUGH	0.35	34
TOMORROW	0.34	33
QUESTION	0.34	33
PROCEEDING	0.34	33
HIGHER	0.34	33
COUNTRYSIDE	0.34	33
COSTUME	0.34	33
CHILDREN	0.34	66
CAROL	0.34	33

Word	Normalized	Absolute
WE	0.33	32
WALKING	0.33	32
FAVOR	0.33	32
CREAM	0.33	32
BUSINESS	0.33	32
WOULD	0.32	31
WITH	0.32	31
SLOWLY	0.32	31
GUESS	0.32	31
DRIVING	0.32	31
BROKEN	0.32	31
PAM	0.31	30
HAVEN'T	0.31	30
ENJOY	0.31	30
DIG	0.31	30
USE	0.30	29
TABLE	0.30	29
SLOPE	0.30	29
MONEY	0.30	58
GIVES	0.30	29
FORMS	0.30	29
DROP	0.30	29
STUDY	0.29	28
STOCKINGS	0.29	28
MEDICAL	0.29	28
LEAP	0.29	28
THAT	0.28	82
NOISE	0.28	27
FARM	0.28	27
OUT	0.27	26
WILL	0.26	25
WALL	0.26	25
THAN	0.26	49
SWING	0.26	25
NEEDED	0.26	25
MINE	0.26	25
CHEAP	0.26	25
ALWAYS	0.26	25

Word	Normalized	Absolute
SUCH	0.25	24
SUBWAY	0.25	24
JENNIFER'S	0.25	24
JANE	0.25	24
FISH	0.25	24
YOUNG	0.24	23
THE	0.24	562
MUCH	0.24	23
MICHAEL	0.24	23
FIVE	0.24	23
ENOUGH	0.24	23
DISTANCE	0.24	23
COMBINE	0.24	23
ARM	0.24	23
WE'LL	0.23	22
SING	0.23	22
PEOPLE	0.23	22
I'D	0.23	44
HOUSE	0.23	22
GAS	0.23	22
FREEWAY	0.23	22
BIG	0.23	44
HER	0.22	42
FIRST	0.22	21
DANCE	0.22	21
AFTER	0.22	42
TWO	0.21	20
THEY	0.21	40
SMILES	0.21	20
SAME	0.21	20
RIGHT	0.21	20
LAKE	0.21	20
HAVE	0.21	40
CALL	0.21	20
ALL	0.21	40
YOUR	0.20	38
PLEASE	0.20	39
ITEM	0.20	19

Word	Normalized	Absolute
FELT	0.20	19
CAN	0.20	19
THERE	0.19	18
SHOULD	0.18	17
IT	0.18	35
HIGHWAY	0.18	17
UP	0.17	16
AN	0.17	16
WHEN	0.16	30
SPEND	0.15	14
MORE	0.15	42
JEFF	0.15	14
GO	0.15	14
COMES	0.15	14
BEST	0.15	28
SHE	0.14	13
SENSE	0.14	13
EACH	0.14	13
BEFORE	0.14	41
YOU	0.12	69
SOUND	0.12	12
HIS	0.12	12
HIGH	0.12	12
A	0.12	108
SMALL	0.11	11
ONE	0.11	11
I	0.11	33
DECEMBER	0.11	11
BUY	0.11	21
AND	0.11	41
NOT	0.10	28
MEETING	0.10	10
MAY	0.10	28
HE	0.10	10

Word	Normalized	Absolute
BEGAN	0.10	10
AT	0.10	20
TOO	0.09	9
TIME	0.09	18
TAKE	0.09	9
ONLY	0.09	9
NOW	0.09	9
NICE	0.09	9
MADE	0.09	9
DAY	0.09	17
ARE	0.09	35
OF	0.08	37
MEAN	0.08	8
MAKES	0.08	8
GROWS	0.08	8
FROM	0.08	8
FOR	0.08	69
AS	0.08	22
WHY	0.07	14
BY	0.07	7
BUT	0.07	7
NEED	0.06	6
ME	0.06	6
IN	0.06	28
IF	0.06	23
WAY	0.05	5
OR	0.05	5
WAS	0.04	7
ON	0.04	18
NO	0.04	4
BE	0.04	21
MY	0.03	3
IS	0.03	5
TO	0.02	15

A.2 FAU LME Non-Native Database

Information about the speaker level human rating (HumRating), number of words considered as mispronounced (Miscnt), the relative share of words considered mispronounced (MisRatio), the number of utterances (#utr) and words (#words), the optimal warping factor α for VTLN, and word (WA) and phoneme accuracy (PA) for each speaker in the FAU LME Non-Native Children Speech Corpus.

ID	HumRating	Miscnt	MisRatio	#utr	#words	VTLN α	WA	PA
w014	1	2	0.01	62	388	0.90	44.85	25.64
m101	1	24	0.04	289	624	1.00	28.21	-5.74
m031	1	5	0.03	29	156	1.00	46.79	9.96
m023	2	4	0.02	30	208	0.95	40.38	9.67
w028	2	6	0.02	62	390	0.90	38.46	2.07
m011	2.5	14	0.08	22	165	1.05	33.33	-20.68
w020	2	17	0.05	46	328	1.0	26.83	0.19
w018	2	5	0.02	56	316	0.90	23.42	-0.46
m002	2	3	0.02	29	182	1.00	31.32	2.48
m008	2	6	0.03	30	214	0.90	15.42	-5.15
m025	2	7	0.02	60	365	0.95	33.70	0.93
m027	2	5	0.01	62	395	0.95	31.65	-0.99
w006	2	5	0.01	61	420	0.90	35.48	1.53
m030	2	6	0.02	62	394	0.95	27.66	2.35
w207	2	18	0.04	79	440	0.95	36.82	12.86
m005	2	10	0.05	32	195	1.00	28.72	3.43
m102	2	57	0.09	289	641	1.00	22.46	-12.90
m210	2	17	0.10	100	177	1.00	32.20	-20.70
m221	2	16	0.09	100	172	1.00	25.00	-31.55
w225	2	6	0.04	100	168	0.95	26.79	-28.11
w032	3	5	0.01	60	361	0.90	26.59	-8.41
m217	3	31	0.10	127	324	0.95	29.01	-7.24
m218	3.5	45	0.12	134	373	0.90	32.71	-15.54
w224	3	26	0.15	100	174	1.00	17.82	-51.23
w223	3	18	0.11	100	170	0.95	27.06	-43.24

ID	HumRating	Miscnt	MisRatio	#utter	#words	VTLN α	WA	PA
w222	3	26	0.05	143	496	1.00	30.04	-8.74
w001	3	11	0.03	62	392	0.90	30.10	3.58
w003	3	17	0.04	62	389	0.90	30.85	-7.61
m103	3	99	0.09	349	1083	0.95	22.07	-11.13
m029	3	14	0.07	29	190	1.00	35.26	-14.39
w216	3	55	0.10	79	531	0.95	12.05	-22.22
m021	3	6	0.03	31	209	1.00	36.36	-1.70
w213	3	37	0.08	147	465	1.00	37.20	-8.44
w212	3	27	0.15	100	175	0.90	14.29	-37.78
m013	3	20	0.05	61	384	1.05	30.47	-10.80
w022	3	16	0.04	56	358	0.90	29.89	-14.44
w024	3	25	0.07	61	379	0.90	33.77	2.72
m009	3	22	0.05	62	402	1.05	28.36	-15.19
w209	3	67	0.16	49	423	0.95	14.42	-7.89
w104	3	38	0.06	289	636	1.00	23.78	-7.68
w208	3.5	70	0.14	76	485	0.90	23.51	-8.40
m007	3	5	0.03	29	157	0.90	27.39	-16.20
w203	3	47	0.09	76	497	0.90	24.55	-9.81
w205	3	40	0.19	30	212	0.95	32.08	-15.36
m211	4	61	0.13	141	474	1.00	27.43	-21.04
w202	4	61	0.22	58	277	0.95	23.47	-14.91
w201	4	47	0.21	28	220	0.90	15.45	-42.98
w206	4	77	0.24	46	326	1.05	28.83	-9.15
m019	4	19	0.09	32	203	1.00	17.24	-16.52
w016	4	9	0.04	30	216	0.95	21.76	-8.38
w214	4	33	0.10	37	333	0.95	20.72	-25.94
m220	4	65	0.15	141	445	1.05	15.96	-26.42
w010	4	27	0.07	62	398	0.90	27.14	-8.17
w215	4	37	0.14	41	270	0.90	21.11	-9.08
m204	5	43	0.19	35	229	1.20	12.23	-15.63
m219	5	23	0.21	41	112	0.90	17.86	-47.06
w004	5	17	0.11	23	156	0.90	34.62	-7.76

A.3 TIMIT Corpus

Table A.3 shows how the narrow transcriptions were converted into wider transcriptions using less phoneme symbols. The conversion was necessary to be able to compute the forced-alignment with the native acoustic model trained with speech data from the WSJ corpus. The conversion rules are derived from the phoneme symbol documentation included in the TIMIT corpus.

Original Symbol	Target symbol
ax-h	ax
eng	ix ng
el	ax l
en	ax n
em	ax m
hv	hh
q	t
ux	uw
nx	n

Appendix B

Human Evaluation

B.1 Evaluation Instructions

- The aim of this experiment is to obtain an assessment of proficiency of non-native speakers in terms of pronunciation and fluency.

- First we would like you to listen to 22 utterances. Each sentence is uttered by a different speaker. Please assign a level of proficiency to each utterance. Each level of proficiency should be selected at least once. Level 1 (green) should be used for maximum proficiency, Level 5 (red) for minimum proficiency present among speakers. The purpose of this step is to give a feeling of how to use the grading scale.

- Then you will start evaluation for a subset of 100 non-native speakers. There are 48 different utterances per speaker. They are presented in random order. You can listen twice to an utterance if it seems necessary to you. First we would like to ask you to mark any mispronounced words. Please tolerate any pronunciation variations which may exist in standard British or standard American English. Please consider only phonetic errors and ignore wrong lexical stress.

- There may be some words a speaker has (completely) misread. Please also mark these words as mispronounced.

- Next, we would like you to select a level of overall proficiency for the utterance by considering both pronunciation and fluency. Badly pronounced words, misread words, strong non-native accent, long pauses between words, stuttering, etc. should have an influence on your assessment. Please ignore sentence intonation for your evaluation.

- If there is an utterances which was not playable, please skip the sentence by selecting *NONE*, any proficiency level and clicking on the submit button. Please write down the number of any sentence for which the utterance was not playable on this instruction sheet.

B.2 Evaluator Information

Table B.1 gives information about the human evaluators, who marked mispronounced words and assigned ratings to each utterance of the non-native speakers. The evaluators {3,16} grew up and went to school in Canada, all other evaluators in the U. S.

Table B.1: *Information about the evaluators.*

Evaluator ID	Working Places	Teaching Experience	Background in Phonetics
2	Private School (Japan)	9 months	no
3	Private School (Japan)	6 months	yes
5	Private School / Companies (Japan)	5 years	yes
6	Private School / Companies / Privately	4-5 years	no
7	Privately (Japan)	2-3 years	some
8	Privately (Japan)	1-2 years	yes
9	University (Japan)	5 years	yes
10	Public School (Japan)	3 years	some
11	Private School (Japan, Canada)	2-3 years	yes
12	Private School (Japan)	17 months	no
13	Private School (Japan)	18 months	no
15	Privately (Japan)	3-4 years	no
16	Private Schools / Companies (Taiwan, Hong Kong, Japan, Canada)	8 years	no
17	Company (Japan)	6 years	some
20	Various places (Europe,Asia) Creation of teaching material (e.g. audio CDs for TOIEC test)	9 years	yes

Table B.2 shows, which rater had to evaluate the utterances of which speakers. All evaluators processed all of the requested 1152 utterances (24 speakers times 48 sentences). The result of utterance rating to score correlation analysis carried out separately for each evaluator is shown in Table B.3.

Table B.2: *Which rater evaluated which non-native speaker?*

Evaluator IDs	Non-native speaker IDs
8 12 16 20	M082 M027 M077 F018 M044 M088 M010 M080 M072 M064 M030 M049 M025 M033 M093 F012 M023 F014 M014 M055 M061 M028 M056 M076
5 9 13 17	M029 M021 M038 M057 F021 M047 M066 F023 M011 M054 F009 M075 M031 M034 M067 M051 M015 M032 M062 M006 F008 M065 M090 M086
2 6 10	F010 M071 M052 M045 M024 M081 F025 M092 M016 M078 M073 M040 F013 M046 M053 F022 M084 M091 M039 F019 M013 M036 M050 M037
3 7 11 15	M074 M043 M026 M070 M060 F024 F026 M022 M041 F011 M087 M058 M042 M069 M012 F020 M085 M068 M089 M063 M059 M083 M001 M035

Table B.3: *Separate human rating to score correlation analysis for each evaluator.*

EvalID	L7	K3	R4	P2	D2	X1	M3	E1
2	-0.36	-0.50	+0.32	+0.21	-0.40	-0.43	-0.37	+0.23
3	-0.30	-0.42	+0.27	+0.23	-0.30	-0.38	-0.27	+0.17
5	-0.33	-0.40	+0.30	+0.28	-0.37	-0.32	-0.35	+0.23
6	-0.39	-0.45	+0.36	+0.28	-0.38	-0.38	-0.43	+0.30
7	-0.32	-0.39	+0.30	+0.24	-0.31	-0.37	-0.35	+0.19
8	-0.42	-0.49	+0.40	+0.34	-0.47	-0.39	-0.37	+0.36
9	-0.39	-0.39	+0.37	+0.33	-0.37	-0.34	-0.37	+0.24
10	-0.46	-0.55	+0.43	+0.29	-0.45	-0.45	-0.44	+0.28
11	-0.21	-0.39	+0.18	+0.14	-0.25	-0.41	-0.20	+0.13
12	-0.33	-0.34	+0.31	+0.33	-0.34	-0.29	-0.28	+0.29
13	-0.28	-0.30	+0.27	+0.21	-0.31	-0.26	-0.28	+0.16
15	-0.12	-0.31	+0.09	+0.11	-0.17	-0.29	-0.16	+0.11
16	-0.40	-0.48	+0.37	+0.32	-0.47	-0.46	-0.33	+0.30
17	-0.19	-0.26	+0.17	+0.17	-0.25	-0.27	-0.20	+0.15
20	-0.43	-0.51	+0.42	+0.34	-0.49	-0.42	-0.38	+0.35

B.3 Screenshots

http://lab.slt.atr.co.jp/cgi-bin/xtcinca/start.py

Hello Tobias Cincarek !

1. Listen to the following sentences to get an impression of proficiency of some non-native speakers.
2. Assign a level of proficiency to each speaker. Each level of proficiency should be selected once.
3. Start evaluation.

	Proficiency Level (Pronunciation, Fluency)				
Listen to ...	1	2	3	4	5
utterance of speaker 1	○	○	○	○	○
utterance of speaker 2	○	○	○	○	○
utterance of speaker 3	○	○	○	○	○
utterance of speaker 4	○	○	○	○	○
utterance of speaker 5	○	○	○	○	○
utterance of speaker 6	○	○	○	○	○
utterance of speaker 7	○	○	○	○	○
utterance of speaker 8	○	○	○	○	○
utterance of speaker 9	○	○	○	○	○
utterance of speaker 10	○	○	○	○	○
utterance of speaker 11	○	○	○	○	○
utterance of speaker 12	○	○	○	○	○
utterance of speaker 13	○	○	○	○	○
utterance of speaker 14	○	○	○	○	○
utterance of speaker 15	○	○	○	○	○
utterance of speaker 16	○	○	○	○	○
utterance of speaker 17	○	○	○	○	○
utterance of speaker 18	○	○	○	○	○
utterance of speaker 19	○	○	○	○	○
utterance of speaker 20	○	○	○	○	○
utterance of speaker 21	○	○	○	○	○
utterance of speaker 22	○	○	○	○	○

Start evaluation from sentence 0 + 1 Start evaluation

http://lab.slt.atr.co.jp/cgi-bin/xtcinca/eval_utterance.py

Evaluation of utterance 2 of 1152

1. Click on hyperlink *PLAY* to replay the utterance.
2. Mark mispronounced words. Click on checkbox for *NONE* if there are not any.
3. Select level of proficiency considering pronunciation and fluency.
4. Go on to the next utterance by clicking on the submit button.

