

**Using the HTK Hidden Markov Model
Toolkit Speech Recogniser to Analyse
Prosody in a Corpus of German Spoken
Learners' English**

Toshifumi Oba

*Submitted in accordance with the requirements for
the degree of Master of Science*

**The University of Leeds
School of Computing**

January, 2003

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgments

First and foremost, I would like to thank my supervisor, Eric Atwell, for his generous and enthusiastic guidance. His insightful and constructive advice helped me focus and gave me an enormous confidence. Without his unfailing supports, this research would have not been achieved. I am deeply grateful to him.

I received a tremendous benefit from the Interactive Spoken Language Education (ISLE) project for reuse of a speech corpus. I am grateful to everyone involved in the project: Leeds University, Universität Hamburg, Università di Milano-Bicocca, Entropic Ltd., Ernst Klett Verlag GmbH, and Dida*El S.R.L.

‘Summer University - Speech Recognition 2002’ of Otto-von-Guericke-Universität Magdeburg helped my understanding of the Hidden Markov Model Toolkit (HTK) speech recogniser. I am deeply grateful especially to Professor A. Wendemuth, Doctor S. E. Krüger and a PhD student M. Katz for their kind support and permission to use script files for the HTK manipulation.

I also would like to thank North-West Centre for Linguistics (NWCL). Lectures of their research training course provided me a precious opportunity to learn basic knowledge of linguistics.

I am also grateful to Peter Howarth, an English language teaching researcher at the Leeds University. He not only voluntarily judged German spoken learners’ English intonation abilities, but also gave me constant instruction in linguistics aspects of English language.

Finally, I would like to thank my family in Japan for financial support and giving me an opportunity to study at Leeds University.

Abstract

The Interactive Spoken Language Education (ISLE) project collected a corpus of audio recordings of German and Italian spoken learners' English, in which subjects read aloud samples of English text and dialogue selected from typical second language learning exercises (Atwell et al, 2000b; 2003; Menzel et al, 2000). The audio files were time-aligned to graphemic and phonetic transcriptions, and speaker errors were annotated at the word and the phone-level, to highlight pronunciation errors such as phone realisation problems and misplaced word stress assignments.

This thesis describes reuse of the ISLE corpus in experiments using speech recognition to investigate prosody in the German speakers. There were three main stages: prosodic annotation of the English text in the corpus, following a model devised for speech synthesis; native speakers' assessments of the intonation abilities of the 23 German speakers; and speech recognition experiments using the Hidden Markov Model Toolkit (HTK) (Young et al, 2001).

Prosodic annotation was done following the set of instructions or informal algorithm in (Knowles, 1996), to predict 'model' intonation patterns for written English text, to be passed to a speech synthesiser. We hoped to use these predicted intonations as a 'native speaker target', against which to compare learners' actual intonation patterns, so we investigated automated methods to extract intonation features from the learners' speech-files. Unfortunately, we were unable to automatically predict markup equivalent to the synthesiser cues, so could not directly compare the learners' against this model.

Instead, we turned to expert human evaluation: a computational linguistics researcher and an English language-teaching (ELT) researcher subjectively assessed the intonation of the recorded utterances from German learners of English, by listening to the recorded utterances, comparing against the 'model' marked-up script, and counting perceived intonation errors. Using these judgments, the learners were divided into two groups: 'poor' intonation group and 'good' intonation group. Speakers with exceptionally poor pronunciation were excluded in this grouping by referring to data in the corpus so that results of the following recognition experiments are independent from pronunciation ability.

Finally, the Hidden Markov Model Toolkit (HTK) was used to train monophone and triphone Hidden Markov Models for 'poor' intonation group and 'good' intonation group separately. In each case, the training set excluded test speakers; this was achieved via cross-validation, repeating the experiment, taking out a different test-subset each time, and averaging the results. Every trained model was tested with test speakers from both 'poor' and 'good' intonation groups. Results reveal that recognition accuracy becomes higher when models are trained with a group of same intonation ability as test speakers. Cross-merging experiments confirm that these results are consistent. Additional experiments secure the dominance of prosodic factors and irrelevancy of pronunciation abilities to the results.

Table of Contents

Abbreviations	v
Conventions	iv
List of Figures	v
List of Tables	vi
Chapter 1: Introduction	1
1.1 Introduction	1
1.2 Corpora and Prosodic Annotation	4
1.2.1 Corpora	4
1.2.2 Corpora and Prosodic Annotation	5
1.2.2.1 Spoken English Corpus (SEC)	5
1.2.2.2 London-Lund Corpus (LLC)	5
1.2.2.3 Hong Kong Corpus of Spoken English (HKCSE)	6
1.2.2.4 EUSTACE	6
1.2.3 Selection of Annotation Instruction and Tool	7
1.3 The UKTK	8
1.4 Examples of Speech Recognition Research	9
1.5 Outline of the Thesis	10
Chapter 2: Prosodic Annotation	11
2.1 Instructions for Prosodic Annotation	11
2.1.1 Step 1: Prosodic Parsing	12
2.1.2 Step 2: Assembling Blocks	14
2.1.3 Step 3: Lightning Rules	15
2.1.4 Step 4: Tone Types	16
2.2 Result of Prosodic Annotation	17
2.3 Analysis of Annotation Result	18
2.4 Modification of Prosodic Annotation	21

Chapter 3:	Human Evaluation of Intonation Abilities and Grouping of German Speakers	23
3.1	Amount of Human Evaluation	24
3.2	Process of Human Evaluation	25
3.3	Results from Human Evaluation	26
3.4	Analysis of Evaluation Results	28
3.5	Grouping German Speakers	29
3.5.1	Exclusion of Pronunciation Factors	30
3.5.2	Exclusion of Intermediate Speakers	30
3.5.3	Rules to Group Speakers	31
3.5.4	Grouping I	32
3.5.5	Grouping II	32
3.5.6	Grouping III	33
Chapter 4:	The HTK Speech Recognition Experiments for Analysing Prosody	34
4.1	Script and Configuration Files for the HTK Experiments	35
4.1.1	Configuration File for Training	36
4.1.2	Configuration File for Recognition Test	37
4.1.3	Files Required	38
4.2	Training Model and Parameter Investigation	38
4.2.1	Experimental Conditions	39
4.2.2	Grammar Scale Factor (GSF)	39
4.2.3	Training Models	40
4.2.4	Word Insertion Penalty (WIP)	41
4.3	Reduction of Training Speakers	42
4.3.1	Experimental Conditions	43
4.3.2	Reduction of Training Speakers	43
4.3.3	Adjustment of WIP	44
4.4	The HTK Experiments for Prosodic Analysis	45
4.4.1	Experimental Conditions	46
4.4.2	Experiment I – Grouping I	46
4.4.3	Experiment II – Grouping II	48

4.4.4	Experiment III – Grouping III	49
4.5.	Dominance of Prosodic Factors and Irrelevance of Pronunciation Abilities	50
4.5.1	Recognition of Prosodic Keywords	51
4.5.2	Experiment to Prove Irrelevance of Pronunciation Abilities	53
4.6	Analysis of Experiments Results	54
Chapter 5:	Conclusions	58
5.1	Summary	58
5.1.1	Prosodic Annotation	58
5.1.2	Human Evaluation and Grouping	59
5.1.3	The HTK Experiments	60
5.2	Contributions	62
5.2.1	Contribution to Speech Recognition Research	62
5.2.2	Contribution to HTK Research	62
5.2.3	Contribution to Foreign Language Learning	63
5.3	Future Work	63
	References	65
	Appendix 2 Final Version of Prosodic Annotation	71
	Appendix 3.1 Score Sheets of Human Evaluation	77
	Appendix 3.2 Pronunciation Abilities of 23 German Spoken Learners’ English	79
	Appendix 3.3 Listing of German Speakers in ‘Good’ Englishg Intonation Order	81
	Appendix 4 Files Required for the HTK Speech Recognition Experiments	85

List of Tables

Table 1.1	Features of Various Speech Recognition Research	10
Table 2.1	Grammatical Tags	12
Table 2.2	Inflectional Types: Noun	12
Table 2.3	Inflectional Types: Verb	12
Table 2.4	Accentual Types	13
Table 2.5	Transition Marks	13
Table 2.6	Tone Types	17
Table 3.1	Number of Sentences with Intonation Error(s)	27
Table 4.1	Affect of GSF to Recognition Accuracy	40
Table 4.2	Affect of WIP to Recognition Accuracy against Triphone MMs	42
Table 4.3	Affect of WIP to Recognition Accuracy against Monophone MMs	42
Table 4.4	Reduction of Training Data and Recognition Accuracy: IP = -20.0	43
Table 4.5	Reduction of Training Data and Recognition Accuracy: IP = -40.0	43
Table 4.6	Reduction of Training Data and Word Error Types: IP = -20.0	44
Table 4.7	Reduction of Training Data and Word Error Types: IP = -40.0	44
Table 4.8	Balance of Word Error Types: WIP = -60.0	45
Table 4.9	Result: ‘Good’ Intonation Speakers (Triphone Models) <1>	47
Table 4.10	Result: ‘Poor’ Intonation Speakers (Triphone Models) <1>	47
Table 4.11	Result: ‘Good’ Intonation Speakers (Monophone Models) <1>	47
Table 4.12	Result: ‘Poor’ Intonation Speakers (Monophone Models) <1>	47
Table 4.13	Result: ‘Good’ Intonation Speakers (Triphone Models) <2>	48
Table 4.14	Result: ‘Poor’ Intonation Speakers (Triphone Models) <2>	48
Table 4.15	Result: ‘Good’ Intonation Speakers (Monophone Models) <2>	48
Table 4.16	Result: ‘Poor’ Intonation Speakers (Monophone Models) <2>	48
Table 4.17	Result: ‘Good’ Intonation Speakers (Triphone Models) <3>	49

Table 4.18	Result: ‘Poor’ Intonation Speakers (Triphone Models) <3>	49
Table 4.19	Result: ‘Good’ Intonation Speakers (Monophone Models) <3>	50
Table 4.20	Result: ‘Poor’ Intonation Speakers (Monophone Models) <3>	50
Table 4.21	Result of Recognised Keywords: ‘Good’ Intonation Test Speakers	52
Table 4.22	Result of Recognised Keywords: ‘Poor’ Intonation Test Speakers	52
Table 4.23	Result: ‘Good’ Intonation Speakers (Triphone Models) <4>	54
Table 4.24	Result: ‘Poor’ Intonation Speakers (Triphone Models) <4>	54
Table 4.25	Result: ‘Good’ Intonation Speakers (Monophone Models) <4>	54
Table 4.26	Result: ‘Poor’ Intonation Speakers (Monophone Models) <4>	54
Table A3.1	Score Sheet of Evaluator I	77
Table A3.2	Score Sheet of Evaluator II	78
Table A3.3	Number of Pronunciation Errors	80

List of Figures

Figure 2.1	Example of Annotation by Step 1	13
Figure 2.2	Example of Annotation by Step	15
Figure 2.3	Final Version of Tagged Example Sentence	17
Figure 4.1	File of Recognised Words	52
Figure A4.1	Scenario File: 'all_sent.snr'	85
Figure A4.2	Word List: 'wordlist'	87
Figure A4.3	Monophone Pronunciation Dictionary: 'mono.dic'	87
Figure A4.4	Monophone List: 'monophon.lst'	87
Figure A4.5	HMM Definition: 'hmm.def'	87
Figure A4.6	Lists of all wave files and feature files: 'w2f.all'	88
Figure A4.7	List of All Feature Files: 'feat.all'	89
Figure A4.8	List of Feature Training Files: 'feat.train'	89
Figure A4.9	List of Feature Test Files: 'feat.test'	89
Figure A4.10	List of All Lables: 'text.all'	89
Figure A4.11	0-gram Lattice: 'zero.lat'	90
Figure A4.12	2-gram Lattice: 'bigr.lat'	90

Abbreviations

GSF	Grammar Scale Factor
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
ISLE	Interactive Spoken Language Education
WIP	Word Insertion Penalty

Conventions

Two words, ‘prosody’ and ‘intonation’, were equally used in this thesis. However, ‘intonation’ was preferably used in compound nouns such as ‘intonation ability’ and ‘intonation error’.

Each appendix was numbered corresponding to a chapter which the appendix implemented, e.g. Appendix 2 is a supplement of Chapter 2.

Throughout the whole thesis, ‘we’ and ‘our’ refer to ‘author’ and ‘author’s’.

Chapter 1

Introduction

1.1 Introduction

Intonation is important in human communication to help the listener to understand the meaning and attitude of the speaker (Brown, 1977; O'Connor, 1970). Language students and teachers see intonation as a part of the structure of the language (Tench, 1996). However, acquisition of proper intonation is difficult for non-native speakers and requires repeated practice (O'Connor, 1970). For example, the Interactive Spoken Language Education (ISLE) project found that intonation was the biggest problem for German learners of English, but the project did not tackle intonation (Atwell et al, 1999).

Intonation is also important for speech recognition to retrieve correct semantic and syntactic information and to successfully identify words (Rodman, 1999; Werner and Keller, 1994). It not only influences acoustic models, but can also contribute for language models; Taylor et al (1997; 1998) focused on the relationship of intonation and 'move types' of the dialogue to constrain the number of possible language models. However, intonation is generally limited to be that of native-speakers in speech recognition research.

Speech recognition research has been done for non-native speakers too. The works can be categorized into two groups: multilingual speech recognition; and use of speech recogniser for foreign language learning. Uebler (1998) investigated bilingual (Italian and German) and multi-dialectal speech recognition by assuming the two languages being one. Stemmer et al (2001) developed acoustic models of foreign words which appeared in German dialog, such as English words in film titles. In the FLUENCY project, Eskenazi (1996) used a speech recogniser to detect foreign speakers' pronunciation errors for second language training. This work also involved prosody looking at a correlation of pronunciation and prosody errors. However, little

research has been undertaken on speech recognition targeted on non-native speakers' intonation.

The HTK (Hidden Markov Model Toolkit) is a free and portable toolkit for building and manipulating Hidden Markov Models (HMMs) primarily for speech recognition research, although it has been widely used for other topics such as speech synthesis, character recognition and DNA sequencing (HTK3, 2000; Young et al, 2001). Taylor et al (1998) exploited the HTK-based speech recogniser to provide word hypotheses constrained by 'move types' correlated with intonation. However, direct analysis of prosody using the HTK has not been the focus of any past research.

The ISLE project exploited the IHAPI HMM-based speech recogniser to improve the performance of computer-based English learning systems, such as providing clear feedback by specifying error words and phones (Atwell et al, 1999; 2003). The project collected a corpus of audio recordings of 23 Italian, 23 German Spoken Learners' and 2 native speakers' English, in which subjects read aloud samples of English text and dialogue selected from typical second language learning exercises, such as pronunciation and stress placement training using minimal pairs and polysyllabic words. (Atwell et al, 2000b; 2003; Menzel et al, 2000). The audio files were time-aligned to graphemic and phonetic transcriptions, and speaker errors were annotated at the word- and the phone-level, to highlight pronunciation errors such as phone realisation problems and misplaced word stress assignments. We re-used the first three blocks of the corpus, Block A through C, which contained 82 sentences edited from 'The Ascent of Everest' (Hunt, 1996). As the rest of the corpus generally consisted of shorter sentences or just words without a uniform topic, the first three blocks were the best for prosodic analysis using a speech recogniser.

Our research analysed German spoken learners' English prosody re-using the ISLE speech corpus by using the HTK-based speech recogniser. There were three main stages to the research: prosodic annotation of the English text in the corpus, following a model devised for speech synthesis; native speakers' assessments of the intonation abilities of the 23 German speakers; and speech recognition experiments using the HTK.

Prosodic annotation was done following the set of instructions or informal algorithm in (Knowles, 1996), to predict 'model' intonation patterns for written English text, to be passed to a speech synthesiser. The annotation was done to all 27 sentences of Block A of the ISLE corpus. All the tags were added by hand using Windows Excel. The process consisted of four steps: prosodic parsing to divide the text into blocks; assembling blocks to form a potential tone group; adapting lightening rules to merge the groups into an actual tone group; deciding the type of the tone group. The annotation was generally achieved by simple mappings of each step such as from grammatical tags and transition markers to assembling of tone units. The marked-up prosody was compared with the native speakers' recordings and some of the patterns were modified. We hoped to use the modified prosodic patterns as a 'native speaker target', against which to compare learners' actual prosodic patterns, so we investigated automated methods to extract prosodic features from the learners' speech-files. Unfortunately, we were unable to automatically predict markup equivalent to the synthesiser cues, so could not directly compare the learners against this model.

Instead, we turned to expert human evaluation: a computational linguistics researcher and an English language-teaching (ELT) researcher subjectively assessed the intonation of the recorded utterances from German learners of English, by listening to the recorded utterances, comparing against the 'model' marked-up script, and counting perceived intonation errors. The judgments were used to partition the speakers: the speakers were divided into two groups by assigning the upper half of them, who made fewer intonation errors, to a 'good' intonation group; and the rest to a 'poor' intonation group. Three different groupings were done: two groupings based on each of the two evaluators; and the third based on agreement of the two evaluators. Speakers with exceptionally poor pronunciation as indicated by the ISLE corpus pronunciation markup were excluded in this grouping so that results of the following experiments would be independent from pronunciation ability.

Finally, the speech recognition experiments were undertaken using the HTK. Before starting the main experiments for prosodic analysis, we made two preparation experiments: for investigating best training models and several parameters for recognition tests; and for checking the influence of reducing training speakers to the

recognition accuracy. In the main experiments, the HTK was used to train monophone and triphone Hidden Markov Models for a 'poor' intonation group and a 'good' intonation group separately. In each case, the training set excluded test speakers and each model was tested with the test speakers from both 'good' and 'poor' intonation group. This was achieved via cross-validation, repeating the experiment, taking out a different test-subset each time, and averaging the results. The whole process was repeated three times taking a different grouping. Results reveal that recognition accuracy becomes higher when models are trained with a group of same intonation ability as test speakers. Cross-merging experiments confirm that these results are generally consistent.

We believe the analysis contributes to speech recognition research and foreign language learning technology. Our HTK experiments found better training models for German speakers with 'good' and 'poor' intonation speakers separately. Its results should help to find an effective way to train English speech recognition systems for German speakers with various English intonation abilities.

The use of Speech recognition has been investigated in Computer-Assisted Language Learning (CALL) such as in Eskenazi (1996) and Witt and Young (1997). However, the ISLE project reported that detection of errors and providing feedback to the learners were not well developed in existing language learning software (Atwell et al, 1999). For example, a well-known software, 'Tell Me More' of Auralog, improved the detection and the feedback for pronunciation practice by pointing out error phonemes and showing a 3D animation to visualize the 'model' articulation. However, its technology for intonation practice is still poor. Eskenazi (1999) mentions that a visual display is more effective than oral instructions for intonation practice and 'Tell Me More' displays a waveform and pitch curve, which traces the amplitude and frequency variations, respectively, of both user's voice and a 'model' utterance. However, it neither points out the placement of the intonation errors, nor provides suggestions to improve intonation, leaving the comparison tasks to the users. Investigation of Germans speakers' English prosody should help to improve these technologies, by pointing out their common weakness in English intonation, which should be included in exercises.