PLAY	Mispronounced?
HIS	☐
SHOULDER	☐
FELT	☐
AS	☐
IF	☐
IT	☐
WERE	☐
BROKEN	☐
NONE	☐

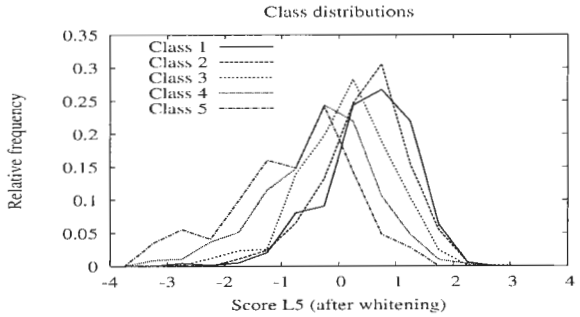
Proficiency level (Pronunciation, Fluency)				
1	2	3	4	5
○	○	○	○	○

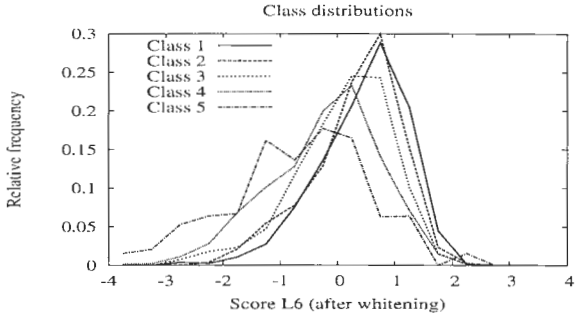
Submit Reset

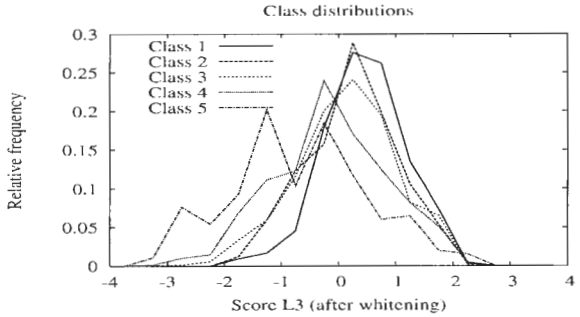
Appendix C

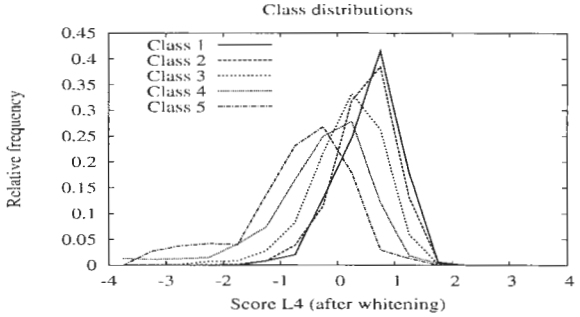
Gallery of Scores

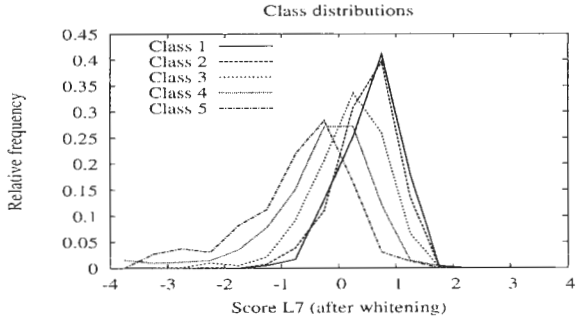
C.1 Utterance Scores

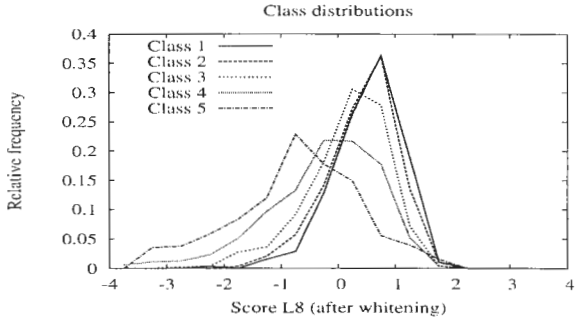


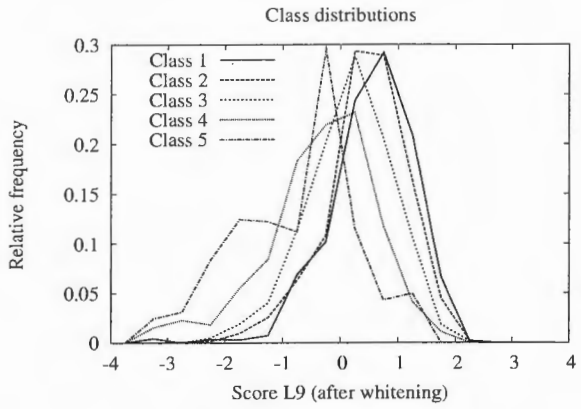


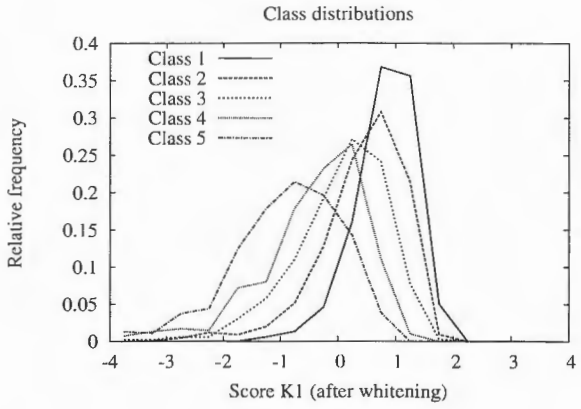


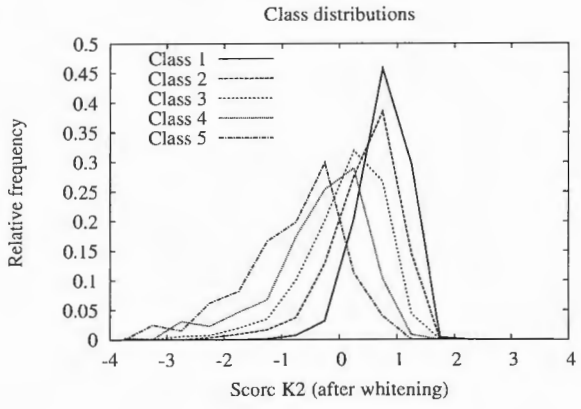


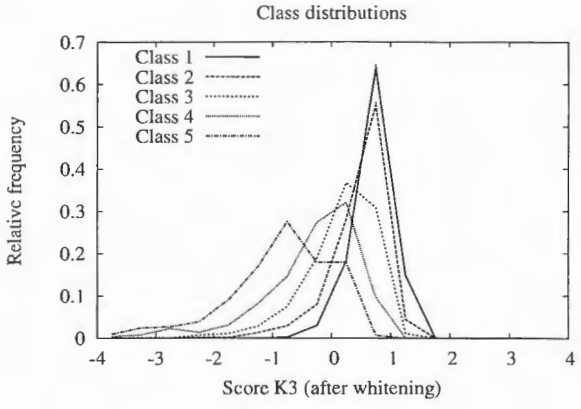


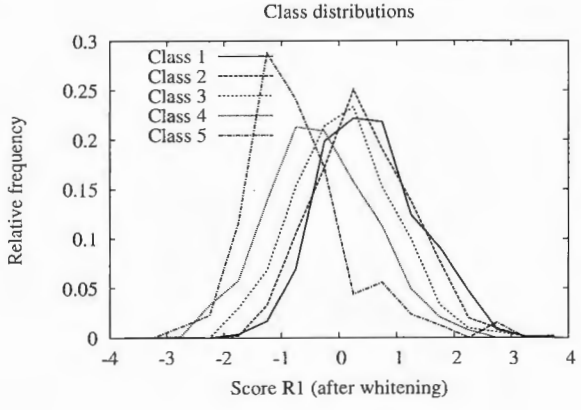


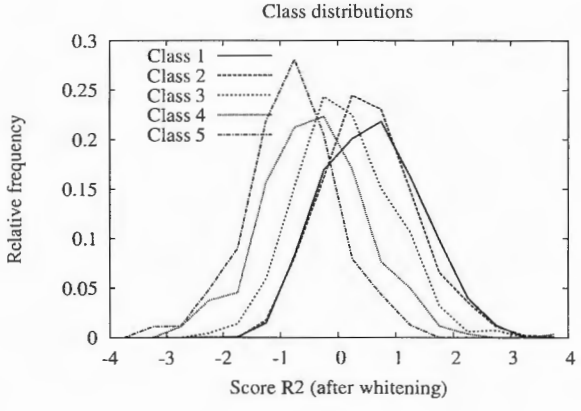

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score L9 (after whitening)

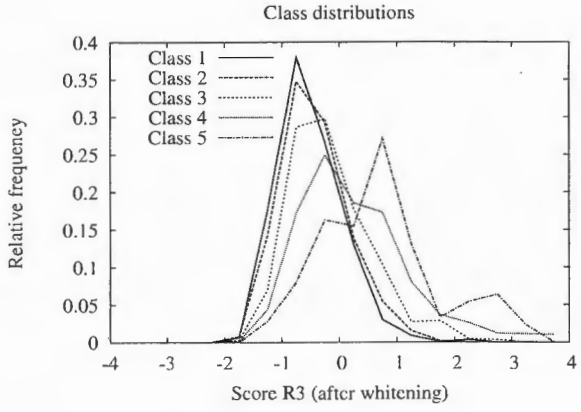

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score K1 (after whitening)

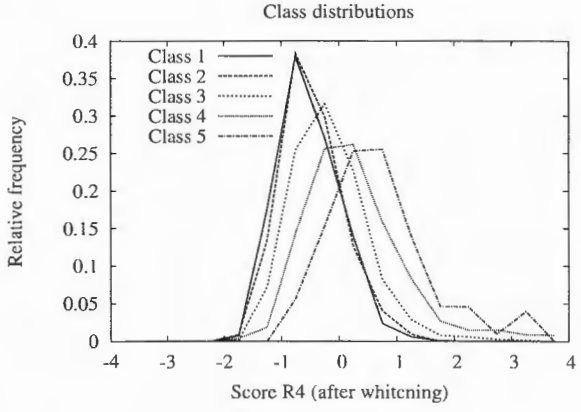

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score K2 (after whitening)

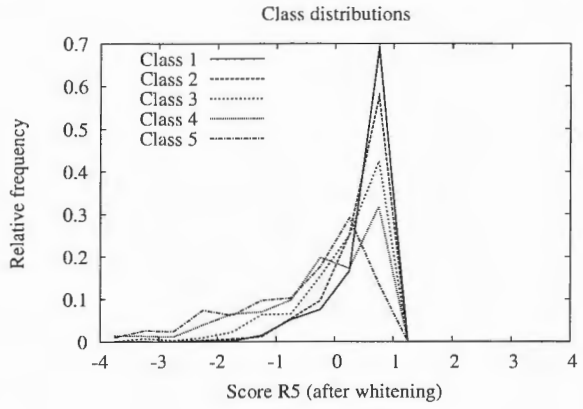

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score K3 (after whitening)

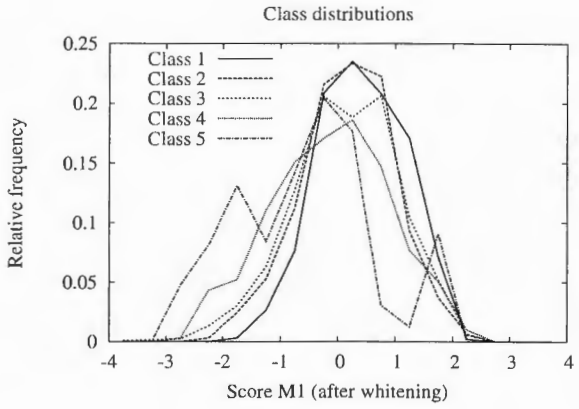

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score R1 (after whitening)

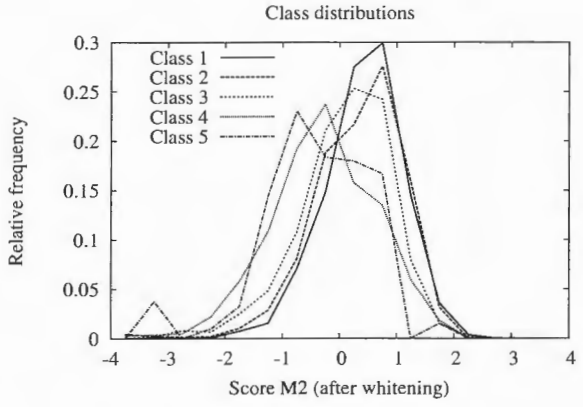

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score R2 (after whitening)

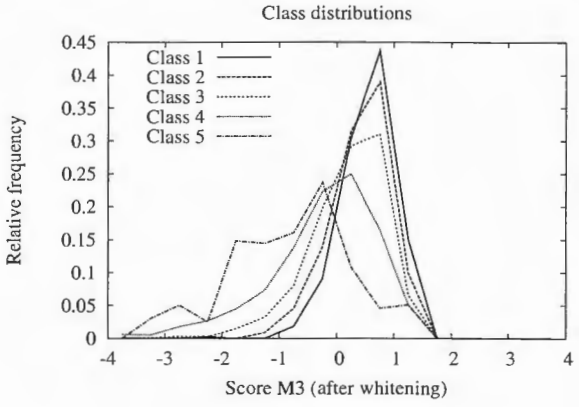

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score R3 (after whitening)

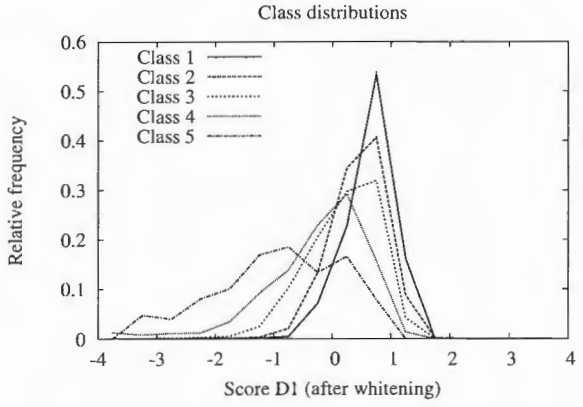

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score R4 (after whitening)

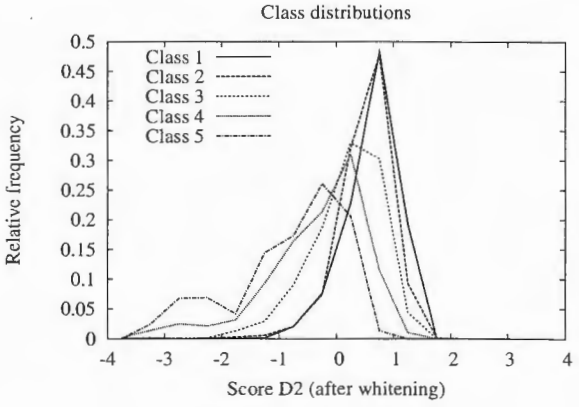

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score R5 (after whitening)

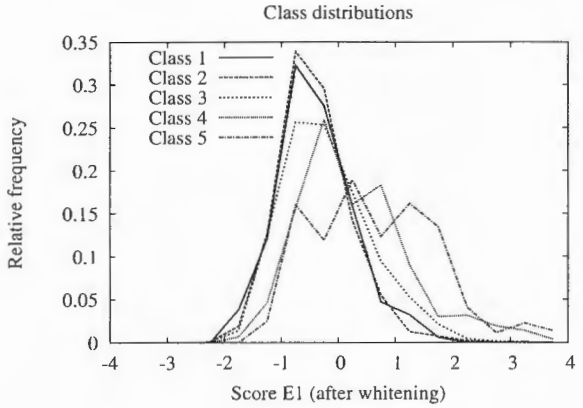

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score M1 (after whitening)