1.2 Corpora and Prosodic Annotation

1.2.1 Corpora

Corpora are any collections of text and/or speech, and are used as a basis of statistical processing of natural language (Jurafsky and Martin, 2000). There are various kinds of corpora: tagged or untagged; monolingual or multilingual; balanced or specialized. For example, one of the largest and best-known corpora, the British National Corpus (Warwick, 1997), consists of 100 million words of written (about 90%) and speech (about 10%) data collected from modern British English which covers a variety of styles and subjects.

There are arguments about pros and cons of using corpora in research. For example, Chomsky (1957; 1965) argues that empirical evidence does not give deep insight into language structure, and that statistical models are of no interest on linguistics. However, Leech (1992) argues that corpus is useful for quantitative and empirical research.

1.2.2 Corpora and Prosodic Annotation

Annotation is to add linguistic information to text and/or speech data in order to enrich and extract information from the corpus (McEnery and Wilson, 1999). There is a variety of annotations such as grammatical, prosodic and discourse annotations. Annotation can be also done at different levels such as sentence, word- and syllabic-levels. Prosodically annotated corpora are useful for phonetics and speech processing researchers. We summarize several prosodically annotated corpora and one corpus designed for analysis of utterance duration.

1.2.2.1 *Spoken English Corpus (SEC)*

The Spoken English Corpus (SEC) was originally collected for a database of intonation assignment programs, and consists of 52,000 words of contemporary spoken British English BBC radio broadcasts (Taylor and Knowles, 1998). The corpus is designed in various forms: unpunctuated transcriptions; orthographic transcriptions; grammatically tagged versions; prosodic transcriptions.

The SEC corpus clearly distinguishes prosodic features of stressed syllable and accented syllables; the latter are stressed and have independent pitch movement, but the former do not have pitch movement. The corpus also deals with 14 pitch directions: simple tones such as ‘fall’ and ‘rise’; complex tones such as ‘fall-rise’ and ‘rise-fall’; and distinctions of high and low to these tones; high and low tones start higher and lower than the preceding pitches, respectively.

1.2.2.2 *London-Lund Corpus (LLC)*

The London-Lund Corpus (LLC) consists of 100 spoken and 100 written texts, computerized and amended from the Survey of English Usage (SEU) corpus which aimed at providing accurate descriptions of grammar of adult educated speakers of English since 1959 (Svartvik, 1990). Among the 100 texts, 17 were recorded from spoken delivery of written material, such as news broadcasts, plays and scripted speeches, and these are prosodically annotated.

According to Svartvik (1990), prosodic annotation of the LLC corpus has a lot more features than the SEC corpus.

- Tone unit boundaries
- Location of onset and nucleus
- Direction of pitch
- Length of pause
- Degree of stress, loudness and tempo

- Modification in voice quality (e.g. pitch and tension)
- Paralinguistic features (e.g. whisper and creak)

1.2.2.3 *Hong Kong Corpus of Spoken English (HKCSE)*

The Hong Kong Corpus of Spoken English (HKCSE) is a collection of 50 hours of naturally-occurring English conversations between Hong Kong Chinese and native English speakers, and consists of four sub-corpora: conversations, academic discourses, business discourses and public discourses (Cheng and Warren, 1999).

Prosodic annotation was added to the business sub-corpus to analyse the use of intonation by the two sets of speakers. Their recent research focused on the occurrence of two tones (rise and rise-fall) associated with the assertion of dominance, looking at a variety of discourse types in business. It was found that the choice of the two tones was at least partly determined by both text types and designated roles of speakers.

1.2.2.4 *EUSTACE*

The Edinburgh University Speech Timing Archive and Corpus of English (EUSTACE) is designed for examining durational effects in speech, which are controlled for length and phonetic content (White and King, 2003). It was recorded from 6 British English speakers with a total of 4,608 sentences.

This corpus contains 16 test syllables with each controlled by various terms:

- Number of syllables in a word; 1 (test syllable only), 2 or 3 syllables in the word.

- Placement of the test syllable in a word; the first, second or third syllables from the beginning or from the end of the word.
- Length of the utterance; where the test syllable is close to the edge of utterance, utterance length covaries with word length and varies whilst word length is fixed.

The corpus is annotated with the following features.

- Utterance production labels; e.g. identify incorrect placement of emphasis.
- Test syllable labels; the beginning and the end of onset, nucleus and coda of the test syllable.
- Additional syllable labels; the beginning and the end of additional syllables in the word which contains a test syllable.

White (2002) used this corpus to investigate durational processes associated with suprasyllabic speech structure. He found that accentual lengthening did not interacted the word-initial lengthening, and that locus of accentual lengthening was the word and especially significant at word edges. This corpus can be used to investigate deeper or other relationships of duration and prosody, such as duration and accentual types.

1.2.3 Selection of Annotation Instruction and Tool

The prosodic features of the SEC corpus are the basis of prosodic rules for the text-to-speech system of Knowles (1996), which was used for prosodic annotation in our research. In the annotation instructions, the number of the pitch types is reduced to 5: ‘fall’, ‘rise’, ‘fall-rise’, ‘high rise’, ‘low-rise’ and ‘level’. Prosodic features in the SEC corpus and the instructions for prosodic annotation are a lot simpler compared with

the LLC corpus. However, as our purpose was to define a ‘model’ English prosodic pattern of native speakers, to compare against that of German speakers, in order to judge their intonation abilities, it was not worth considering as many prosodic features as in the LLC corpus.

ToBI is a well-known algorithm for prosodic annotation, which has been developed not only for English, but also for other languages such as German and Japanese. ToBI is especially convenient to retrieve instances of the same type of the event from a large corpus, as it captures categories of prosodic phenomena (Silverman et al, 1992). However, with the same reason as above, we did not require such a sophisticated algorithm.

MATE is a tagger toolkit which allows multiple levels of annotation (MATE, 1998). However, as we needed to annotate only 27 sentences (See Section 2.1), it was not worth investigating such an automatic annotation tool. Instead, we simply used Windows Excel which satisfied our needs as an annotation environment.

1.3 The HTK

We exploited the Hidden Markov Model Toolkit (HTK) (Young et al, 2001) for analysing German speakers’ English intonations. The HTK is a toolkit for building Hidden Markov Models (HMMs) which is mainly used for speech recognition research. It has been developed at Cambridge University Engineering Department (CUED) since 1989, although it was once purchased by Entropic Research Laboratory (ERL) in 1995¹ and by Microsoft in 1999² (Evermann, 2002; Woodland, 2000). The toolkit is available for both UNIX and Windows machines.

The HTK is a set of modules which can be called from both command line and script file(s). For speech recognition experiments, it requires speech files, prototype HMMs, scenario file of the speech data, and its lexicon at various levels such as word, monophone and triphone transcriptions.

¹ The HTK had been distributed with ERL since 1993.

The Master Label File (MLF) is the most characteristic feature of the HMMs. Each label file, which the HTK compares against its utterance, should have a transcription of one utterance with the location where the speech data is stored. All the label files can be stored in a single MLF. This is more efficient as the MLF allows multiple labels to share one label, if the labels contain the same transcription, but are stored in different locations (e.g. same utterance from several speakers), by substituting their locations with ‘*’.

In the actual processing, the HTK firstly parameterizes features of speech data to various forms such as Linier Predictive Coding (LPC) and Mel-Cepstrum. Then, it will estimate the HMM parameters using the Baum-Welch Algorithm for training. Recognition tests are executed by estimating the best hypothesis from given feature vectors and from a language model. Results are given with recognition percentage as well as numbers of deletion, substitution and insertion errors.

The HTK was an optimal toolkit for our speech recognition research. It is a well-known and free toolkit which has been widely used not only for speech recognition research, but also for other pattern recognition such as hand-writing recognition and facial recognition. Although many speech recognition researches have already exploited the HTK, it was worth investigating the HTK for prosodic analysis, especially of non-native speakers’ utterances, as the other research dealt with pronunciation factors and with native speakers.

1.4 Examples of Speech Recognition Research

As mentioned in Section 1.1, the majority of speech recognition research focuses on pronunciation of native speakers. While there have been relatively few studies on speech recognition dealing with prosody or non-native speakers, those for non-native speakers’ prosody have received even less attention.

² Microsoft later returned the license to CUED.

Table 1.1 shows a variety of speech recognition research and which fields are involved in this research. Research fields to be considered are pronunciation, prosody, native speaker, non-native speaker, monolingual, multilingual and the HTK. ‘Y’ represents a field which a research deals with: otherwise ‘N’ is given. More precisely, ‘Y’ is marked in each column as follows:

- Non/Mul: when speech recogniser are for non-native speakers’ utterances or for multi-languages;
 - German: when speech recogniser deals with German speakers’ English;
 - Inton: when speech recogniser considers intonation features;
 - G/B: when the research deals with ‘goodness’ and ‘badness’ of intonation;
 - HTK: when the research exploits the HTK based-speech recogniser.
- These terms do not have to be the main focus of the research, e.g. if the research at least considers the feature, then ‘Y’ is given.

Table 1.1: Features of Various Speech Recognition Research

Research Reference	Non-N	German	Inton	G/P	HTK
(Taylor, 1998)	N	N	Y	N	Y
(Uebler, 1998)	Y	N	N	N	N
(Stemmer, 2001)	Y	Y	N	N	N
(Teixeira, 1996)	Y	Y	N	N	Y
(Hansen, 1995)	Y	Y	Y	Y	N
(Yan and Vaseghi, 2002)	N	N	Y	N	Y
(Jurafsky et al, 1994)	Y	Y	N	N	N
(Berkling et al, 1998)	Y	N	N	N	Y
(Oba and Atwell, 2003)	Y	Y	Y	Y	Y

1.5 Outline of the Thesis

Chapter 2 explains prosodic annotation of 27 written English sentences of the ISLE corpus. The Annotation instructions of Knowles (1996) are summarized showing an actual annotation of one of the sentence as an example. After showing results of all

the sentences with tone types, we modify the tone types by comparing against two native speakers' recordings.

Chapter 3 shows results of human evaluation of 23 German speakers' English intonation abilities comparing the assessment results from two evaluators. Three different groupings of the 23 speakers by their intonation abilities are given based on each of the two judgements and agreement of the two evaluators.

Chapter 4 explains the HTK experiments for prosodic analysis. Before the main experiments, we show two preparation experiments for finding the most suitable training models and parameters for recognition tests. Results from the main experiments are followed by prosodic analysis.

Chapter 5 summarizes overall results from our research. Then, we discuss the results and contribution of the work, and conclude with possible future work.

Chapter 2

Prosodic Annotation

Prosodic annotation was done following the set of instructions or informal algorithm in (Knowles 1996), to predict ‘model’ intonation patterns for written English text. We hoped to use these predicted intonations as a ‘native speaker target’, against which to compare learners’ actual intonation patterns.

The instructions generally consisted of simple mappings such as from grammatical tags and transition markers to block assembling. However, there were many cases, in which decisions on block merging to form the final tone groups, were not clear. We left them as separate blocks in many such cases, and this tended to predict more tone groups than in intonation patterns produced by native speakers in their recordings of utterances. These redundant intonation patterns were deleted, so that the patterns should be ignored in later evaluations of German speakers’ English intonation abilities.

All the tags were added ‘by hand’ using Windows Excel. Although an automatic PoS³-tagger such as AMALGAM-tagger (Atwell et al, 2000a) could have been used for the grammatical tagging, it was not worth investigating an additional work of file-format conversion for our small sample of 27 sentences.

Section 2.1 explains the instructions for prosodic annotation showing an example sentence. Section 2.2 shows a result of the prosodic annotation, while Section 2.3 analyses the result. Modification of the prosodic annotation is done in Section 2.4.

2.1 Instructions for Prosodic Annotation

³ PoS: part-of-speech

This section summarizes instructions for annotation explained in (Knowles, 1996), which we followed to add a prosodic annotation to 27 sentences consisted of 429 words in Block A of the ISLE corpus. The outline is explained with actual annotation to one of the 27 sentences, ‘It is in fact a story of many years, in which many men tried to climb that mountain’ (Souter et al, 1999). The process consisted of four steps: prosodic parsing to divide the text into blocks; assembling blocks to form a potential tone group; adapting lightening rules to merge the groups into an actual tone group; deciding the type of the tone group. The four steps are explained in Sub-Section 2.1.1 through 2.1.4.

2.1.1 Step 1: Prosodic Parsing

The first step was to make a prosodic parsing of the text by putting a ‘p-tag’ to each word and a transition marker between successive words. The ‘p-tag’ showed grammatical and accentual types of the words, which were considered as prosodic properties, while the transition marker indicated the closeness of successive words and was referred to in the lightning rules of Step 3.

Table 2.1 is a list of these grammatical tags. Table 2.2 and Table 2.3 show tags of inflectional types additionally put with nouns and verbs, respectively, whose types possibly influence prosodic behaviours of the word (Knowles, 1996).

Table 2.1 Grammatical Tags

Tag	Type	Example	Tag	Type	Example
A	Adverb	<i>Always</i>	NA	Adverbial noun	<i>kindliness</i>
C	Conjunction	<i>And</i>	NC	Common noun	<i>mountain</i>
E	Exsistential 'there'	<i>There</i>	NF	Foreign word / Formula	<i>maron</i>
I	Proposition	<i>In</i>	NI	Pronoun	<i>they</i>
J	Adjective	<i>High</i>	NT	Title	<i>professor</i>
K	Numeral	<i>Two</i>	NP	Proper noun	<i>London</i>
Q	Qualifier	<i>Very</i>	NZ	Character	<i>Superman</i>
U	Interjection	<i>Oh</i>	VA	Auxiliary verb	<i>can</i>
W	Wh-word	<i>When</i>	VI	Main noun	<i>climb</i>
X	Not	<i>Not</i>			

Table 2.2 Inflectional Types: Noun

Verb

Tag	Type	Example
S	Subject	<i>I</i>
O	Object / Oblique	<i>me</i>
\$	Possessive	<i>my</i>

Table 2.3 Inflectional Types:

Tag	Type	Example
F	Finite	<i>took</i>
I	Infinitive	<i>take</i>
P	Participle	<i>taking</i>

According to (Knowles, 1996), each prosodic block is typically constructed with the size of noun phrases, prepositional phrases, adverbial phrases and verb phrases; however the type of the block is determined by the grammatical tags of its final word, e.g. a block ending with a noun word is called a ‘noun block’.

The grammatical tags were useful to distinguish ‘lexical’ or ‘grammatical’ words, which decided accentual types. Knowles (1996) mentions that lexical words such as adverbs, adjectives and nouns except pronouns and titles keep their accents, while grammatical words, such as verbs, determiners and prepositions, tend to lose the accents. The instructions defined four accentual types depending on the lexical or grammatical words as seen in Table 2.4.

Table 2.4 Accentual Types

Tag	Type	Criteria	Example
A	Accented	Lexical words	<i>mountain</i>
S	Stressed	Grammatical words without weak-forms	<i>professor</i>
U	Unreduced quality	Grammatical words with weak-forms, but do not have such a vowel as in 'W'	<i>may, on, your, when</i>
W	Weak	Grammatical words with weak-forms whose vowel is reduced to shwa or /l/	<i>has, the On, and</i>

The grammatical tags were also useful to decide the types of transition markers, although it required more careful observation of the context. Six different level transition markers were given in the annotation rules put to the words as shown in Table 2.5, in which tags are shown in the order of closeness of successive words. Figure 2.1 shows the example sentence with the annotation described in this step.

Table 2.5 Transition Marks

Tag	Type	Criteria
_	Word	Other cases within the block
	Colligation	Ending of 'noun-noun', 'adjective-noun', 'qualifier-qualified' ⁴ within the block
,	Comma	Block ending with verb or possessive

⁴ For example, the comma was given after ‘very hot’.

;	Semi-colon	Block ending with adverb, 'not', or postmodification ⁵
:	Colon	Block ending with right bracket, comma or ending of noun block
.	Period	Block ending with punctuations except right bracket and comma

It		Is		In		fact		a		story	
NIS	:	VMF	,	I		NCO	:	D	_	NCO	:
W		W		W		A		W		A	

Of		Many		years,		In		which		many	
I		J	_	NCO	:	I		W	:	J	_
W		A		A		W		U		A	

Men		Tried		To		climb		that		mountain.	
NCS	:	VMF		I		VMI	,	J	_	NCO	.
A		S		W		S		A		A	

Figure 2.1 Example of Annotation by Step 1

- ❖ Each of the wide cells in the first columns contains one word of the example sentence, while short cells in the columns show boundaries of blocks with '|', which we added.
- ❖ Grammatical tags including inflectional types are shown in wide cells of the second columns, while transition markers are given to short cells in the columns.
- ❖ The third columns show accentual types of the words.

2.1.2 Step 2: Assembling Blocks

The second step of the annotation rules was to assemble blocks by reducing the number of accents in each block based on three deaccentuation rules explained in his earlier work (Knowles, 1987). Nucleus and onset are the last and the first, if there is more than one, accented syllables in a block, and the former is a place where intonation takes a fall, while intonation stays at high level on the latter (Knowles, 1996).

1. grammatical words with weak forms,
2. the second elements of compound nouns, and

⁵ For example, in the case of 'University of Leeds', 'University' was given the semi-colon.

3. any potential accents between onset and nucleus.
(Knowles 1996: 159)

Based on these rules, we renewed accentual types defined in Step 1 by putting ‘A’ for those words whose accents were retained and ‘D’ for those which lost accents. ‘W’ and ‘U’ were automatically down to ‘D’ due to the first deaccentuation rules. In the example sentence, there were no blocks which contained more than two potential accents, that is, those given ‘A’ or ‘S’. As a result, all of these words received a new tag, ‘A’. If the accented word contained multiple syllables, then ‘A’ was given to only an accented syllable, and ‘D’ was given to the rest of syllables in the word as seen in ‘mountain’ of the example sentence. Syllabification was done by referring to an EFL dictionary (Matsuda, 1999). The tags for accentual types were added to the example sentence as seen in Figure 2.2.

It		is		in		fact		a		story	
NIS	:	VMF	,	I		NCO	:	D	_	NCO	;
W		W		W		A		W		A	
D		D		D		A		D		A	

of		many		years,		in		which		many	
I		J	_	NCO	:	I		W	:	J	_
W		A		A		W		U		A	
D		A		A		D		D		A	

men		tried		to		climb		that		mountain.	
NCS	:	VMF		I		VMI	,	J	_	NCO	.
A		S		W		S		A		A	
A		A		D		D		A		AD	

Figure 2.2 Example of Annotation by Step 2

- ❖ The newly added tags are written in bold in the bottom columns.

2.1.3 Step 3: Lightning Rules

The third step of the annotation rules was to form tone groups by merging blocks reducing the block boundaries and suppressing accents, based on lightning rules. Two block merger rules were given in the instructions.

- A block which has no accents cannot stand on its own as a tone group, and must be attached to a neighboring block.
- A sequence of two blocks with one accent each can be combined to form one tone group with onset and nucleus.