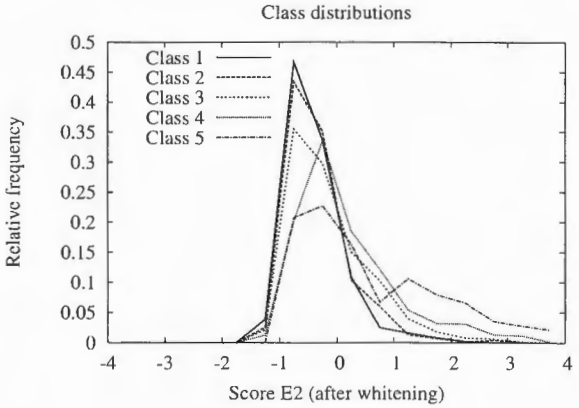

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score M2 (after whitening)

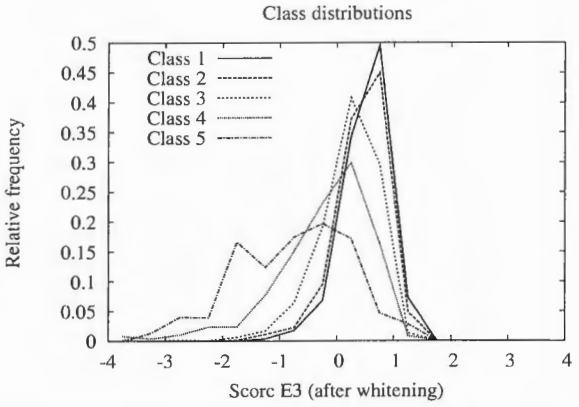

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score M3 (after whitening)

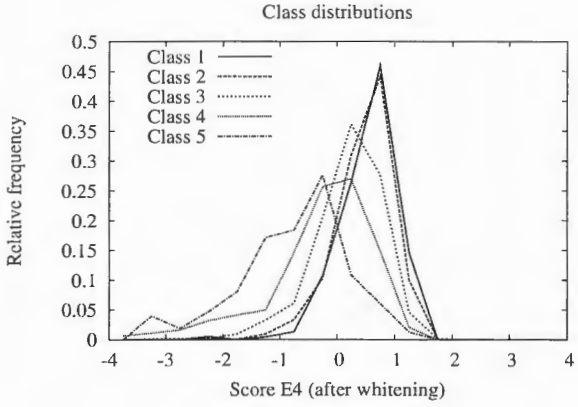

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score D1 (after whitening)

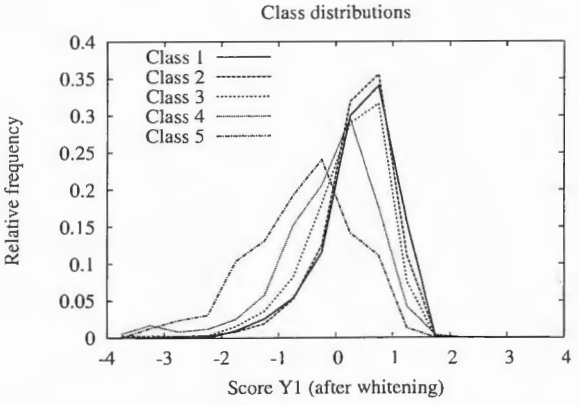

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score D2 (after whitening)

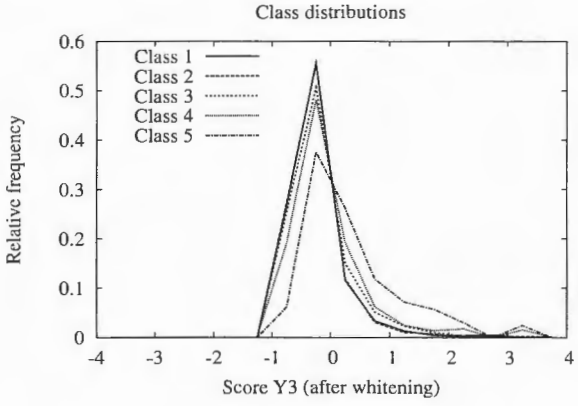

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score E1 (after whitening)

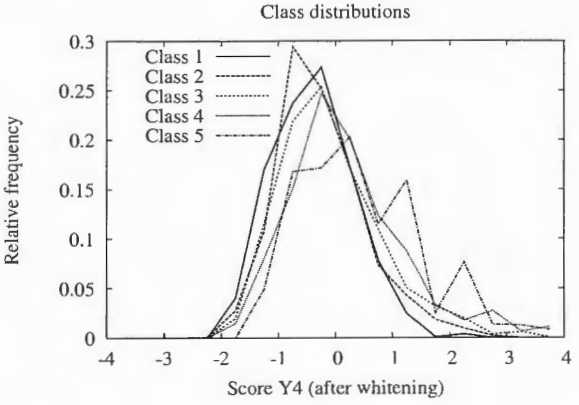

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score E2 (after whitening)

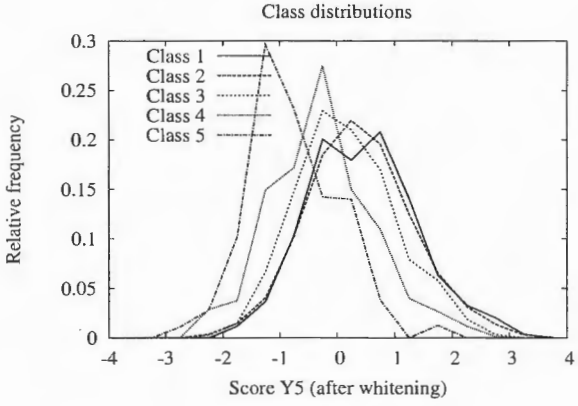

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score E3 (after whitening)

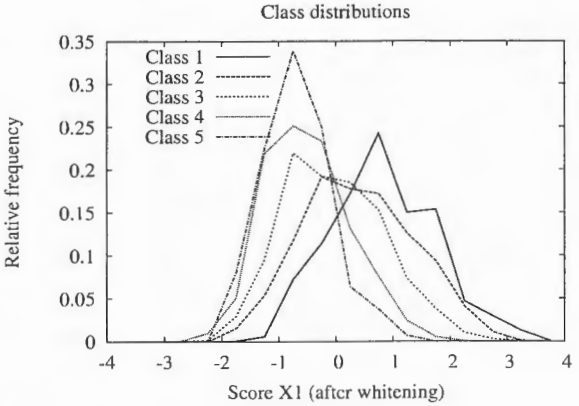

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score E4 (after whitening)

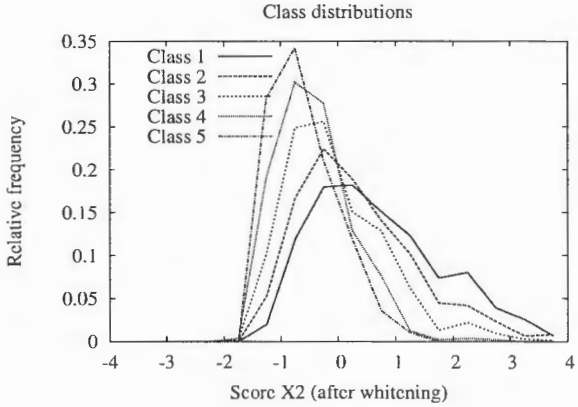

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score Y1 (after whitening)


Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score Y3 (after whitening)

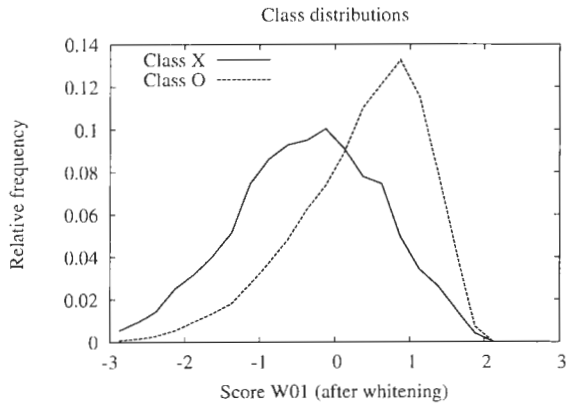

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score Y4 (after whitening)

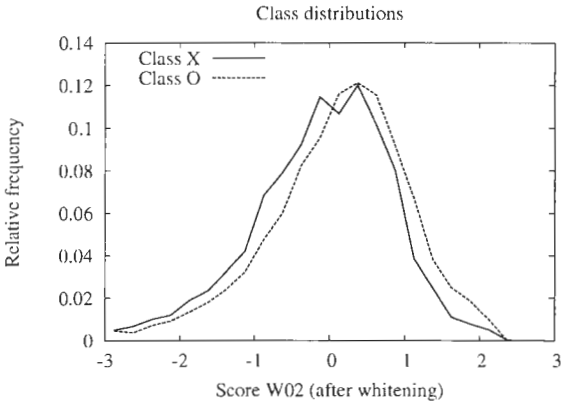

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score Y5 (after whitening)

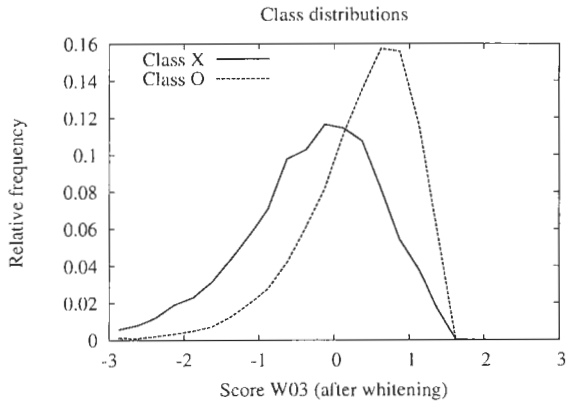

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score X1 (after whitening)

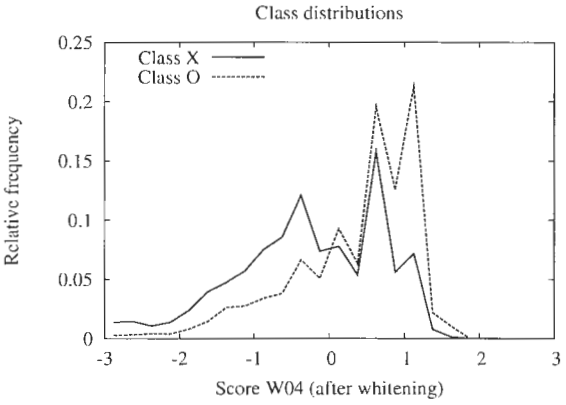

Class distributions
Class 1
Class 2
Class 3
Class 4
Class 5
Relative frequency
Score X2 (after whitening)

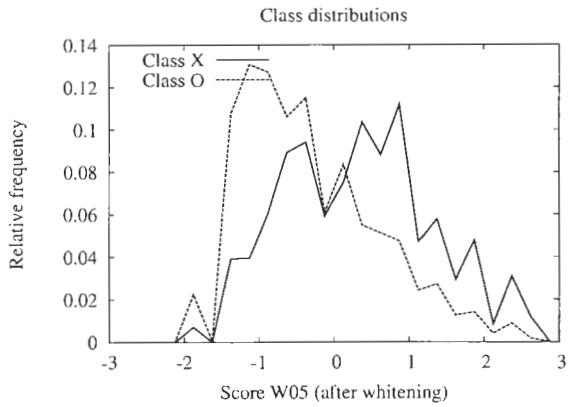
C.2 Word Scores

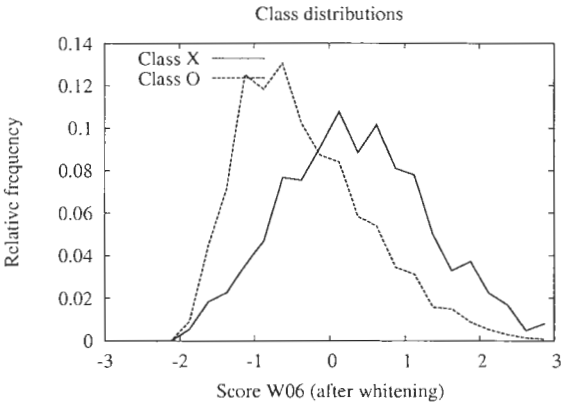


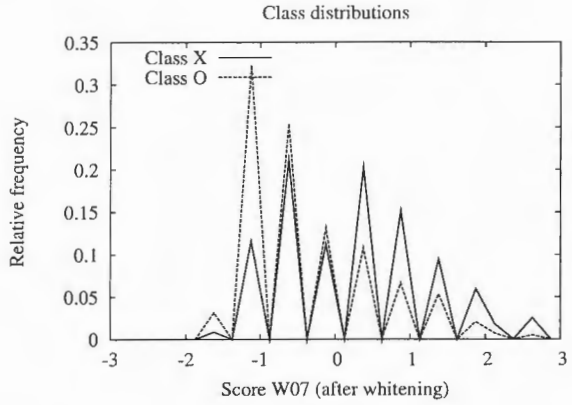


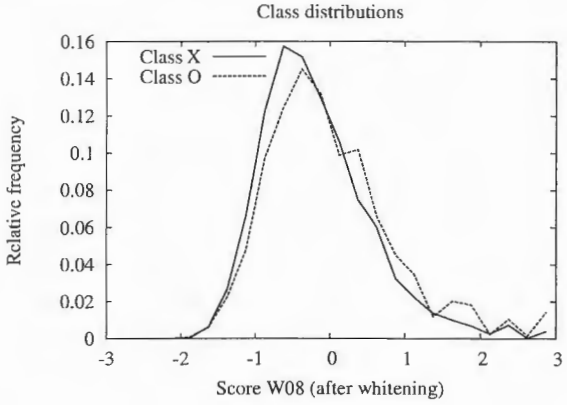


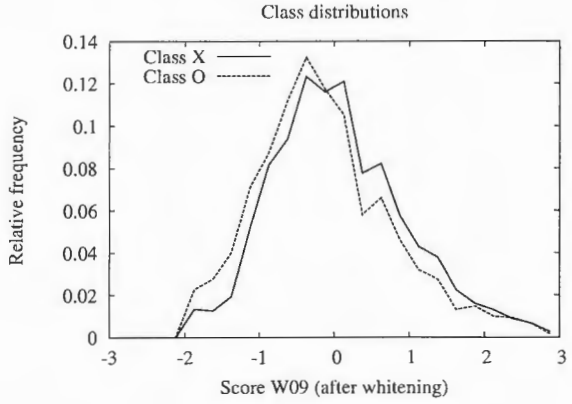


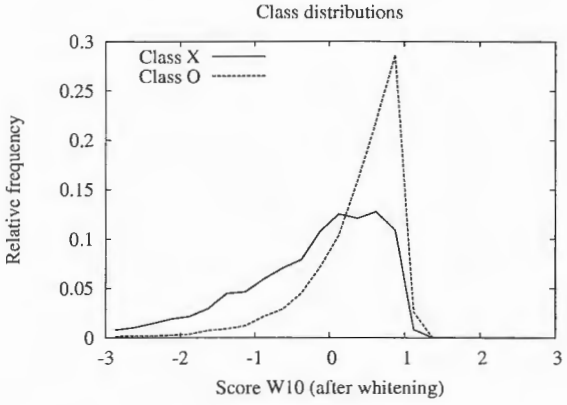


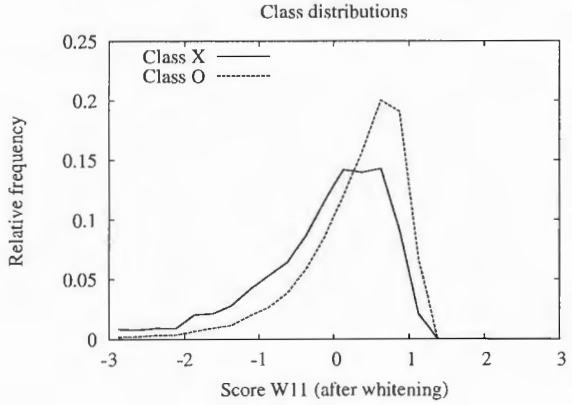


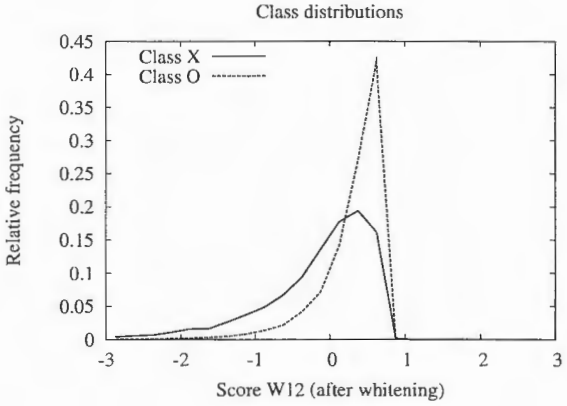

Class distributions
Class X
Class O
Relative frequency
Score W07 (after whitening)