(Knowles 1996: 160)

There were also three accent suppressing rules, in which the number of accents were still kept up to two.

- If the first block had onset and nucleus, and if the second had only nucleus, then the nucleus of the first block would be suppressed.
- If the first block had only nucleus, and if the second had onset and nucleus, then either nucleus of the first block or the onset would be suppressed depending on their original accentual types, that is, ‘A’, was given a priority to retain its accent compared with a lower rank, ‘S’.
- Even if both blocks had onset and nucleus, they could be merged to one tone group.
 - As this third rule was ambiguous in the way of reducing accents, we considered the original accentual types first. If all of them were the same, then we uniformly suppressed nucleus of the first block and onset of the second.

In the example sentence, the first two blocks which had no accents were merged with the third block because of the first rule. Although ‘of many years’ already had both onset and nucleus, the block was merged with ‘a story’, as a transition marker between the two blocks, ‘;’, indicated their relationships were close, and because of the meaning of the context. It was not clear from the second rule which of ‘story’ and ‘many’ should be suppressed, as both of their original accentual types were ‘A’. However, it was not a crucial matter, as nucleus, ‘years’, was more important as a place where intonation would take a fall. ‘story’ was suppressed in this case, as ‘many’ seemed to have the emphasis on its meaning, and therefore to retain its accent. ‘in which’ and ‘many men’ were merged because of the first rule. ‘tried to climb’ and ‘that mountain’ were also merged, as their transition marker showed them close enough. As the original accentual type of ‘tried’ and ‘climb’ was ‘S’, and as that of ‘that’ and ‘mountain’ was a higher rank, ‘A’, the two verbs were suppressed, and ‘that’ retained its accent.

The example sentence was given a new ‘A’ and ‘D’ for those which retained accents and those which lost accents, respectively. Tagging of the example sentence from this step will be shown in Figure 2.3 with tone types in the next sub-section.

2.1.4 Step 4: Tone Types

The final step was to decide a type of each tone group. According to Knowles (1996), it was simply decided by the status of each block, that is, a transition marker of the final word in the block, as seen in Table 2.6. The final version of annotation of the example sentence is shown in the Table 2.9.

Table 2.6: Tone Types

Tag	Type	Transition Mark
F	Fall	Period '.'
FR	Fall-Rise	Colon ':'
HR	High rise	Semi-colon ';'
LR	Low rise	Comma ','
L	Level	Colligation ' '

It		is		in		fact		a		story	
NIS	:	VMF	,	I		NCO	:	D	_	NCO	;
W		W		W		A		W		A	
D		D		D		A		D		A	
D		D		D		A	 	D		D	
							FR				

of		many		years,		in		which		many	
I		J	_	NCO	:	I		W	:	J	_
W		A		A		W		U		A	
D		A		A		D		D		A	
D		A		A	 	D		D		A	
					FR						

men		tried		to		climb		that		mountain.	
NCS	:	VMF		I		VMI	,	J	_	NCO	.
A		S		W		S		A		A	
A		A		D		A		A		AD	
A	 	D		D		D		A		AD	
	FR										F

Figure 2.3 Final Version of Tagged Example Sentence

- ❖ The final accentual types and merged boundaries, ‘||’, which we added, are written with bold characters in the 2nd columns from the bottoms.
- ❖ Tone types are written in bold in bottom columns.

2.2 Result of Prosodic Annotation

This section shows results of the prosodic annotation of the 27 sentences done by following instructions of Knowles (1996). Each sentence is given with nuclei, and tone types in angle brackets, in bold. Underlines show that the tone types were cancelled by the modification in Section 2.4. The final version of the annotation to the first 10 sentences, which were used for judging German speakers’ English intonation abilities in Chapter 3, will be given in Appendix 2.

(A_01) This is the **story** <HR> of how two **men** <FR> reached the top of **Everest** <FR> on the twenty-ninth of May nineteen fifty-**three** <FR> and came back **safely** <HR> to their friends **below** <F>.

(A_02) Yet this will not be the whole **story** <F>.

(A_03) The ascent of **Everest** <FR> was not the work of one **day** <FR>, nor even of those few unforgettable **weeks** <FR> in which we prepared and climbed that **summer** <F>.

(A_04) It is in **fact** <FR> a story of many **years** <FR>, in which many **men** <FR> tried to climb that **mountain** <F>.

(A_05) The first important **expedition** <FR> was sent to **Everest** <FR> in nineteen twenty-**one** <F>.

(A_06) Then followed eleven large **expeditions** <FR>, mostly from Britain, America and **Switzerland** <F>.

(A_07) In nineteen twenty-**four** <FR> and nineteen thirty-**three** <FR>, British climbers nearly reached the **top** <F>.

(A_08) In all these **attempts** <FR>, several people **died** <F>.

(A_09) Most of these **expeditions** <FR> had tried to climb the **mountain** <FR> from the **north** <F>.

(A_10) **Then** <FR>, in nineteen forty-nine for the first **time** <FR>, foreigners were allowed to **enter** <LR> the Kingdom of **Nepal** <F>.

(A_11) This opened **up** <HR> the south side of the **mountain** <F>.

(A_12) In nineteen **fifty** <FR> a small Anglo-American **expedition** <FR> went to have a **look** <F>.

(A_13) The south side looked very hard to **climb** <FR>, but in **mountaineering** <FR> it is a **golden rule** <FR> to try, try and try **again** <FR>, however difficult the

mountain **seems** to be <F>.

(A_14) So in the next **summer** <FR> another small **expedition** <FR>, British this **time** <FR>, went out to **Everest** <F>.

(A_15) It found a possible route to the **top** <FR>, and climbed a difficult part of that **route** <F>.

(A_16) In nineteen fifty-two <FR> a Swiss expedition was **sent** <FR>, and two of the reached a **point** <FR> only three hundred metres from the **top** <FR> before they had to turn **back** <F>.

(A_17) Our own **expedition** <FR>, **therefore** <FR>, was simply the latest in a whole **series** <F>.

(A_18) The earlier **expeditions** <FR> gave us valuable **knowledge** <F>.

(A_19) The Swiss in **particular** <FR>, when they **returned** <FR>, gave us all the **information** <FR> that they had gained on **Everest** <F>.

(A_20) Our expedition to Everest <FR> was not a competition <FR>, in which we tried to do better than anyone else <F>.

(A_21) Mountaineering is not like **that** <F>.

(A_22) It is more like a relay **race** <FR>, in which each **runner** <FR> hands the baton on to the next **one** <FR> at the end of his **lap** <F>.

(A_23) The Swiss last **year** <FR> received that baton from the earlier British **climbers** <F>; and after running a brilliant **lap** <FR>, they handed the baton on to **us** <F>.

(A_24) As it **happened** <FR>, we were the last runners in this particular **race** <F>.

(A_25) But we might have failed to **finish** <F>.

(A_26) In that **case** <FR>, we would have handed on the **baton** <FR> to our French **friends** <FR> who were preparing to take up the **challenge** <F>.

(A_27) Our **opponent** <FR> was not other **climbers** <FR>, but Everest **itself** <F>.

2.3 Analysis of Annotation Result

This section analyses results of the prosodic annotation comparing with predictions mentioned in instructions of Knowles (1996). Then we discuss difficulties of the instructions which we faced in the actual annotation process.

By the prosodic annotation, the 27 sentences received 1 ‘low rise’, 3 ‘high rise’, 52 ‘fall-rise’ and 28 ‘fall’ patterns. As can be seen, many of tone types were ‘fall-rise’ or ‘fall’ patterns. As the final tone group of each sentence received a ‘fall’ pattern, most of tones appearing in the other parts of the sentence were ‘fall-rise’ patterns. Knowles (1996) predicts that his instructions produce too many ‘fall-rise’ patterns. He mentions that this tendency might be because the mapping is too simple or the scope of lightning rules needs to be improved.

We found several difficulties with the instructions after the actual annotation process was started. Firstly, it was not always straightforward to decide a grammatical tag for each word. For example, the distinction between adverb and preposition in verb phrases, such as ‘on’ in ‘each runner handed the baton on to the next one’ in a sentence ‘A_22’, was not obvious. This problem was solved by a simple rule pointed by an English language teaching researcher: if the word is moved, and if the sentence maintains its meaning, then the word is an adverb, otherwise it is a preposition. ‘on’ was an adverb in the above case, as ‘each runner handed on the baton to the next one’ still makes sense. However, non-linguists generally do not have this sort of knowledge; and even grammarians recognise this as a problematic area (e.g. Johansson et al, 1986).

Secondary, block boundaries were sometimes ambiguous, especially in the case of a verb phrase again. As ‘hand on’ is a verb phrase which has a specific meaning, ‘we would have handed on the baton’ in a sentence ‘A_26’ was divided into ‘we’, ‘would have handed on’ and ‘the baton’ leaving ‘handed’ and ‘on’ in one block in this sentence. However, the pair was broken off in the ‘A_22’, as there was a noun block ‘the baton’ between the two words.

The lightning rule was the most difficult step in the prosodic annotation. According to the rule, even if both of two successive blocks already had onset and nucleus, they were allowed to merge and to form a single tone group. This flexibility turned out to be ambiguous in the actual merging process. In the case of a sentence ‘A_10’, ‘foreigners were allowed to enter the Kingdom of Nepal’ was divided between ‘enter’ and ‘the’. The verb block, ‘were allowed to enter’ was merged with a subject, ‘foreigners’. However, in another sentence ‘A_09’, a verb block ‘had tried to climb’ was separated from its subject, ‘most of these expeditions’, because the subject was already merged from ‘most’ and ‘these expeditions’ by suppressing ‘these’. Instead, the verb block was merged with an object, ‘the mountain from the north’, which already had nucleus and onset.