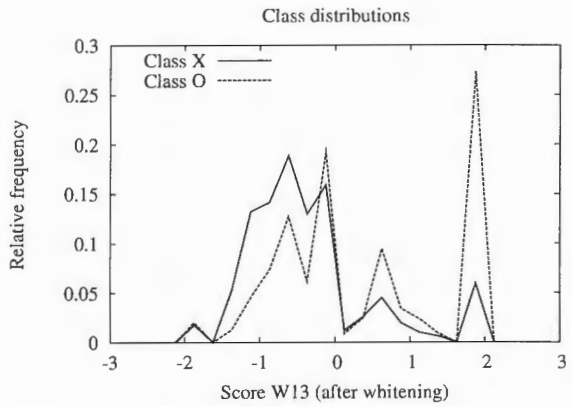

Class distributions
Class X
Class O
Relative frequency
Score W08 (after whitening)

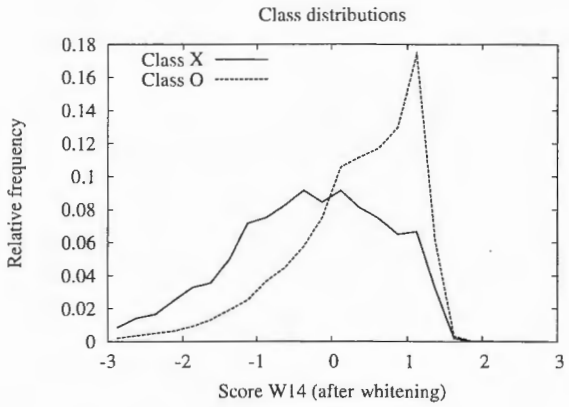

Class distributions
Class X
Class O
Relative frequency
Score W09 (after whitening)

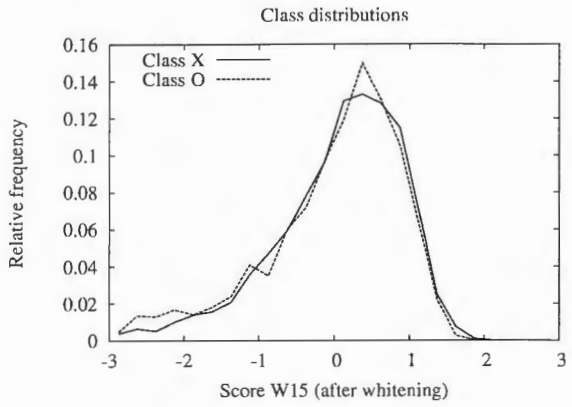

Class distributions
Class X
Class O
Relative frequency
Score W10 (after whitening)

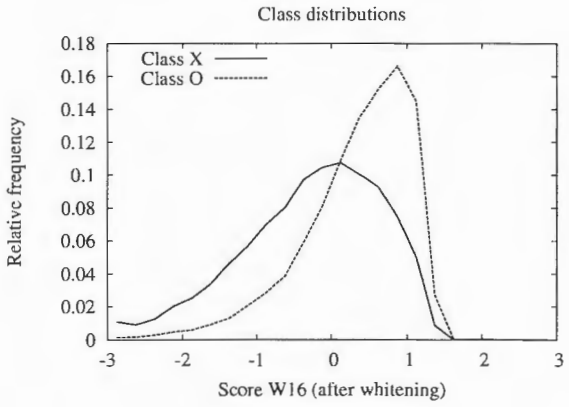

Class distributions
Class X
Class O
Relative frequency
Score W11 (after whitening)

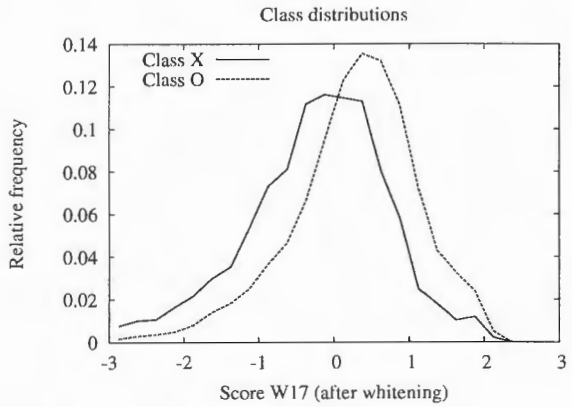

Class distributions
Class X
Class O
Relative frequency
Score W12 (after whitening)

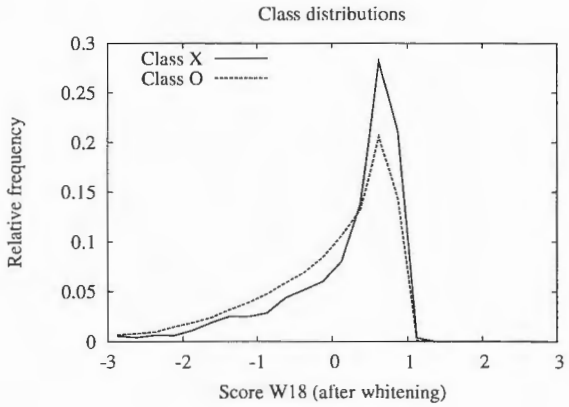

Class distributions
Class X
Class O
Relative frequency
Score W13 (after whitening)

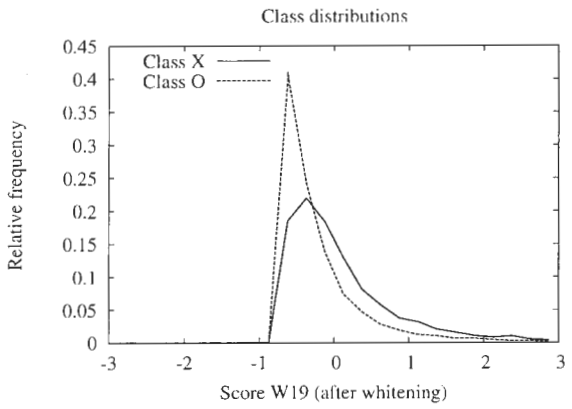

Class distributions
Class X
Class O
Relative frequency
Score W14 (after whitening)

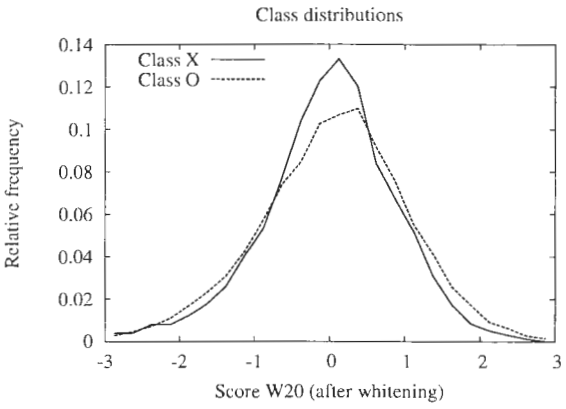

Class distributions
Class X
Class O
Relative frequency
Score W15 (after whitening)

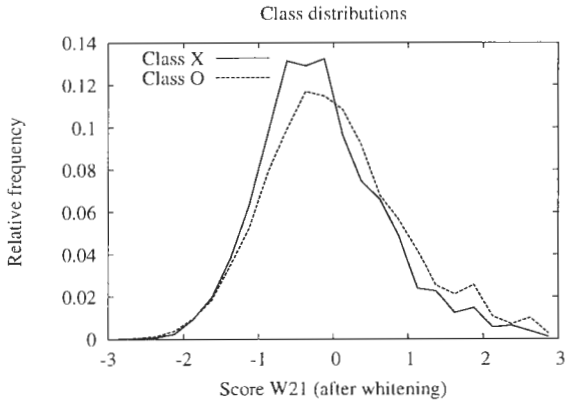

Class distributions
Class X
Class O
Relative frequency
Score W16 (after whitening)

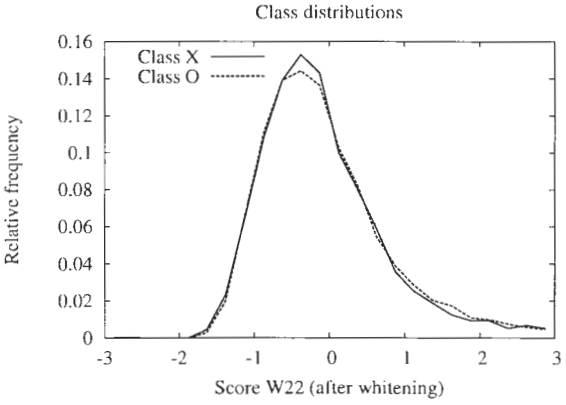

Class distributions
Class X
Class O
Relative frequency
Score W17 (after whitening)

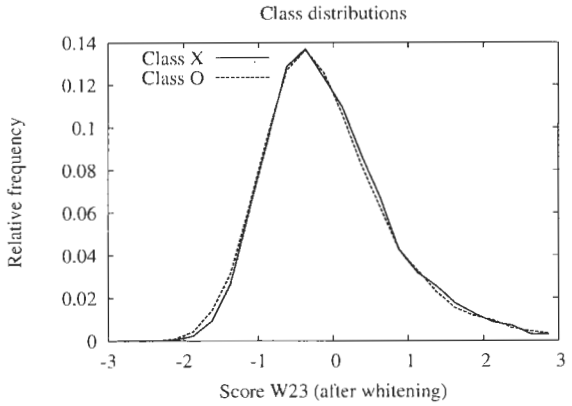

Class distributions
Class X
Class O
Relative frequency
Score W18 (after whitening)

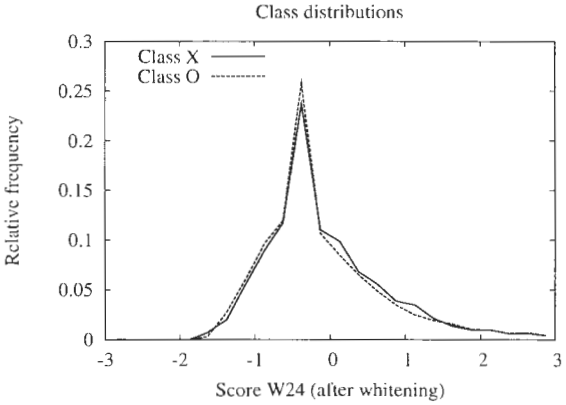

Class distributions
Class X
Class O
Relative frequency
Score W19 (after whitening)

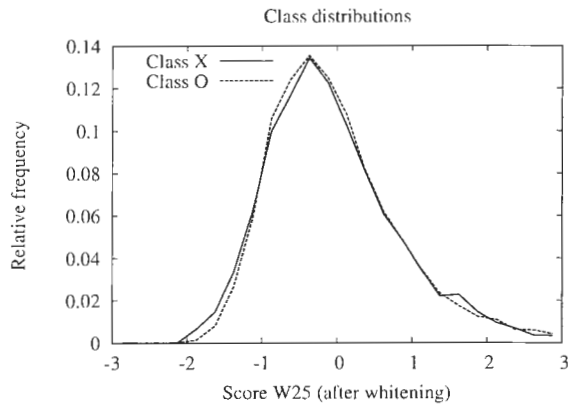

Class distributions
Class X
Class O
Relative frequency
Score W20 (after whitening)

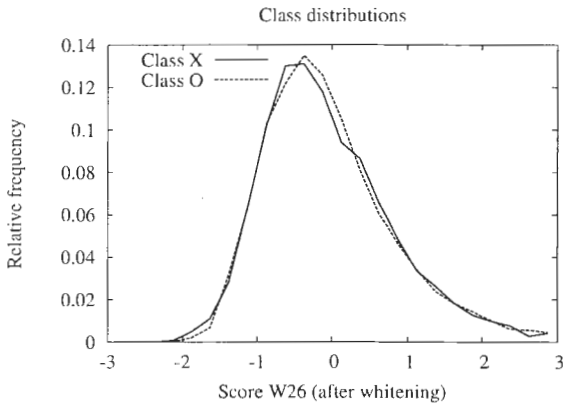

Class distributions
Class X
Class O
Relative frequency
Score W21 (after whitening)

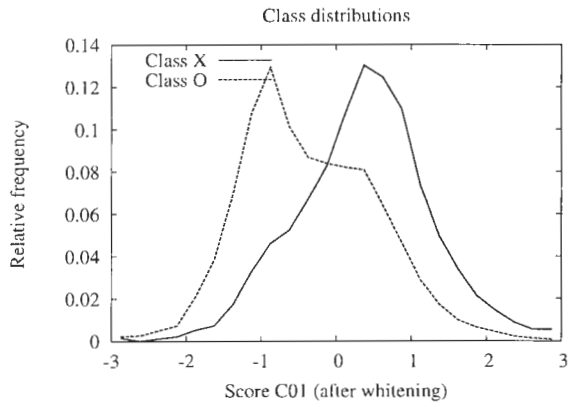

Class distributions
Class X
Class O
Relative frequency
Score W22 (after whitening)

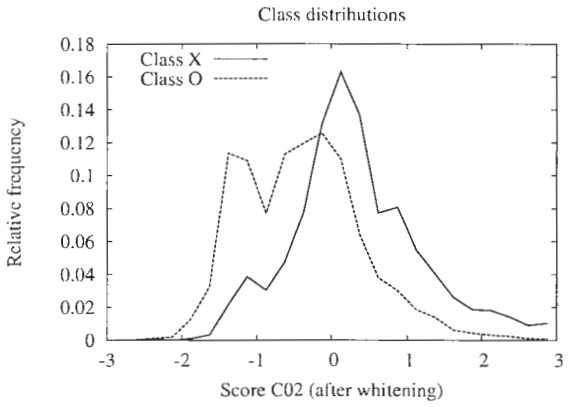

Class distributions
Class X
Class O
Relative frequency
Score W23 (after whitening)

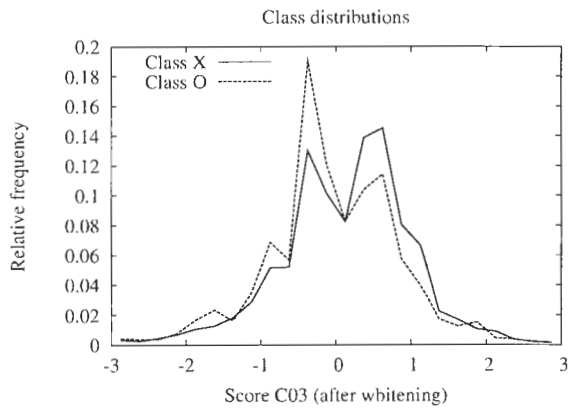

Class distributions
Class X
Class O
Relative frequency
Score W24 (after whitening)

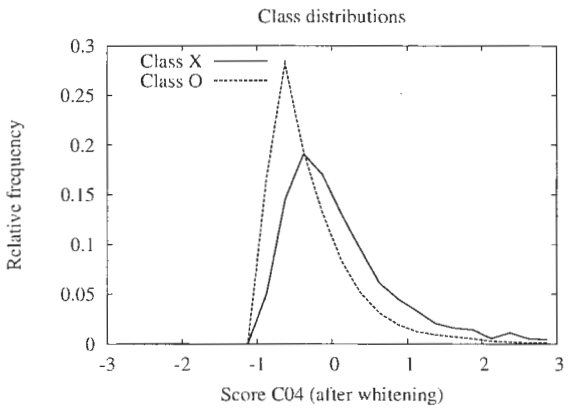

Class distributions
Class X
Class O
Relative frequency
Score W25 (after whitening)