Overall, many ‘fall-rise’ patterns were produced as predicted in the instructions. This tendency might be especially significant in our annotation. As mentioned in the previous paragraph, lightning rules were sometimes not clear if blocks should be merged. In many of these cases, we left them as individual tone groups. This resulted in creating more tone groups and more ‘fall-rise’ patterns.

2.4 Modification of Prosodic Annotation

Tone types produced from the prosodic annotation were compared with two native speakers’ recordings in the ISLE speech corpus. As mentioned in the previous section, we hesitated to merge blocks when the decision was not clear from lightning rules, and therefore tended to produce more tone groups than actual native speakers’ utterances.

The comparison was done for the first 10 of the annotated 27 sentences, which were used for judging German speakers' English intonation abilities. See Section 3.1 for reasons why the 10 sentences were used in the judgements. If tone types in the 10 sentences did not appear in or were different from even one of the native speakers' utterances, then the tone types were ignored in the judgments.

The tone types deleted from the 10 sentences were underlined in Section 2.2. By the comparison, 1 'low rise', 2 'high rise' and 4 'fall-rise' patterns were cancelled. 'rise' patterns produced from the prosodic annotation tended to be deleted. As a result, the remaining tone types in the 10 sentences were 15 'fall-rise' and 10 'fall' patterns.

Chapter 3

Human Evaluation of Intonation Abilities and Grouping of German Speakers

This chapter describes human evaluation of 23 German speakers' English intonations and how these speakers were grouped by the evaluation. In the previous chapter, prosodic annotation was added to the written English text of the ISLE corpus, and it was modified by a comparison with recorded native speakers' utterances. We hoped to use the intonations as a 'native speaker target', against which to compare learners' actual intonation patterns, so we investigated automated methods to extract intonation features from the learners' speech files. Unfortunately we were unable to automatically predict markup equivalent to the synthesiser cues, so could not directly compare the learners against this model.

Instead, we turned to expert human evaluation of a computational linguistics researcher (Evaluator I) and an English language-teaching (ELT) researcher (Evaluator II). These two evaluators subjectively assessed the recorded utterances from German learners of English, by listening to the recorded utterances, comparing against the 'target' marked-up script, and counting perceived intonation errors. Each evaluator made 230 judgments (10 utterances from 23 speakers), but about one third of the judgments were not agreed between two evaluators. The judgments were used to partition the speakers: the speakers were divided into two groups by taking the upper half of them, who made fewer errors based on one evaluator's judgment, to a 'good' intonation group; and the rest to a 'poor' intonation group. This partition was done for each evaluator; many speakers were categorized into the same group by both evaluators. Therefore, it can be said that this human evaluation was successful enough as a method of grouping German speakers by their intonation abilities.

These groups were separately used to train different HMMs in the HTK speech recognition experiments as explained in the next chapter. Speakers with exceptionally poor pronunciation as indicated by the ISLE corpus pronunciation markup were excluded in this grouping so that results of the following experiments would be independent from pronunciation ability. Although 23 German speakers were generally divided into the same intonation ability group by both evaluators, considering the high rate of disagreement between two evaluators, the HTK experiments were repeated three times using different two groups each time: grouping based on Evaluator I; grouping based on Evaluator II; grouping based on the agreement between two evaluators.

The amount and the process of the evaluation are explained in Section 3.1 and Section 3.2, respectively. Results from two evaluators are shown in 3.3 followed by analysis of the results in Section 3.5. Then speakers are grouped by these intonation scores in Section 3.5.

3.1 Amount of Human Evaluation

We discuss the amount of human evaluation in this section. The evaluation was not taken place to all of ‘model’ intonation patterns defined by prosodic annotation considering the workload of human evaluators. This section explains how much and how it was reduced for reasonable reasons introducing a discussion with the evaluator.

In the previous chapter, prosodic annotation was added to all 27 sentences of Block A of the ISLE corpus. This brought about putting 1 ‘low rise’, 3 ‘high rise’, 52 ‘fall-rise’ and 28 ‘fall’ patterns to the written English text before the modification by comparing with 2 native speakers’ recordings.

Evaluator I suggested that evaluation would be too time-consuming if all the patterns had to be judged for 23 speakers. Although it might be possible to receive more accurate feedback if all the patterns were judged, after the discussion with the

evaluator, the number of sentences to be assessed was reduced to 10, the first 10 of 27 sentences, for following reasons.

- After Evaluator I experimentally compared several speakers' utterances against the 'model' script, he realized that 10 sentences were enough to roughly decide a 'good' or 'poor' intonation speaker.
- This human evaluation would require intensive concentration, which would not allow evaluators to make a large number of judgments: ideally the evaluation should take less than one hour of focused concentration.

Before the discussion with Evaluator I, we planned to ask evaluators to access every intonation pattern in each sentence. However, this plan was also changed to lighten the workload of the evaluators, so that evaluator would make just one judgment to each utterance, by marking a correct if every intonation pattern of the utterance was the same as the 'model' script, and marking an error if one or more intonation patterns were different from it, in other words, number of errors in the sentence was not counted.

In the end, evaluators were required to assess the first 10 sentences of Block A of the ISLE corpus, which consisted of 8 to 30 words per sentence with the total of 157 words, and 15 fall-rise and 10 fall patterns, making one judgment on each sentence. This reduced amount of human evaluation satisfied the optimal length for the evaluators.

3.2 Process of Human Evaluation

This section explains how human evaluation was done by two evaluators. The evaluators, Evaluator I and Evaluator II, subjectively assessed intonation of 23 German spoken learners' English, by listening to their recorded utterances in the ISLE corpus, comparing against a 'native speaker target' marked-up in the previous chapter, and counting perceived intonation errors. Details of the human evaluation process are explained mentioning assessment environment and actual scoring rules.

The evaluation took place in a fairly quiet room of Leeds University School of Computing. We manipulated speech-files, in each of which utterance of one sentence was stored, playing back Microsoft 2000 Windows Media Player for Windows XP. Evaluator I directly listened to the recorded utterances, while Evaluator II used a headphone at his request.

The assessments of two evaluators were individually taken place to precede the process smoothly. The evaluators were not allowed to compare their judgments so as not to be influenced by the other until all the assessments were complete.

The judgments were preceded in the following steps.

- 1) Compare intonation patterns of each utterance against the 'model' script.
 - Evaluators were allowed to listen to the utterance repeatedly if it was necessary.
- 2) If all the intonation patterns in the utterance agreed to the 'model' script, mark a tick in a correspondent box on a provided scoring sheet. Otherwise, mark a cross in the box.
 - Evaluators were asked to add a question mark next to the tick or the cross if the decision was not confident enough.
 - As mentioned in the previous section, the number of intonation errors in each sentence was not counted. If the utterance contained error(s), one cross was marked regardless the number of errors in the sentence.
- 3) Repeat all 10 sentences of the same speaker before moving to the next one to make it easier to roughly recognise each speaker's intonation ability.

The process was efficient enough to keep the evaluators' concentration at high level throughout their judgments and to obtain steady feedback from the evaluators.

The evaluation took approximately one hour for each evaluator, which was considered as optimal time in a discussion with the evaluator.

3.3 Results from Human Evaluation

This section shows results from two human evaluators who assessed 23 German spoken learners' English intonations following the process described in the previous section. They marked a correct or an error to each utterance of the first 10 sentences of Block A of the ISLE corpus by comparing their intonation from the recordings against a 'model' pattern. Every speaker was given a result to each of these 10 utterances independently from two evaluators.

The number of sentences, in which each evaluator judged that there was at least one intonation error, is show in Table 3.1. The table also shows the numbers in optimistic and pessimistic judgments as well as how many judgments were agreed between two evaluators. The numbers in optimistic and pessimistic judgments were given by subtracting 'indefinite error marks' from and by summing 'indefinite correct marks' from the original number, respectively. See Appendix 3.1 for a score sheet from each evaluator which shows every judgment on 10 utterances for 23 speakers.

- As mentioned in the evaluation process, evaluators were allowed to put a question mark along with a correct and error mark when they were not confident enough for the decision. The former was called 'indefinite error mark' and the latter was called 'indefinite correct mark'.

Table 3.1 Number of Sentences with Intonation Error(s)

Speaker	Evaluator I				Evaluator II				Agr
	Err	Opt	Pst	Group	Err	Opt	Pst	Group	
SESS0006	6	6	6	P	5	5	5	P	7
SESS0011	1	0	2	G	4	4	4	G	7
SESS0012	6	5	8	-	3	2	3	-	5
SESS0015	5	4	6	P	1	1	2	G	6
SESS0020	0	0	2	G	5	4	5	M	5
SESS0021	0	0	0	G	7	7	7	P	3
SESS0161	1	0	3	G	5	4	5	M	5

SESS0162	3	2	5	M	3	3	3	G	8
SESS0163	6	4	6	-	3	3	3	-	5
SESS0164	4	3	4	M	4	4	6	M	8
SESS0181	5	3	6	P	6	6	6	P	5
SESS0182	3	0	5	M	4	4	4	G	7
SESS0183	7	6	8	P	7	7	7	P	6
SESS0184	4	3	5	P	9	9	9	P	5
SESS0185	1	0	2	G	4	4	4	G	5
SESS0186	3	0	5	G	4	4	4	G	7
SESS0187	4	3	4	M	5	4	5	M	9
SESS0188	2	0	3	G	2	2	2	G	8
SESS0189	6	4	8	P	10	10	10	P	6
SESS0190	3	2	5	P	5	4	5	P	8
SESS0191	4	2	5	-	4	4	4	-	4
SESS0192	4	3	6	P	5	5	6	P	6
SESS0193	0	0	2	G	4	4	5	G	9
Sum	78	50	106	-	109	104	114	-	144

- ❖ Err: Number of sentences, in which the evaluator judged intonation error(s); ‘definite errors’
- ❖ Opt = Err – (number of indefinite error sentences); ‘optimistic errors’
- ❖ Pst = Err + (number of indefinite correct sentences); ‘pessimistic errors’
- ❖ Agr: Number of sentences on which two evaluators’ judgments agreed.
- ❖ ‘G’, ‘M’ and ‘P’ represent ‘good’, ‘intermediate’ and ‘poor’ intonation groups, respectively.