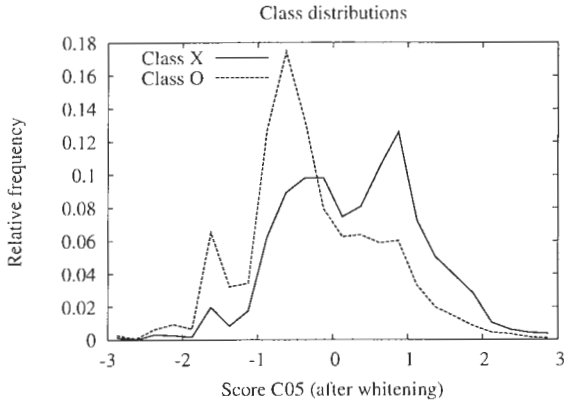

Class distributions
Class X
Class O
Relative frequency
Score W26 (after whitening)

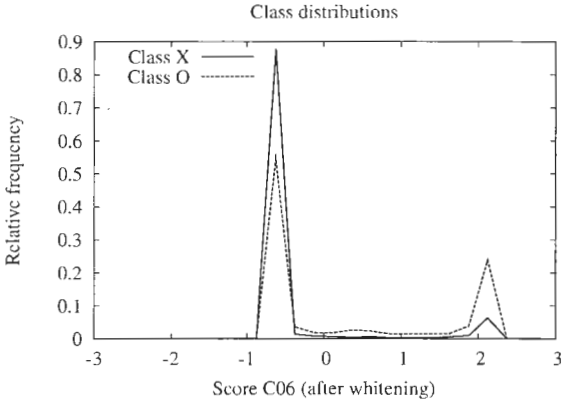

Class distributions
Class X
Class O
Relative frequency
Score C01 (after whitening)

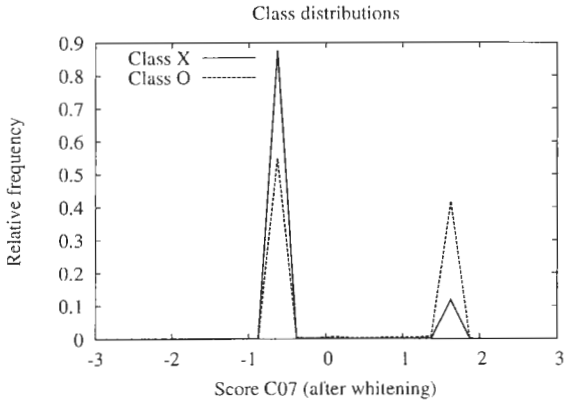

Class distributions
Class X
Class O
Relative frequency
Score C02 (after whitening)

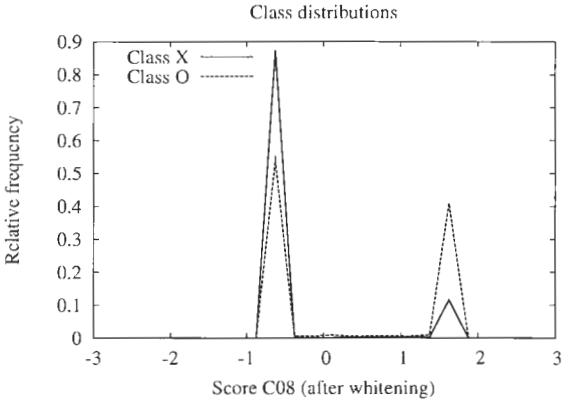

Class distributions
Class X
Class O
Relative frequency
Score C03 (after whitening)

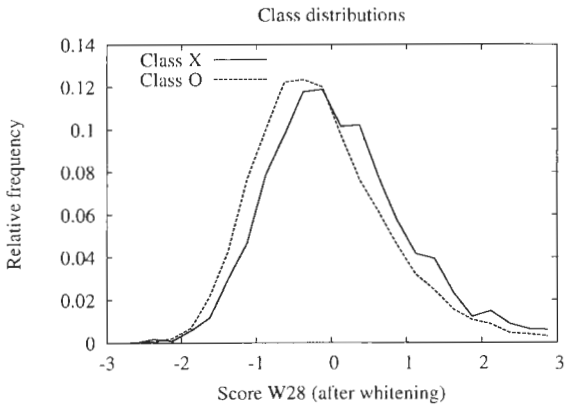

Class distributions
Class X
Class O
Relative frequency
Score C04 (after whitening)

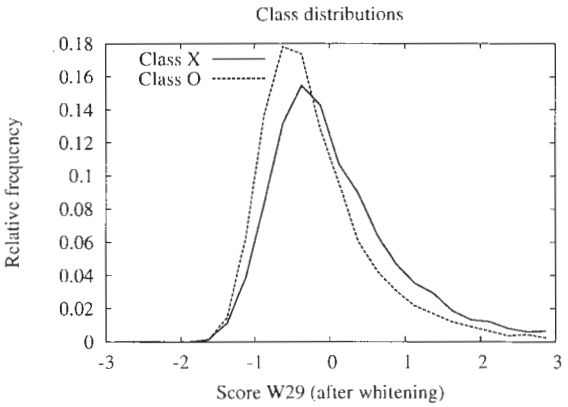

Class distributions
Class X
Class O
Relative frequency
Score C05 (after whitening)


Class distributions
Class X
Class O
Relative frequency
Score C06 (after whitening)


Class distributions
Class X
Class O
Relative frequency
Score C07 (after whitening)


Class distributions
Class X
Class O
Relative frequency
Score C08 (after whitening)


Class distributions
Class X
Class O
Relative frequency
Score W28 (after whitening)


Class distributions
Class X
Class O
Relative frequency
Score W29 (after whitening)

C.3 Inter-Score Correlations

The following tables show the inter-score correlations and the correlation between human ratings and scores for the ATR SLT data at the utterance level.

	HR	L3	L4	L5	L6	L7	L8	L9	K1
HR	+1.00	-0.24	-0.41	-0.34	-0.28	-0.42	-0.37	-0.35	-0.48
L3	-0.24	+1.00	+0.25	+0.16	+0.56	+0.25	+0.56	+0.19	+0.04
L4	-0.41	+0.25	+1.00	+0.86	+0.28	+1.00	+0.54	+0.84	+0.31
L5	-0.34	+0.16	+0.86	+1.00	+0.25	+0.85	+0.32	+0.95	+0.23
L6	-0.28	+0.56	+0.28	+0.25	+1.00	+0.28	+0.91	+0.29	+0.14
L7	-0.42	+0.25	+1.00	+0.85	+0.28	+1.00	+0.54	+0.83	+0.33
L8	-0.37	+0.56	+0.54	+0.32	+0.91	+0.54	+1.00	+0.36	+0.23
L9	-0.35	+0.19	+0.84	+0.95	+0.29	+0.83	+0.36	+1.00	+0.25
K1	-0.48	+0.04	+0.31	+0.23	+0.14	+0.33	+0.23	+0.25	+1.00
K2	-0.50	+0.08	+0.44	+0.27	+0.17	+0.46	+0.34	+0.29	+0.97
K3	-0.52	+0.13	+0.60	+0.44	+0.21	+0.62	+0.41	+0.45	+0.91
R1	-0.34	+0.56	+0.51	+0.37	+0.89	+0.51	+0.92	+0.40	+0.18
R2	-0.37	+0.22	+0.91	+0.88	+0.23	+0.91	+0.42	+0.84	+0.24
R3	+0.37	-0.54	-0.57	-0.35	-0.87	-0.57	-0.98	-0.39	-0.22
R4	+0.39	-0.21	-0.98	-0.84	-0.24	-0.98	-0.49	-0.81	-0.29
R5	-0.34	+0.28	+0.69	+0.26	+0.21	+0.69	+0.57	+0.30	+0.26
P1	+0.33	-0.36	-0.71	-0.27	-0.25	-0.71	-0.63	-0.31	-0.25
P2	+0.32	-0.46	-0.54	-0.25	-0.25	-0.54	-0.50	-0.28	-0.21
D1	-0.45	+0.57	+0.51	+0.23	+0.37	+0.53	+0.59	+0.28	+0.43
D2	-0.46	+0.06	+0.69	+0.50	+0.13	+0.70	+0.37	+0.52	+0.54
X1	-0.45	+0.06	+0.30	+0.26	+0.21	+0.30	+0.26	+0.28	+0.63
X2	-0.38	+0.18	+0.22	+0.17	+0.36	+0.22	+0.37	+0.20	+0.49
M1	-0.22	+0.88	+0.14	+0.00	+0.45	+0.13	+0.48	+0.06	+0.03
M2	-0.28	-0.02	+0.62	+0.67	+0.08	+0.61	+0.17	+0.67	+0.21
M3	-0.40	+0.75	+0.71	+0.50	+0.47	+0.71	+0.66	+0.53	+0.20
Y1	-0.26	+0.16	+0.61	+0.68	+0.08	+0.61	+0.16	+0.77	+0.20
Y2	-0.21	+0.10	+0.42	+0.46	+0.42	+0.41	+0.41	+0.54	+0.16
Y3	+0.17	-0.02	-0.43	-0.46	+0.02	-0.43	-0.06	-0.56	-0.16
Y4	+0.20	-0.07	-0.14	-0.15	-0.10	-0.18	-0.11	-0.15	-0.38
Y5	-0.24	+0.14	+0.61	+0.69	+0.01	+0.62	+0.11	+0.75	+0.16
E1	+0.30	-0.21	-0.45	-0.20	-0.38	-0.46	-0.57	-0.30	-0.27

	K2	K3	R1	R2	R3	R4	R5	P1	P2
HR	-0.50	-0.52	-0.34	-0.37	+0.37	+0.39	-0.34	+0.33	+0.32
L3	+0.08	+0.13	+0.56	+0.22	-0.54	-0.21	+0.28	-0.36	-0.46
L4	+0.44	+0.60	+0.51	+0.91	-0.57	-0.98	+0.69	-0.71	-0.54
L5	+0.27	+0.44	+0.37	+0.88	-0.35	-0.84	+0.26	-0.27	-0.25
L6	+0.17	+0.21	+0.89	+0.23	-0.87	-0.24	+0.21	-0.25	-0.25
L7	+0.46	+0.62	+0.51	+0.91	-0.57	-0.98	+0.69	-0.71	-0.54
L8	+0.34	+0.41	+0.92	+0.42	-0.98	-0.49	+0.57	-0.63	-0.50
L9	+0.29	+0.45	+0.40	+0.84	-0.39	-0.81	+0.30	-0.31	-0.28
K1	+0.97	+0.91	+0.18	+0.24	-0.22	-0.29	+0.26	-0.25	-0.21
K2	+1.00	+0.96	+0.26	+0.33	-0.34	-0.42	+0.44	-0.45	-0.33
K3	+0.96	+1.00	+0.33	+0.49	-0.42	-0.59	+0.49	-0.53	-0.37
R1	+0.26	+0.33	+1.00	+0.49	-0.93	-0.50	+0.51	-0.50	-0.48
R2	+0.33	+0.49	+0.49	+1.00	-0.49	-0.93	+0.57	-0.53	-0.49
R3	-0.34	-0.42	-0.93	-0.49	+1.00	+0.57	-0.60	+0.65	+0.51
R4	-0.42	-0.59	-0.50	-0.93	+0.57	+1.00	-0.67	+0.69	+0.52
R5	+0.44	+0.49	+0.51	+0.57	-0.60	-0.67	+1.00	-0.92	-0.77
P1	-0.45	-0.53	-0.50	-0.53	+0.65	+0.69	-0.92	+1.00	+0.72
P2	-0.33	-0.37	-0.48	-0.49	+0.51	+0.52	-0.77	+0.72	+1.00
D1	+0.54	+0.57	+0.51	+0.41	-0.60	-0.50	+0.64	-0.71	-0.57
D2	+0.63	+0.71	+0.31	+0.60	-0.40	-0.69	+0.58	-0.59	-0.39
X1	+0.61	+0.57	+0.23	+0.23	-0.24	-0.26	+0.23	-0.21	-0.19
X2	+0.48	+0.44	+0.35	+0.17	-0.35	-0.18	+0.20	-0.19	-0.20
M1	+0.08	+0.08	+0.46	+0.09	-0.45	-0.09	+0.30	-0.35	-0.45
M2	+0.24	+0.35	+0.19	+0.61	-0.19	-0.59	+0.28	-0.24	-0.23
M3	+0.33	+0.44	+0.63	+0.63	-0.69	-0.69	+0.65	-0.73	-0.66
Y1	+0.23	+0.36	+0.17	+0.60	-0.19	-0.59	+0.22	-0.25	-0.21
Y2	+0.18	+0.25	+0.43	+0.41	-0.42	-0.40	+0.18	-0.17	-0.17
Y3	-0.19	-0.29	-0.04	-0.38	+0.08	+0.42	-0.14	+0.19	+0.09
Y4	-0.35	-0.34	-0.09	-0.12	+0.09	+0.12	-0.06	+0.06	+0.06
Y5	+0.19	+0.32	+0.16	+0.71	-0.17	-0.64	+0.23	-0.23	-0.24
E1	-0.37	-0.39	-0.50	-0.36	+0.58	+0.44	-0.59	+0.59	+0.42
E2	-0.30	-0.33	-0.42	-0.30	+0.49	+0.37	-0.48	+0.50	+0.36
E3	+0.40	+0.54	+0.80	+0.66	-0.88	-0.75	+0.58	-0.67	-0.51
E4	+0.41	+0.60	+0.49	+0.88	-0.54	-0.96	+0.56	-0.60	-0.46

	D1	D2	X1	X2	M1	M2	M3	Y1	Y2
HR	-0.45	-0.46	-0.45	-0.38	-0.22	-0.28	-0.40	-0.26	-0.21
L3	+0.57	+0.06	+0.06	+0.18	+0.88	-0.02	+0.75	+0.16	+0.10
L4	+0.51	+0.69	+0.30	+0.22	+0.14	+0.62	+0.71	+0.61	+0.42
L5	+0.23	+0.50	+0.26	+0.17	+0.00	+0.67	+0.50	+0.68	+0.46
L6	+0.37	+0.13	+0.21	+0.36	+0.45	+0.08	+0.47	+0.08	+0.42
L7	+0.53	+0.70	+0.30	+0.22	+0.13	+0.61	+0.71	+0.61	+0.41
L8	+0.59	+0.37	+0.26	+0.37	+0.48	+0.17	+0.66	+0.16	+0.41
L9	+0.28	+0.52	+0.28	+0.20	+0.06	+0.67	+0.53	+0.77	+0.54
K1	+0.43	+0.54	+0.63	+0.49	+0.03	+0.21	+0.20	+0.20	+0.16
K2	+0.54	+0.63	+0.61	+0.48	+0.08	+0.24	+0.33	+0.23	+0.18
K3	+0.57	+0.71	+0.57	+0.44	+0.08	+0.35	+0.44	+0.36	+0.25
R1	+0.51	+0.31	+0.23	+0.35	+0.46	+0.19	+0.63	+0.17	+0.43
R2	+0.41	+0.60	+0.23	+0.17	+0.09	+0.61	+0.63	+0.60	+0.41
R3	-0.60	-0.40	-0.24	-0.35	-0.45	-0.19	-0.69	-0.19	-0.42
R4	-0.50	-0.69	-0.26	-0.18	-0.09	-0.59	-0.69	-0.59	-0.40
R5	+0.64	+0.58	+0.23	+0.20	+0.30	+0.28	+0.65	+0.22	+0.18
P1	-0.71	-0.59	-0.21	-0.19	-0.35	-0.24	-0.73	-0.25	-0.17
P2	-0.57	-0.39	-0.19	-0.20	-0.45	-0.23	-0.66	-0.21	-0.17
D1	+1.00	+0.77	+0.34	+0.33	+0.54	+0.17	+0.71	+0.31	+0.09
D2	+0.77	+1.00	+0.42	+0.29	+0.04	+0.43	+0.46	+0.48	+0.18
X1	+0.34	+0.42	+1.00	+0.83	+0.17	+0.42	+0.27	+0.20	+0.23
X2	+0.33	+0.29	+0.83	+1.00	+0.29	+0.33	+0.30	+0.11	+0.22
M1	+0.54	+0.04	+0.17	+0.29	+1.00	+0.21	+0.76	+0.07	+0.06
M2	+0.17	+0.43	+0.42	+0.33	+0.21	+1.00	+0.51	+0.50	+0.36
M3	+0.71	+0.46	+0.27	+0.30	+0.76	+0.51	+1.00	+0.42	+0.28
Y1	+0.31	+0.48	+0.20	+0.11	+0.07	+0.50	+0.42	+1.00	+0.19
Y2	+0.09	+0.18	+0.23	+0.22	+0.06	+0.36	+0.28	+0.19	+1.00
Y3	-0.24	-0.42	-0.14	-0.06	+0.03	-0.37	-0.25	-0.87	+0.09
Y4	-0.21	-0.25	-0.14	-0.12	+0.00	-0.03	-0.07	-0.13	-0.08
Y5	+0.28	+0.46	+0.14	+0.05	+0.03	+0.47	+0.41	+0.86	+0.18
E1	-0.61	-0.56	-0.26	-0.25	-0.21	-0.19	-0.45	-0.37	-0.19
E2	-0.51	-0.45	-0.24	-0.21	-0.21	-0.18	-0.41	-0.33	-0.18
E3	+0.60	+0.53	+0.30	+0.33	+0.38	+0.40	+0.76	+0.49	+0.50
E4	+0.47	+0.67	+0.30	+0.21	+0.12	+0.63	+0.68	+0.64	+0.44

	Y3	Y4	Y5	E1	E2	E3	E4
HR	+0.17	+0.20	-0.24	+0.30	+0.28	-0.43	-0.41
L3	-0.02	-0.07	+0.14	-0.21	-0.21	+0.50	+0.25
L4	-0.43	-0.14	+0.61	-0.45	-0.39	+0.77	+0.96
L5	-0.46	-0.15	+0.69	-0.20	-0.18	+0.62	+0.89
L6	+0.02	-0.10	+0.01	-0.38	-0.32	+0.72	+0.29
L7	-0.43	-0.18	+0.62	-0.46	-0.39	+0.77	+0.96
L8	-0.06	-0.11	+0.11	-0.57	-0.48	+0.86	+0.50
L9	-0.56	-0.15	+0.75	-0.30	-0.30	+0.69	+0.87
K1	-0.16	-0.38	+0.16	-0.27	-0.21	+0.28	+0.31
K2	-0.19	-0.35	+0.19	-0.37	-0.30	+0.40	+0.41
K3	-0.29	-0.34	+0.32	-0.39	-0.33	+0.54	+0.60
R1	-0.04	-0.09	+0.16	-0.50	-0.42	+0.80	+0.49
R2	-0.38	-0.12	+0.71	-0.36	-0.30	+0.66	+0.88
R3	+0.08	+0.09	-0.17	+0.58	+0.49	-0.88	-0.54
R4	+0.42	+0.12	-0.64	+0.44	+0.37	-0.75	-0.96
R5	-0.14	-0.06	+0.23	-0.59	-0.48	+0.58	+0.56
P1	+0.19	+0.06	-0.23	+0.59	+0.50	-0.67	-0.60
P2	+0.09	+0.06	-0.24	+0.42	+0.36	-0.51	-0.46
D1	-0.24	-0.21	+0.28	-0.61	-0.51	+0.60	+0.47
D2	-0.42	-0.25	+0.46	-0.56	-0.45	+0.53	+0.67
X1	-0.14	-0.14	+0.14	-0.26	-0.24	+0.30	+0.30
X2	-0.06	-0.12	+0.05	-0.25	-0.21	+0.33	+0.21
M1	+0.03	+0.00	+0.03	-0.21	-0.21	+0.38	+0.12
M2	-0.37	-0.03	+0.47	-0.19	-0.18	+0.40	+0.63
M3	-0.25	-0.07	+0.41	-0.45	-0.41	+0.76	+0.68
Y1	-0.87	-0.13	+0.86	-0.37	-0.33	+0.49	+0.64
Y2	+0.09	-0.08	+0.18	-0.19	-0.18	+0.50	+0.44
Y3	+1.00	+0.08	-0.65	+0.31	+0.29	-0.35	-0.47
Y4	+0.08	+1.00	-0.14	+0.10	+0.08	-0.13	-0.15
Y5	-0.65	-0.14	+1.00	-0.35	-0.27	+0.41	+0.63
E1	+0.31	+0.10	-0.35	+1.00	+0.70	-0.54	-0.37
E2	+0.29	+0.08	-0.27	+0.70	+1.00	-0.51	-0.34
E3	-0.35	-0.13	+0.41	-0.54	-0.51	+1.00	+0.80
E4	-0.47	-0.15	+0.63	-0.37	-0.34	+0.80	+1.00

Appendix D

Software

This chapter describes the software and further necessary details to reproduce all experimental results in this work. As speech recognition toolkit HTK V3.2 was employed. File formats are explained in Section D.1. Sections D.2 and D.3 describe software libraries and scripts in RUBY, PYTHON and PERL written by the author or taken from public sources. Most scripts display a help message if they are executed on the commandline without any arguments. Look at the HISTORY files and scripts in the `scripts` directory in each experiment directory for usage examples. In Section D.5 the connections between experiment directories and experimental results shown in tables and figures are given. Section D.6 describes the implementation of the `pronfex` module.

D.1 File formats

Input and output files used by various scripts and libraries have to obey a certain format. The abbrevations of the file formats are:

- MLF: HTK-style Master Label File
- UAF: Unlabeled Ascii Feature file
- LAF: Labeled Ascii Feature file
- PDS: Phone(me) Duration Statistic file
- PCS: Phoneme Confusion Statistic file
- GDP: Gaussian Density Parameter file
- GMP: Gaussian Mixture Parameter file
- LRP: Linear Regression Parameter file
- CRF: Classifier Result File

- LTP: LDA Transformation Parameter file
- KTP: KLT Parameter file
- WTP: Whitening Transform Parameter file
- PLM: Phoneme Language Model file
- LAB: Label file
- RES: Result file
- APF: Adjustment Parameter File

The file formats are described in the following subsections. `<string>` denotes variables, `{<string>}` none or multiple repetitions of `<string>`, `(<string>)` one or more repetitions of `<string>` and values in `[<string>]` are optional. Other symbols do not have a special meaning.

D.1.1 MLF

MLF are generated by `HVite` for the forced-alignment or recognition results. An MLF contains one or more entries of the form:

```
''<path>/<file>.<suffix>''
{<segment>}
.
```

Instead of specifying the `<path>` the wildcard `*` can be used. `<file>` is the name of the acoustic features file. The `<suffix>` is usually `lab` for reference label files and `rec` for files containing of phoneme or word recognition output. Each `<segment>` has the format:

```
[<start> [<end>]] (<name> [<score>]) [<comment>]
```

`<start>` and `<end>` are the starting and ending times in $\frac{1}{10}$ microseconds of the acoustic segment, `<name>` and `<auxname>` a state, (phoneme) model or word label and `<score>` the corresponding acoustic score. However, scripts and binaries for extraction of pronunciation features, which are described later, assume the following format:

```
[<start> <end>] [<name1> [<score1> [<name2>]]]
```

`<name1>` and `<name2>` are phoneme or word labels.

D.1.2 UAF

An unlabeled ASCII features file has one or more lines in the following format:

```
(<featval>)
```

`<featval>` is the value of a feature.

D.1.3 LAF

A labeled ASCII feature file has one or more lines in the following format:

```
<class> (<featval>)
```

`<class>` is the class label of each pattern.

D.1.4 PDS

A file containing a phone(me) duration statistic has a header of the form

```
DTYPE <type>
```

followed by one or more lines in the format

```
<symbol> <count> <parameter1> <parameter2>
```

`<symbol>` is the phone(me) symbol and `<count>` the occurence frequency. `<type>` can have the values `log` for log-normal density parameters, `normal` for normal density parameters and `hist` for histogram parameters. However, the pronfex module does not support the format `hist`. A phoneme duration statistic can be obtained from a MLF with the script `phondurstat.rb`.

D.1.5 PCS

A phone(me) confusion statistic file contains phone(me) labels in the first line. These are followed by the phoneme confusion probabilities. There is one line per row of the confusion matrix. Two PDS files, one containing the phoneme confusion matrix for correctly pronounced words (`O`), the other for mispronounced words (`X`), can be generated from a forced-alignment MLF, a recognition MLF with phoneme segment labels and files containing word labels and the corresponding markings with `O` and `X` by the script `confusionmatrix.rb`.

D.1.6 GDP

A Gaussian density parameter file contains the parameters μ and Σ for one multivariate Gaussian. The header format is

```
GAUSS <type>
<dim>
```

<type> may be FULL, for a full covariance matrix Gaussian, or DIAG for a diagonal covariance matrix. <dim> is the feature dimension. The header is followed by the mean vector μ and the full covariance matrix Σ or only its diagonal. A GDP file can be generated with nvk-make.rb.

D.1.7 GMP

A Gaussian mixture model parameter file contains the parameters of one or more multivariate Gaussians and mixture weights. The header format is

```
GAUSSMIX <type>
<dim> <ndens>
(<weight>)
```

<type> has the same meaning as for the GDP file format. <dim> is the feature dimension, <ndens> the number of mixture components. There are as many weights <weight> as there are mixture components. The header is followed by the mean vectors μ_i and the covariance matrix Σ_i of each Gaussian. GMP files can be generated with gmm-train.py and can be used by gmm-test.py for classification.

D.1.8 LRP

A linear regression parameter file contains the parameters a_i of a linear regression function $f(\boldsymbol{x}) = \sum a_i x_i$. There is one line per parameter starting with the bias term. Such a parameter file can be generated by linreg-estimate.rb and can be used by linreg-transform.rb.

D.1.9 CRF

A classifier result file contains the classification result generated by scripts nvk.rb or gmm-test.py. There is one line per classified sample. Each line has the format

```
<class> (<score>)
```

<class> is the class index starting from 1 up to the number of classes. The class index is followed by the logarithmic probabilities <score>, i.e. the values of the log-likelihood probability function for each class.

D.1.10 LTP

A LDA transformation parameter file contains the Eigenvalues in descending order from the LDA in the first line and the Eigenvectors of the LDA transformation in the following lines. A LTP file can be generated by the script `lda-train.py` from one LAF or multiple UAFs. It is employed by `lda-test.py` to transform a LAF or multiple UAFs.

D.1.11 KTP

A KLT parameter file contains the Eigenvalues in descending order from the PCA in the first line, the source and target dimensions in the second line, and the Eigenvectors of the KLT in the following lines. A KTP file is generated by the script `pca-train.py` from one UAF or HTK-style feature files.

D.1.12 WTP

A whitening transform parameter file contains a mean vector `<mean>` and the vector of standard deviations `<stddev>` of each component. This file can be generated with the script `wht-train.py` from a sample.

```
m <mean>
s <stddev>
```

D.1.13 PLM

The phoneme language model file in HTK ARPA format.

D.1.14 LAB

A label file contains reference labels. There is one label per line and sample. The labels may have any value. By combining a LAB and a UAF file with the UNIX command `paste`, a LAF file is generated.

D.1.15 RES

A result file contains classification or scoring results. There is one line per sample. Result files are the output of the scripts `linreg-transform.rb`, `softscore.rb` and `adjustscores.rb`.

D.1.16 APF

An adjustment parameter file contains the parameters of the linear and the multiplicative polynomial transformation for score adjustment. The parameters can be estimated with `adjustscores.rb` with a reference label file (`LAB`) and a score result file (`RES`).

D.2 Libraries

Table D.1 gives an overview to libraries written in RUBY or PYTHON consisting of functions often used in various scripts.

D.3 Scripts

The names and purposes of scripts are shown in Tables D.2, D.3, D.4 and D.5. The scripts 101-105 call HTK Tools for feature extraction, acoustic model training, etc. Scripts 106-113 are for editing MLF or extracting information from MLF files. Scripts 201-209 are for processing features, e.g. LDA, PCA or whitening. Scripts 301-307 are for training/testing based on a Gaussian classifier and for Linear Regression. The scripts 501-520 in Table D.5 are for stream-based processing of data files, i.e. they read lines from standard input and write lines to standard output.

D.4 Speech Recognizer

HTK V3.2 is employed for acoustic model training and speech recognition. Table D.6 shows the experiment directories w.r.t. each trained acoustic model.

The topology of each model is the same. There is one 3-state HMM for each of the monophones 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, oy, p, r, s, sh, t, th, uh, uw, v, w, y, z and zh. Moreover, there is a combined silence HMM sil/sp and a 1-state short pause HMM sp. Each state has a 16 Gaussian mixture density with diagonal covariance matrix. The HTK feature kind employed is MFCC_E_D_A_Z with 39 acoustic features in total. The HTK configuration file for acoustic feature extraction is

```
SOURCEFORMAT = NOHEAD
SOURCEKIND   = WAVEFORM
SOURCERATE   = 625
TARGETFORMAT = HTK
TARGETKIND   = MFCC_E_D_A_Z
TARGETRATE   = 100000
WINDOWSIZE   = 200000
USEHAMMING   = T
PREEMCOEF    = 0.97
NUMCHANS     = 26
CEPLIFTER    = 22
NUMCEPS      = 12
SAVEWITHCRC  = F
DELTAWINDOW  = 2
ACCWINDOW    = 2
THIRDWINDOW  = 2
```

Table D.1: *RUBY and PYTHON software libraries.*

ID	Lib-Name	Author/Status	Purpose
L1	uci.rb	T.Cincarek	Command line parsing
L2	misc.rb		Miscellanenous functions
L3	htk2.rb		HTK related stuff
L4	matvec.rb		Matrix and vector functions
L5	log.rb		Simple logging facility
L6	spkdata.rb		SLT Non-native DB speaker information
L7	pronscore.rb		Some functions for pronunciation FEX
L8	matrix-algebra.rb	S.Hara	Matrix algebra
L9	numarray.py	Open Source	Matrix algebra V0.7
L10	stats.py	T.Cincarek	Mean, Co-Variance, Correlation, a.s.o.
L11	vq.py		Clustering algorithms (k-means, LBG)
L12	gmm.py		GMM training (EM and FJ algorithm)
L13	pca.py		Principal Component Analysis (PCA)
L14	lda.py		Linear Discriminant Analysis (LDA)
L15	misc.py		Miscellaneous functions
L16	uci.py		Command line parsing

Table D.2: *Scripts for HTK Tools and MLF.*

ID	Script	Purpose
101	htk-fex.rb	Acoustic feature extraction (HTK)
102	htk-buildmonophAM.rb	Monophone AM training from scratch (HTK)
103	htk-makelm.rb	Build language model (HTK)
104	htk-phonerecognizer.rb	Phone(me) recognition (HTK)
105	htk-lmscore.rb	Calculate LM score (HTK)
106	mlf-getcolumn.rb	Get certain columns from MLF
107	mlf-stripsilsp.rb	Strip sp/sil symbols from MLF
108	mlf-edit.rb	Edit MLF (delete, insert, substitute)
109	mlf-getid.rb	Get utterance IDs from MLF
110	mlf-score.rb	Get AM score from MLF; calculate LM score for MLF
111	mlf-nbestfex.rb	Pronunciation feature extraction (C06-C08)
112	mlf-wordfex.rb	Pronunciation feature extraction (W01-W27, C01-C05)
113	mlf-utrfex.rb	Pronunciation feature extraction (Utterance)

Table D.3: *Scripts for feature processing.*

ID	Script	Purpose
201	pca-train.py	PCA (Eigenvector estimation)
202	pca-test.py	KLT
203	lda-train.py	LDA (Eigenvector estimation)
204	lda-test.py	LDA (Transformation)
205	wht-train.py	Whitening (Mean/Variance estimation)
206	wht-test.py	Whitening (Transformation)
207	selectfeatures.rb	Select features by index or name from LAF/UAF
208	combinedata.rb	Combine multiple UAFs to one LAF
209	splitdata.rb	Split one LAF into multiple UAFs; optional resampling

Table D.4: *Scripts for classification.*

ID	Script	Purpose
301	gmm-train.py	GMM Training
302	gmm-test.py	GMM Classification
303	nvk.rb	Gaussian classifier; floating search
304	nvk-make.rb	Build Gaussian classifier
305	linreg-estimate.py	Estimation of LR function
306	linreg-transform.rb	Application of LR function
307	linreg-feateval.rb	Feature evaluation with LR