➤ Grouping will be explained in Section 3.5.

3.4 Analysis of Evaluation Results

In this section, we analyse results from human evaluation shown on Table 3.1. Details of statistics based on the results are explained concerning agreement and disagreement of the two evaluators' judgments.

Two evaluators frequently disagreed with their judgments as described later in this section; however, most importantly results from human evaluation were able to roughly show intonation ability of every speaker. When 23 speakers were divided into ‘good’ or ‘poor’ intonation groups based on the human evaluation, many of these speakers were categorized into the same group by both evaluators as explained in the next section. In groupings, both groups were given 8 speakers, leaving 3 exceptionally

poor pronunciation and 4 intermediate intonation speakers. The evaluators grouped the same 7 speakers and 5 speakers into a ‘good’ and ‘poor’ intonation groups, respectively.

Agreement of two evaluators’ judgments of intonation at each utterance was not high. They assessed 10 utterances from each of 23 speakers with the total of 230 judgments. Table 3.1 shows that these evaluators’ judgments agreed on 144 sentences, which was about 63%. An extreme case was a speaker SESS0021, who received 7 intonation error marks from Evaluator II, but none from Evaluator I, the evaluators agreed on only 3 utterances. Interestingly, although their judgments on another speaker SESS0191 agreed on only 4 utterances, this speaker received 4 error marks from both evaluators. It can be explained that this speaker’s intonation patterns were almost on the border between correct and error, as Evaluator I marked 3 indefinite marks.

Disagreement of two evaluators can be also seen from the total number of intonation errors marked by each evaluator. Evaluator II marked 40% more errors compared with Evaluator I: Evaluator I marked 78 errors while Evaluator II marked 109 errors. This was probably because there was a difference of strictness in their judgment norms. In the discussion with Evaluator II after his assessment, he mentioned that German people tended not to have a clear ‘fall’ in a ‘fall-rise’ intonation pattern. Therefore, Evaluator II might mark more errors at ‘fall-rise’ patterns than Evaluator I, although it is not observable from their score sheets in Appendix 3.1 as the evaluators were not required to mark the types of intonation errors. This tendency especially appeared in several speakers such as SESS0020, SESS0021 and SESS0193, who received 4 or more error marks from Evaluator II, but none from Evaluator I.

Another difference between the evaluators was the number of indefinite marks, which were marked along with a correct or error mark when the judgments were not confident enough. The number can be calculated by subtracting ‘Opt’ from ‘Pst’ in Table 3.1. Evaluator I made indefinite judgments 56 times, which was much more frequently than Evaluator II, who marked only 10 times. During the evaluation process, Evaluator II requested us to repeat a playback of the utterance more often than Evaluator I. This resulted in less indefinite marks from Evaluator II.

Unfortunately, there were a large number of disagreed judgments on German spoken learners of English intonation between two evaluators. It seems that human judgments often do not agree as they rely on individual perceptions; in the ISLE project, 5 human annotators marked phone-level pronunciation errors, and they agreed in only 55 % of cases when deciding where and what an error was (Atwell et al, 2003). However, our human evaluation was able to roughly show intonation ability of every speaker, as many speakers were categorized into the same ability group by both evaluators. As it was the main objective of the assessment, it can be said that the human evaluation was successful in this research.

3.5 Grouping German Speakers by Intonation Abilities

This section explains how 23 German speakers were divided into ‘good’ and ‘poor’ intonation groups based on judgments of two human evaluators shown on Table 3.1. Then speakers in each group were divided into 4 sub-groups considering the balance of intonation and pronunciation abilities in these sub-groups. These groupings and sub-groupings were done as the preparation for cross-merging HTK speech recognition experiments explained in the next chapter.

Groupings of the speakers were done in three different ways based on: Evaluator I’s judgment; Evaluator II’s judgment; agreement of two evaluators’ judgments. We named these three groupings as Grouping I, Grouping II and Grouping III, respectively. Many speakers were categorized into the same intonation groups by Grouping I and II as shown in the following of this section. However, as the judgments were agreed only about 63%, and several speakers were categorized into different intonation ability groups, it was safer to do the HTK experiments with these three groupings separately, in order to obtain clear results from the experiments, and to compare three results from them.

Sub-section 3.5.1 through 3.5.3 explains the strategy of groupings. Then, Sub-section 3.5.4 through 3.5.6 shows results from three different groupings.

3.5.1 Exclusion of Pronunciation Factors

Although we could use all 23 speakers in the HTK speech recognition experiments after dividing them by intonation abilities, we considered pronunciation abilities of these speakers to keep the independence of prosodic factors in the HTK speech recognition experiments, as the main objective of overall research was analysing prosody of German spoken learners' English.

The ISLE project found that overall German speakers' English pronunciation was better than that of Italian speakers (Morton, 1999). After investigating the ability of individual German speakers by referring to word-level pronunciation error mark-up in the ISLE corpus, it turned out that the ability also varied among the 23 German speakers. Individual speaker's pronunciation data will be shown in Appendix 3.2. According to the pronunciation data, 3 speakers showed especially poor intonation ability compared with the rest of the German speakers. These 3 speakers, SESS0012, SESS0163 and SESS0191, were excluded from the HTK speech recognition experiments to minimize influences of pronunciation factors.

3.5.2 Exclusion of Intermediate Speakers

After eliminating 3 speakers who had relatively poor pronunciation, there were still 20 speakers to be grouped into a 'good' or 'poor' intonation group. The later HTK speech recognition experiments will show that recognition rate got lower as the number of training speakers became smaller. However, in order to make a clear distinction between 'good' and 'poor' intonation groups and to obtain steady results, the number of speakers for the experiments was reduced to 16 by eliminating 4 speakers, who were listed in the middle of intonation ability order of 20 speakers. These 4 'intermediate' speakers would be different in groupings based on Evaluator I and II. Exclusion of 4 intermediate speakers was done in Grouping III as the selection process was slightly different way as explained in Sub-section 3.5.6.

We decided to cut '4' speakers because:

- If more ‘intermediate’ speakers were excluded, the distinction of between the ‘good’ and ‘poor’ intonation groups would be clearer, but on the other hand, recognition rate would be lower.
- When 16 speakers were remained for groupings, each intonation group received 8 speakers, which was a convenient number for cross-merging experiments with the HTK.

3.5.3 Rules to Group Speakers

20 speakers were listed in order of ‘good’ intonation individually based on judgments of Evaluator I and II according to three priorities below. Number of intonation errors of these speakers can be seen in Table 3.1.

- 1) Fewer intonation errors.
- 2) Fewer optimistic and pessimistic intonation errors.
- 3) Balance of overall pronunciation abilities in ‘good’ and ‘poor’ intonation groups.
 - There was more than one speaker with same numbers of 1) and 2) on borders between ‘good’ and ‘intermediate’ intonation speaker groups, and between ‘intermediate’ and ‘poor’ intonation speaker groups. In the former case, speakers with poorer pronunciation speakers were chosen to a ‘good’ intonation group, and in latter case, those with better pronunciation speakers were taken to a ‘poor’ intonation group to keep the balance of pronunciation abilities in two groups as there was a little tendency that the rest of ‘good’ intonation speakers had slightly better pronunciation than that of ‘poor’ intonation speakers.

After listing 20 speakers in intonation ability order, 8 best and 8 worst speakers from the 20 speakers were categorized into a ‘good’ and ‘poor’ intonation groups leaving ‘intermediate’ 4 speakers. Finally, 8 speakers in each group were partitioned into 4 sub-groups with 2 speakers for each, considering primarily the balance of intonation abilities of the sub-groups, and secondarily that of pronunciation abilities of the sub-groups.

3.5.4 Grouping I

Grouping I was done based on Evaluator I following rules mentioned in Sub-section 3.5.3. This is a list of sub-groups with of ‘good’ and ‘poor’ intonation speakers as well as intermediate speakers for Grouping I. 4 sub-groups of the ‘good’ and ‘poor’ intonation groups were named as ‘Good A’ through ‘Good D’ and ‘Poor A’ through ‘Poor D’, respectively. Number in parenthesis next to each speaker represents an order of intonation ability among 20 speakers. More details of grouping and sub-grouping for Evaluator I will be explained in Appendix 3.3.

Good A: SESS0021 (1), SESS0186 (8)	Poor A: SESS0184 (13), SESS0183 (20)
Good B: SESS0193 (2), SESS0188 (7)	Poor B: SESS0192 (14), SESS0006 (19)
Good C: SESS0020 (3), SESS0161 (6)	Poor C: SESS0190 (15), SESS0189 (18)
Good D: SESS0011 (4), SESS0185 (5)	Poor D: SESS0181 (16), SESS0015 (17)

‘Intermediate’ speakers left out from expeditions:

SESS0162 (9), SESS0182 (10), SESS0187 (11), SESS0164 (12)

3.5.5 Grouping II

Grouping II was done based on Evaluator II and done in the same way as Grouping I. This is a list of subgroups and intermediate speakers for Grouping II. More details of grouping and sub-grouping for Evaluator II will be explained in Appendix 3.3.