Table D.5: *Scripts for various purposes.*

ID	Script	Purpose
401	interpolate.rb	Polynome coefficients with Newton's method
402	cor_bs_confint.rb	Confidence interval of correlation with Bootstrap
403	correlation.rb	Calculate correlation, open correlation
404	performance.rb	Performance for binary classification results
405	softscoreprec.rb	Calculate performance of soft-scoring results
406	result.rb	Calculate confusion matrix, gain, CL, RR
407	recrate.pl	Calculate confusion matrix, gain, CL, CL-A, RR
408	shufflelist.rb	Reorder input lines randomly
409	confinterval.rb	Calculate the confidence interval for the mean
501	col.rb	Grep the n-th column
502	exp.rb	Calculate the e-function
503	format.rb	Format strings
504	docmd.rb	Execution of a command for several arguments
505	round.rb	Rounding of float values to integers
506	sort.pl	(Correct) sorting of floats
507	words.rb	Split lines into word tokens
508	split.rb	Split strings into single characters
509	log.rb	Calculate the logarithm
510	sum.pl	Sum up values
511	mincol.pl, maxcol.pl	argmin, argmax operator
512	geomean.pl	Calculate geometric mean
513	stripnl.rb	Strip newline characters at end of lines
514	abs.pl	Take the absolute value
515	mean.pl	Calculate the mean
516	stddev.pl	Calculate standard deviation
517	variance.pl	Calculate variance
518	sqrt.rb	Take square root
519	upcase.rb	Capitalize input strings
520	downcase.rb	convert to lower case letters

Table D.6: *Acoustic models.*

Model	Directory	Training data (Corpus)
AE	baseline_mono	WSJ0/WSJ1
BE	baseline_mono_be	WSJCAM0
PF_STAR BE	baseline_children	PF_STAR BE

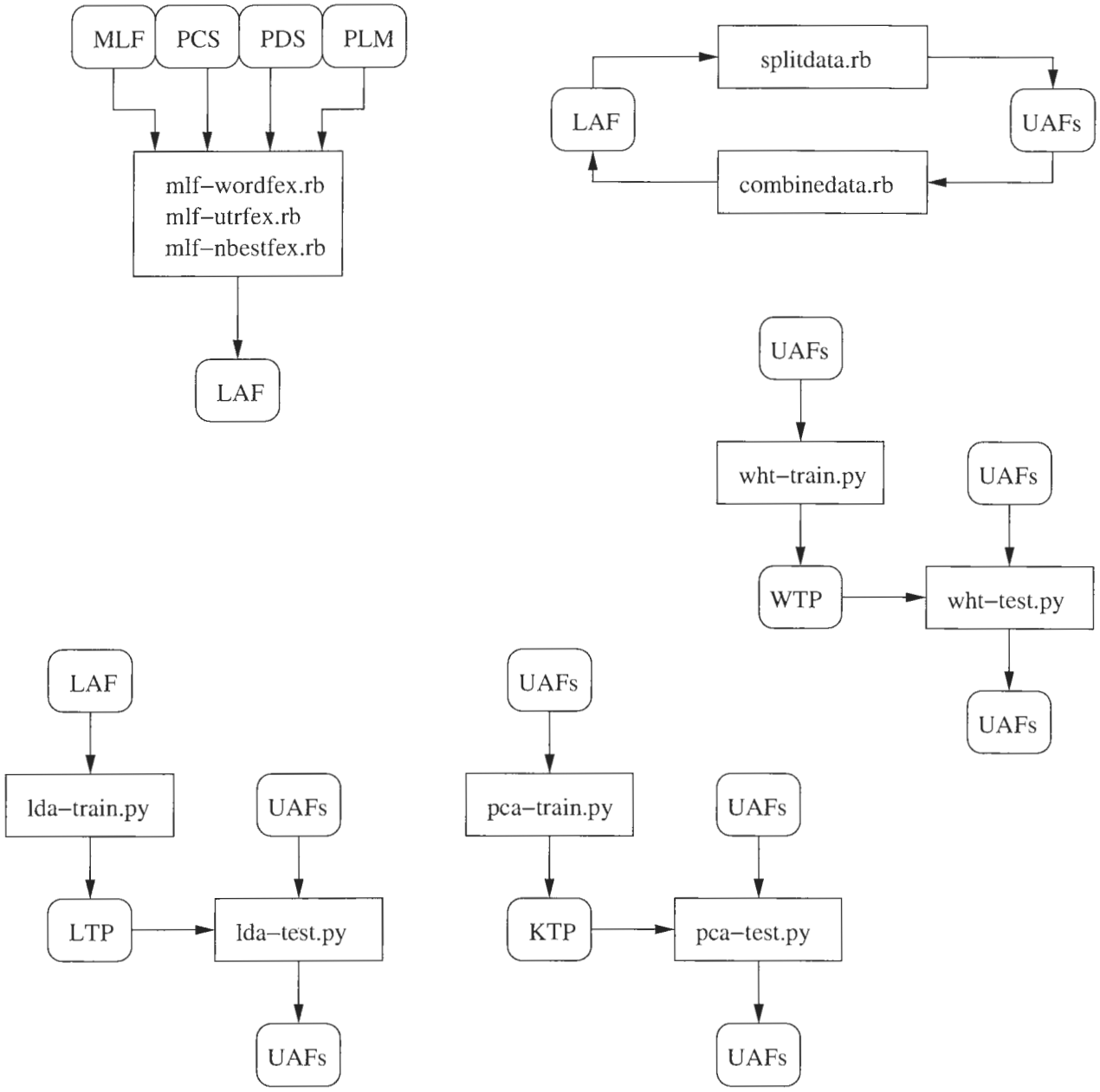


Figure D.1: *Scripts for feature extraction and feature processing.*

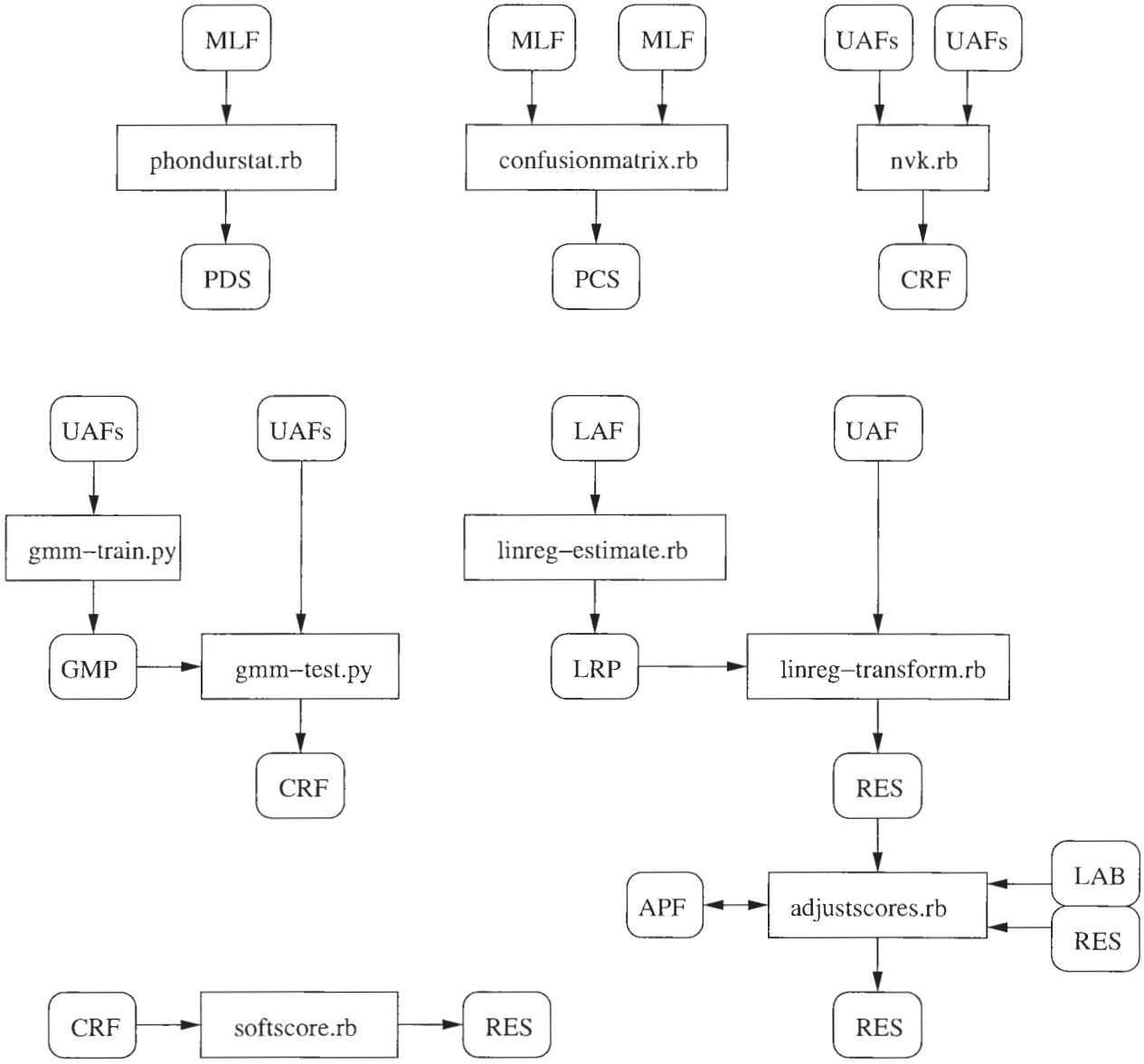


Figure D.2: *Scripts for classification and scoring.*

Additionally, the byte order has to be specified w.r.t. the source data with

```
BYTEORDER     = <order>
```

where `<order>` is `VAX` for little-endian and `SUN` for big-endian. The byte order of the training corpora as shown in Table D.6 is little-endian and big-endian for the ATR SLT data. For feature normalization by VTLN the following lines must be added to the configuration file:

```
WARPFREQ      = <alpha>
WARPLCUTOFF   = 10.0
WARPUCUTOFF   = 8000.0
```

`<alpha>` is the warping factor.

For word recognition a beam size of `200.0`, for unconstrained phoneme recognition a beam size of `500.0` was used, respectively. N-best recognition was carried out for N=100 (`HVite` argument `-n 10 100`).

The directory `asr` contains all necessary acoustic models, configuration files and grammar files for phoneme recognition with the script `htk-phonerecognizer.rb` based on HTK.

The content of the grammar file for unconstrained phoneme recognition is as follows:

```
$phones = 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 | t | th | uh | uw | v | w | y | z ;

(
sil < $phones [sp | sil] > sil
)
```

The grammar can be converted into a recognition network with `HParse`.

D.5 Sources of Tables and Figures

Tables D.7 and D.7 show the names of the directories in which the experiments to generate the results arranged in Tables and Figures were carried out. References to history files, result files and scripts are given.

Table D.7: *Experiment directories for the results shown in all tables (T) and figures (F).*

T/F	Directories	Commands, scripts and/or files
T4.3	lme_adult	HISTORY, spk.{rating,misratio,vtln_wordacc,vtln_phonacc}
F4.1	lme_adult	HISTORY, spk.{rating,misratio}
T5.1	cgi_data	evalres_miscor.rb
T5.2	cgi_data	evalres_utrcor.rb
T5.3	cgi_data	evalres_spkcor.rb
T5.4	cgi_data	rat.{strict,cor_miscnt_utrlab}
T5.5	cgi_data	HISTORY, mispron_hitlist.X, mispron_hitlist_norm.X
T5.6	cgi_data	HISTORY, rat.{strict,wordcor,utrcor,spkcor}
T5.7	cgi_data	HISTORY, evalres_mispronstat_{amean,gmean}.rb
T5.8	cgi_data	HISTORY, evalres_mispronstat_{chain}.rb
T5.9	cgi_data	mispron_phonestat.{C,F,G,I,J}
F5.1	nn_database	HISTORY, README, utrlevel_confint_{0.05,0.1}.data
F5.2	slt_utrcorrelation	HISTORY, spklevel_confint_{0.05,0.01}.data
F5.3	nn_database	HISTORY, README, utrlevel_jn_confint_{0.05,0.1}.data
F6.4	native_stats	HISTORY, phondurstat_{k,ah}_native_{hist,func}.eps
F6.5	slt_utrexp_lr	score_gallery.csh; {X1,K1}_distribution.eps
F6.6	slt_utrcorrelation	HISTORY, analyze_evaluators.rb
F6.7	cgi_data phone_recognition word_recognition	spk.{meanratings,phacc_free,wa_free} phonrec_free_mono.rb eval_zerogram.rb
F6.8	slt_utrcorrelation	HISTORY, analyze_spkcor_avgutr.rb
T6.3	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.4	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.6	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.7	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.8	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.9	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.10	slt_utrcorrelation	RESULTS_UTRCOR, analyze_utrcor.rb
T6.12	cgi_data	spk.{meanratings,mispronratio,phacc_free,wa_free}
T6.13	slt_utrcorrelation	RESULTS_UTRCOR, RESULTS_SPKCOR
F7.3	nn_database	analyze_rating2miscnt.rb, polynome.eps
T7.1	nn_database	analyze_rating2miscnt.rb
F8.1	slt_utrexp_lr	lr_eval.csh K3,X1,M3,D2; rating_score.eps
T8.4	slt_utrexp_lr	HISTORY, lr_eval.csh, cv_lr_eval.csh
T8.5	slt_native_utr	README, eval.csh K3,X1,M3,D2, nvk_eval.csh E3,K2,X1
T8.6	slt_utrexp_fs	HISTORY, RESULT_FS_{COR,GAINMATRIX1,GAINMATRIX2}
T8.7	slt_utrexp_fs	HISTORY, {nvk,cv_nvk}_eval.csh, cv_result_table.csh
T8.8	slt_utrexp_fs	nvk_eval.csh E3,K2,X1

Table D.8: *Experiment directories for the results shown in all tables (T) and figures (F).*

T/F	Directories	Commands, scripts and/or files
F8.2	slt_utrexp_lr	spkscore.csh E3,K2,X1
T8.9	slt_utrexp_lr	spkscore1.csh, spkscore2.csh
T8.10	slt_wordexp_fs	HISTORY, RESULT_FS, RESULT_FS_GAINMATRIX
	slt_wordexp_cart	one/RESULT_*
T8.11	slt_wordexp_alt	HISTORY, RESULT_FS, RESULT_FS_GAIN
	slt_wordexp_cart	alt/RESULT_*
	slt_wordexp_gmm	pca.csh, lda.csh, gmm_{pca,lda}_train_test.csh
T8.12	slt_wordexp_alt	cv_nvk_eval.csh
	slt_native_word	HISTORY, eval.csh, nvk_eval.csh
F8.3	slt_wordexp_alt	recall.csh, precision.csh
T8.13	nn_database	RESULT_WORDMARK, evalperf_wordmark.rb
T8.14	slt_wordexp_cls	HISTORY, RESULT_FS_GAIN
T8.15	lme_adult	HISTORY_VTLN, best_{wordrec,phonrec}.rb
	lme_adult_be	HISTORY_VTLN, best_{wordrec,phonrec}.rb
	lme_child	word,phoneme_recognition.rb
F8.4	lme_utrexp	analyze_spkcor_avgutr.rb
T8.16	lme_utrexp	analyze_spkcor_avgutr.rb
	slt_utrcorrelation	RESULTS_SPKCOR, analyze_spkcor.rb
T8.17	lme_wordexp	HISTORY
T8.18	lme_utrexp	HISTORY, spkscore1.csh, spkscore2.csh
F8.5	lme_utrexp	spkscore.csh X1,K1,L7
T8.19	lme_wordexp	HISTORY, RESULT_FS_GAIN
	lme_wordexp/gmm	gmm_lda_test.csh
	lme_wordexp/cart	HISTORY
F8.6	lme_wordexp	precision.csh, recall.csh
T8.20	lme_native_utr2	results_nvk.csh
T8.21	lme_native_utr3	HISTORY

D.6 Pronfex Module

D.6.1 Implementation

The calculation of pronunciation features on word and utterance level with the scripts `mlf-utrfex.rb`, `mlf-wordfex.rb` and `mlf-nbestfex.rb` was re-implemented in C++ as the command line tool `pronfex`. Besides the utterance level feature Y1-Y5, all other word and utterance features can be calculated with `pronfex`. The extraction of the utterance features `M1-M3` and the word features `W14-W16` is based on the HTK tool `LPlex`, i.e. it is employed to calculate phoneme sequence probabilities. Additionally, a few frame level and phoneme level features can also be calculated. The implementation of the `pronfex` module consists of the source files as given in Table D.9. There is a `Makefile` in the source directory. Just execute `make` on the command line and the binary will be build. Compilation was successfully tested with the gcc 3.2.2 compiler.

Source Files	Short Description
lattice.{cc,h}	Read HTK MLF files (forced-alignment, recognition) Read HTK MLF files (N-best word recognition) Calculate Word Posterior Probability (WPP)
main.{cc,h}	Parse command line arguments Read configuration file
mlf.{cc,h}	Determine number of utterances in MLF Determine number of N-best hypothesis in MLF
phonstat.{cc,h}	Read phoneme confusion statistic Read phoneme duration statistic
score.{cc,h}	Frame level feature extraction Phoneme level feature extraction Word level feature extraction Utterance level feature extraction
util.{cc,h}	Helper functions
typedefs.h	Type definitions
Makefile	Makefile for building the `pronfex` binary

Table D.9: *Source files belonging to the implementation of the pronfex tool.*

Additionally to the source files of the implementation, the directory contains also examples of the files necessary in order to extract pronunciation features. There are three MLFs `alignment.mlf`, `recognition.mlf` and `nbestrec.mlf` containing the result of the forced-alignment, the word recognition with phone(me) segment labels, and N-best word recognition. Furthermore there is a configuration file `fex.cfg`, two confusion matrices `confmatrix.O` and `confmatrix.X` in PCS format, a phoneme duration statistic (PDS) `phondurstat.native` and a phone(me) bigram language model `phonbigram.arpa` in HTK ARPA format. By executing the command

```
pronfex -a alignment.mlf -r recognition.mlf -n nbestrec.mlf
        -m phonbigram.arpa -p phondurstat.native -x confmatrix
        -c fex.cfg -l <level>
```

pronunciation features can be extracted for the example files at frame [1], phone(me) [2], word [3] or utterance level [4]. The extracted features and additional information is written to standard output. Depending on the `<level>` of feature extraction, the format of the output changes. The meaning of each output column is summarized in the following Table:

Column	Feature Extraction Level			
	Frame [1]	Phone(me) [2]	Word [3]	Sentence [4]
1	Frame number	Beginning Frame	Beginning Frame	Utterance ID
2	Utterance ID	Ending Frame	Ending Frame	# words
3	Word (alignment)	Utterance ID	Utterance ID	# phone(me)s
4	Phoneme (alignment)	Word (alignment)	Word (alignment)	feature L3
5	Phoneme (recognition)	Phoneme (alignment)	feature W01	feature L4
6	Confidence score [C]	Acoustic score [L]	feature W02	feature L5
7	GOP score [K]	GOP score [K]	feature W03	feature L6
8	-	Actual Duration	feature W04	feature L7
9	-	Expected Duration	feature W05	feature L8
10	-	Duration score [D]	feature W06	feature L9
11	-	Confidence score [C]	feature W07	feature K1
12	-	-	feature W08	feature K2
13	-	-	feature W09	feature K3
14	-	-	feature W10	feature R1
15	-	-	feature W11	feature R2
16	-	-	feature W12	feature R3
17	-	-	feature W13	feature R4
18	-	-	feature W14	feature R5
19	-	-	feature W15	feature P1
20	-	-	feature W16	feature P2
21	-	-	feature W17	feature D1
22	-	-	feature W18	feature D2
23	-	-	feature W19	feature X1
24	-	-	feature W20	feature X2
25	-	-	feature W21	feature M1
26	-	-	feature W22	feature M2
27 ...	-	-	feature W23	feature M3
31	-	-	feature W27	-
32 ...	-	-	feature C01	-
39	-	-	feature C08	-

D.6.2 Usage

```
Pronunciation feature extraction on multiple levels
(c) 2004 Tobias Cincarek, FAU LME, ATR SLT

Usage: ./pronfex

  -l <int:level>    1:frame, 2:phon, 3:word, 4:utter
  -c <file:config>  Configuration for feature extraction
  -a <mlf:align>    HTK MLF file (forced-alignment)
  -r <mlf:recog>    HTK MLF file (recognition)
  -n <mlf:nbest>    HTK MLF file (N-best word)
  -m <arpa:phlm>    Phoneme LM in HTK ARPA format
  -p <file:stat>    Phoneme duration statistic
  -x <file:stat>    Phoneme confusion statistics {.O,.X}
  -d <int:level>    Verbose/debug level
```

-l Specifies the level of feature extraction. 1: frame level feature extraction, 2: phone(me) level feature extraction, 3: word level feature extraction, 4: utterance level feature extraction

-c Specifies the name of the configuration file.

-a Specifies the name of the HTK-style MLF containing the result of the forced-alignment. To generate the MLF, `HVite` has to be executed with option `-m` in order to obtain not only an alignment at the word but also at the phone(me) level.

-r Specifies the name of the HTK-style MLF containing the result of word or phoneme recognition. To generate the MLF, `HVite` has to be executed with option `-m` in order to obtain not only the word hypotheses but also the corresponding phone(me) segments.

-n Specifies the name of the HTK-style MLF containing the result of N-best word recognition. To generate the MLF, `HVite` has to be executed with e.g. `-n 10 100` for 100-best word recognition.

-m Phone(me) language model (LM) in HTK ARPA format for phoneme sequence probability computation (Not yet implemented).

-p Specifies the name of the PDS file with a phoneme duration statistic, which can be generated with `phondurstat.rb`.

-x Specifies the `<basename>` of the PCS files `<basename>.O` and `<basename>.X` with a phoneme confusion statistic for correctly pronounced words (O) and mispronounced words (X). This statistic can be generated with `confusionmatrix.rb`.

-d Debug level. 1: file I/O, 2: parameter values, 3: lattice, 4: features

D.6.3 Usage Example

In the following the necessary steps to extract pronunciation features for utterances and words are explained by an example. The first step is to obtain the forced-alignment and do phoneme and or word recognition on the utterance. An acoustic model file <model> and a file containing the names of all phoneme models <phonlist> have to be available. If <flist> is a file with the name of raw speech files or feature files on each line, <lexicon> the file which contains the pronunciation lexicon, i.e. the mapping of words to sequences of HMM model names, and <label> the MLF with the word level utterance transcriptions, the alignment can be computed by

```
HVite -a -m -b '<silence>' -y lab -i <MLF>
      -C <config> -H <model> -I <label> <lexicon> <phonlist>
```

<silence> is the name of the word always aligned at the beginning and at the end of each utterance. It must be mapped to a silence model in the pronunciation dictionary <lexicon>. The file <config> contains the configuration parameters especially w.r.t. the form of the input speech files. The commands for recognition and N-best recognition are

```
HVite -m -w <network> -i <MLF>
      -C <config> -H <model> <lexicon> <phonlist>

HVite -m -w <network> -i <MLF> -n 10 100
      -C <config> -H <model> <lexicon> <phonlist>
```

The file <network> contains the recognition network. It differs for phoneme recognition and word recognition with and without a language model. An example of how to generate a <network> file for unconstrained phoneme recognition is given in Section D.4. Any results, either for alignment or for recognition, are written to the <MLF> specified with option -i.

For pronunciation feature extraction three additional models are required: A phoneme duration statistic, a phoneme duration statistic for correctly pronounced and mispronounced words and a phoneme bigram LM trained on native data.

The phoneme duration statistic (PDS) can be estimated with phondurstat.rb. Let <flist> be a file which contains the names of MLFs with timing information, i.e. there are three colums, the first two for the beginning and the ending time of each phoneme segment and one column for the phoneme symbol itself. A statistic for each phoneme with the parameters of a log-normal distribution can then be generated with the command

```
phondurstat.rb -r -m <flist> -n <factor> -f <phonstat> -d log
```

Option -r means ROS normalization. <factor> is the duration normalization factor to convert segment times to milliseconds. <phonstat> is the name of the output file containing the phoneme duration statistic. Confer Section D.1 for the file format of <phonstat>.

The phoneme confusion statistics (PCS) for correctly pronounced and mispronounced words can be generated with the command

```
confusionmatrix.rb -a <flist_align>
                   -r <flist_recog>
                   -w <labeldir>
                   -p <phonlist>
                   -m <confstat>
                   -t <factor>
                   -e <markcnt>
```

<flist_align> contains the names of MLFs with phoneme and word level alignment information, <flist_recog> with the recognition result including phoneme segment information, <phonlist> the names of all phoneme symbols, <factor> is the normalization factor to convert segment times to frame indices, <markcnt> is the number of times a word had to be marked by a human evaluator to be considered as mispronounced and <labeldir> is a directory which contains one file per utterance in the MLFs with word level mispronunciation information. Each file in <labeldir> has as many lines as there are words in the reference transcription of the utterance. Each line has the format

```
<word> <misproninfo>
```

<misproninfo> is a string consisting of the characters O and X, one character per human evaluator. The <word> is considered as mispronounced, if there are at least as many characters X as specified by option -e <markcnt>. Two phoneme confusion matrix files are written: <confstat>.O for correctly pronounced words and <confstat>.X for mispronounced words. The matrices are estimated from the frequencies of phoneme confusions at the frame level. Confer Section D.1 for the file format of <confstat>.*.

The phoneme bigram language model can be generated by executing the command

```
htk-makelm.rb -l <network>
              -b <phonlm>
              -w <phonlist>
              -m <MLF>
              -s
```

The file <phonlist> contains the list of phonemes to be covered by the language model. The training data can be specified directly by -m <MLF> or by -f <flist> containing a list of MLFs. There are two output files: A <network> for recognition and an n-gram statistics file <phonlm> in ARPA format, which can be used for calculating phoneme sequence probabilities with the HTK Tool LPlex.

Given the three models (PDS in file <phonstat>, PCS in files <confstat>.O and <confstat>.X, and the phoneme bigram LM in file <arpa>) and at least the MLF of the forced-alignment <mlf_align> with phoneme segment information, pronunciation features can be extracted with pronfex. For example, features at the word level are extracted with the command

```
pronfex -l 3
        -c <config>
        -a <mlf_align>
        -r <mlf_recog>
        -n <mlf_nbest>
        -m <phonlm>
        -p <phonstat>
        -x <confstat>
```

An example of the `pronfex` configuration file `<config>` is

```
factor_2ms      0.0001
frameshift      10.0
sp_symbol       sp
sil_symbol      sil
utt_start       !ENTER
utt_end         !EXIT
wpp_lhscale     0.01
wpp_overlap     0.5
```

`factor_2ms` is the normalization factor to convert segment times to milliseconds. With `frameshift` the window shift can be specified in milliseconds, in order to convert segment times given in milliseconds to frame indices. `sp_symbol` specifies the short pause symbol between words and `sil_symbol` the silence symbol. With `utt_start` and `utt_end` the utterance start and end symbol can be set up, respectively. These symbols have to be present at the beginning and end of each utterance in the `<mlf_align>` file in case of feature extraction at the word level. `wpp_lhscale` is a factor to scale likelihoods for word posterior probability (WPP) computation. `wpp_overlap` defines the necessary degree of overlap between words in order to be considered for WPP computation. Valid values for `wpp_overlap` are between `0.0` and `1.0`, but it should be equal to or greater than `0.5`.

If the MLF with the recognition result and/or the MLF with the N-best recognition result are not available, just specify the MLF `<mlf_align>` of the forced-alignment again for the commandline options `-r` and/or `-n`. Unneccesary to mention, that the features based on the (N-best) recognition result become meaningless and should neither be employed for scoring nor for classification.

List of Figures

List of Tables

Bibliography

[BS97] Ilja N. Bronstein and Konstantin A. Semendjajew. *Taschenbuch der Mathematik.* Harri Deutsch, Frankfurt am Main, Thun, 1997.

[CD02] Stephen Cox and Srinandan Dasmahapatra. High-level approaches to confidence estimation in speech recognition. *IEEE Transactions on Speech and Audio Processing*, 10(7):460–471, 2002.

[CGN] Tobias Cincarek, Rainer Gruhn, and Satoshi Nakamura. Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. To appear.

[Cro51] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 1951.

[CSB97] Catia Cucchiarini, Helmer Strik, and Lou Boves. Automatic evaluation of dutch pronunciation by using speech recognition technology. 1997.

[CSB00] Catia Cucchiarini, Helmer Strik, and Lou Boves. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30:109–119, 2000.

[CSBB00] Catia Cucchiarini, Helmer Strik, Diana Binnenpoorte, and Lou Boves. Pronunciation evaluation in read an spontaneous speech: A comparison between human ratings and automatic scores. 2000.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* John Wiley & Sons, Inc., 2001.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J.R. Statistical Society*, (39):1–38, 1977.

[DWR] Shona M. D'Arcy, Lit P. Wong, and Martin J. Russsel. Recognition of read and spontaneous children's speech using two new corpora. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. To appear.

[ET93] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, 1993.

[FJ02] Mario A.T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[FNDR00] Horacia Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30:121–130, 2000.

[GM00] Ben Gold and Nelson Morgan. *Speech and Audio Signal Processing*. John Wiley & Sons, Inc., 605 Third Avenue, New York, USA, 2000.

[IPA99] *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[Jel97] Frederick Jelinek. *Statistical Methods for Speech Recognition*. Massachusetts Institute of Technology, 1997.

[LR98] Li Lee and Richard Rose. A frequency warping approach to speaker normalization. In *IEEE Transactions on Speech and Audio Processing*, volume 6, pages 49–59, 1998.

[Min04] Nobuaki Minematsu. Automatic scoring of language learners' pronunciations based on the distortion of their universal structures. In *The institute of electronics, information and communication engineers (IEICE)*, 2004.

[NCSB02] Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15:441–447, 2002.

[NFDW00] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 30:83–93, 2000.

[Nie03] Heinrich Niemann. *Klassifikation von Mustern, 2. überarbeitete Auflage im Internet*. Heinrich Niemann, http://www5.informatik.uni-erlangen.de/niemann/homeg.tht/homegli1.html, 2003.

[NMN03] Seiichi Nakagawa, Kazumasa Mori, and Naoki Nakamura. A statistical method of evaluating pronunciation proficiency for english words spoken by japanese. In *Proceedings of the European Conference On Speech Communication and Technology (Eurospeech)*, pages 3193–3196, 2003.

[PFS] *Preparing for future mulit-sensorial interaction research (PF_STAR) project*. http://pfstar.itc.it/.

[Pic93] Joseph W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993.

[RFP+94] Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals. WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition. 1994.

[RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[RSS92] Nimal Ratnayake, Michael Savic, and Jeffrey Sorensen. Use of semi-markov models for speaker-independent phoneme recognition. In *Proceedings of the ICASSP*, volume 1, pages 565–568. IEEE Computer Society Press, 1992.

[SAM] *SAMPA: a computer readable phonetic alphabet.* http://www.phon.ucl.ac.uk/home/sampa/home.htm.

[Sch95] Ernst G. Schukat–Talamazzini. *Automatische Spracherkennung.* Vieweg, Wiesbaden, 1995.

[SSN+02] Georg Stemmer, Stefan Steidl, Elmar Nöth, Heinrich Niemann, and Anton Batliner. Comparison and combination of confidence measures. *Proceedings of the Fifth International Conference on Text, Speech, Dialogue (TSD), Lecture Notes in Artificial Intelligence*, 2448:181–188, 2002.

[Ste01] Stefan Steidl. Konfidenzbewertung von Worthypothesen. Technical report, Studienarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, 2001.

[TFS+00] Carlos Teixeira, Horacio Franco, Elizabeth Shriberg, Kristin Precoda, and Kemal Sönmez. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[TFS+01] Carlos Teixeira, Horacio Franco, Elizabeth Shriberg, Kristin Precoda, and Kemal Sönmez. Evaluation of speaker's degree of nativeness using text-independent prosodic features. In *Workshop on Multilingual Speech and Language Processing*, 2001.

[Veh00] Kimmo Vehkalahti. *Reliability of measurement scales: Tarkkonen's general method supersedes Cronbach's alpha.* PhD thesis, University of Helsinki, Department of statistics, 2000.

[Wel] John C. Wells. *Computer-coding the IPA: a proposed extension of SAMPA.* http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm.

[Wik] *Wikipedia, the free encyclopedia.* http://en.wikipedia.org/.

[WSMN01] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298, 2001.

[WY97] Silke M. Witt and Steve J. Young. Language learning based on non-native speech recognition. In *Proceedings of the European Conference On Speech Communication and Technology (Eurospeech)*, pages 633–636, 1997.

[WY00] Silke M. Witt and Steve J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108, 2000.