Good A: SESS0015 (1), SESS0193 (8) Poor A: SESS0006 (13),
SESS0189 (20)

Good B: SESS0188 (2), SESS0182 (7) Poor B: SESS0190 (14),
SESS0184 (19)

Good C: SESS0162 (3), SESS0185 (6) Poor C: SESS0192 (15),
SESS0183 (18)

Good D: SESS0011 (4), SESS0186 (5) Poor D: SESS0181 (16),
SESS0021 (17)

‘Intermediate’ speakers: SESS0164 (9), SESS0187 (10),
SESS0161 (11), SESS0020 (12)

3.5.6 Grouping III

Grouping III was based on the agreement of Grouping I and II. 7 same speakers were categorized into a ‘poor’ intonation group by both Grouping I and II. These speakers were also grouped in ‘poor’ intonation group in Grouping III.

There were only 5 speakers who were categorized into a ‘good’ intonation group in both of the previous two groupings. In order to equalize the number of speakers in ‘good’ and ‘poor’ intonation groups, we had to add 2 more ‘good’ intonation speakers. There were 4 speakers, SESS0020, SESS0161, SESS0162 and SESS0181, who were categorized into a ‘good’ intonation group by one of the previous two groupings and an ‘intermediate’ group by the other. SESS0161 and

SESS0162, who received fewer total errors from two evaluators than the other 2 speakers, were additionally chosen for this purpose⁶.

Ideally, each group would have 8 speakers as Grouping I and II for convenience of the later cross-merging HTK experiments. However, as there were no speakers, who were categorized into a ‘poor’ group by one of the previous two groupings and into an ‘intermediate’ group by the other, no more speakers could be added to a ‘poor’ intonation group. Therefore, ‘good’ intonation group was also kept with 7 speakers.

This is a list of speakers categorized to ‘good’ and ‘poor’ intonation groups for Grouping III. Speakers in the two groups were named as ‘Good A’ through ‘Good G’, and ‘Poor A’ through ‘Poor G’, respectively.

Good A: SESS0011	Poor A: SESS0006
Good B: SESS0161	Poor B: SESS0181
Good C: SESS0162	Poor C: SESS0183
Good D: SESS0185	Poor D: SESS0184
Good E: SESS0186	Poor E: SESS0189
Good F: SESS0188	Poor F: SESS0190
Good G: SESS0193	Poor G: SESS0192

⁶ SESS0020 received the total errors as many as SESS0161 and SESS0162. However, a difference of two judgments on SESS0020 was the largest among the 3 speakers: Evaluator II marked 6 errors, while Evaluator I marked none. SESS0020 was not chosen for this fact.

Chapter 4

The HTK Speech Recognition Experiments for Analysing Prosody

This chapter explains the HTK speech recognition experiments to analyse German spoken learners' English prosody. German speakers grouped by their intonation abilities in the previous chapter were separately used to train the HMMs and to test the recognition against the trained models. Before starting experiments for intonation analysis, we investigated training models changing acoustic and language models as well as several HTK parameters to find the most reasonable model for the main experiments. As training data were limited, we also made several other pre-experiments to check the relationship between the amount of training data and recognition results reducing the number of training speakers each time.

We investigated two procedures of the HTK experiments: calling every HTK module from a command line; and module-call by provided Perl scripts with configuration files. The main prosodic experiments were repeated changing training and test speakers with different intonation abilities. The use of the script and the configuration files were suitable for the purpose as all they required was changing lists of speakers each time without considering other options.

The purpose of the first HTK experiment was to investigate training models and several parameters for recognition tests. For triphone models, we tested against 1-mixture (single density), 2-mixture and 3-mixture Gaussian component models. We considered two HTK parameters which might have significant influences on recognition results: Grammar Scale Factor (GSF) decides the amount of dependence on a language model in recognition tests; Word Insertion Penalty (WIP) is a fixed value given to each token from the end of a word to the beginning of the next (Young et al, 2001). After investigating training models with different parameters, we decided

to use single density monophone and 3-mixture triphone models with setting GSF to 0.0 and WIP to -60.0 in the later main experiment.

In the second experiment, the HMMs were repeatedly trained and tested reducing training speakers from 17 to 12 and 6, which was the actual number of training speakers in the main experiment. We observed that recognition results gradually went down as the number of the training speakers became smaller. However, the recognition accuracy was not exorbitantly lower even when the HMMs trained with only 6 speakers were used for testing.

Finally, the HTK was used to analyse English prosody of German speakers. The HMMs were separately trained with 6 ‘poor’ and ‘good’ intonation speakers and each model was tested against the remaining speakers from the both groups. This was done via cross-validation, repeating the experiment 4 times for Grouping I and II, and 7 times for Grouping III, taking out a different test-subset each time. Recognition results were generally higher when test speakers’ intonation ability was the same as that of training speakers. Cross-merging experiments confirmed that these results were consistent.

The Perl scripts and configuration files used in the HTK experiments are mentioned in Section 4.1. Section 4.2 describes the HTK experiments for investigating training models and parameters, while Section 4.3 shows the relationship between reduction of training speakers and recognition accuracies. Cross-validation experiments for investigating prosody are explained in Section 4.4, and the analysis of the results follows in Section 4.5.

4.1 Script and Configuration Files for the HTK Experiments

This section explains Perl scripts and configuration files used in the HTK experiments. We also investigated the use of the HTK by calling each module from a command line following Chapter 3 of the HTK Book ‘A Tutorial Example of Using HTK (Young et al, 2001). It allowed setting many options of the HTK tools; however,

it was time-consuming for our experiments, in which changes of training speakers were the main requirements. The use of the Perl scripts and configuration files was more suitable for the purpose without considering the options. These files and the permission for their experimental use were provided by the Institute für Elektronik, Signalverarbeitung und Kommunikationstechnik of Otto-von-Guericke-Universität Magdeburg⁷ in Germany.

A main Perl script invokes the HTK tools with two configuration files separately created for training and recognition tests. The former configuration was used to train monophone and triphone HMMs and the latter was to test the recognition against the models. The other Perl scripts and configuration files were also defined on the two main configuration files for minor tasks such as making directories, and organizing headers of models.

Sub-Section 4.1.1 and 4.1.2 explains commands defined by the configuration files for training and recognition test, respectively. Then Sub-Section 4.1.3 shows requirement files for the configuration files.

4.1.1 Configuration File for Training

This is a list of main tasks executed via the configuration file for training.

- Convert speech data to a parameterised form for the HTK.
- Calculate the global mean and covariance of a set of training data.
- Create a triphone-level pronunciation dictionary from a monophone dictionary.
- Create the Master Label Files with monophone and triphone transcriptions (MLFs).

⁷ World Wide Web: <<http://iesk.et.uni-magdeburg.de/KO/wendemu>>.

- Perform basic Baum-Welch re-estimation of the parameters to create monophone and triphone HMMs.

Several parameters on the configuration file were fixed as on the provided configuration file.

- Type of target parameterized vectors: 12 Mel-Frequency Cepstral Coefficients
- Number of the HMM states: 3
 - More precisely, this was defined in one of required files shown in Sub-Section 4.1.3.
- Single density monophone and 1-, 2- and 3- Gaussian mixture models were created for triphone models.
 - For each of these models, re-estimation of parameters was repeated 4 times.

4.1.2 Configuration File for Recognition Test

This is a list of main tasks executed via the configuration file for training.

- Match parameterized test data against trained HMMs, and output a recognised word-level transcription.
- Compare the recognised transcription against an MLF, and analyse its word-level recognition accuracy as well as number of correctly recognised, deleted, substituted and inserted word errors.

Several parameters on the configuration file were investigated and fixed at following values in the main experiment for prosodic analysis; See Section 4.2 and 4.3 for the investigation of these parameters.

- Grammar Scale Factor (GSF) = 0.0
- Word Insertion Penalty (WIP) = -60.0

4.1.3 Files Required

Following is a list of files required for the scripts and configuration files to execute training and recognition test with the HTK. File names are given in parentheses and details of the files are explained in Appendix 4.1.

- Scenario file: this file contained a written text of all the recorded utterances.
(all_sent.snr)
- Word list: this was a list of all the words which appeared in the scenario file.
(wordlist)
- Monophone pronunciation dictionary: the dictionary consisted of all the words on the word list and their monophone-level pronunciation transcriptions. (mono.dic)
- Monophone list: this list contained all the monophones which appeared in the monophone pronunciation dictionary. (monphon.lst)
 - Each monophone was considered as a single HMM.
- HMM definition: The file defined the number of the HMM states as 3. (hmm.def)
- Lists of all wave files and feature files: Parameterized form of each wave file was named as this list. (w2f.all)

- List of all feature files: This was the list of all the parameterized feature files. (feat.all)
- List of feature training files: The file was the list of the parameterized feature files of training data. (feat.train)
- List of feature test files: This was the list of the parameterized feature files of test data. (feat.test)
- List of all labels: This list was all the label names for the both training and test data. (text.all)
- 0-gram lattice: This was a word network used instead of a 0-gram language model. (zero.lat)
- 2-gram lattice: This was a word network used instead of a 2-gram language model. (bigr.lat)

4.2 Training Model and Parameter Investigation

The first HTK experiment was to investigate training models and several parameters for recognition test. Conditions for this experiment in this section are mentioned in Sub-Section 4.2.1. The GSF parameter is examined in Sub-Section 4.2.2, while the WIP parameter is investigated in Sub-Section 4.2.4 after discussing various training models in Sub-Section 4.2.3.

4.2.1 Experimental Conditions

The HTK experiment in this section was executed with following conditions.

- Training speakers: randomly selected 17 speakers

SESS0161, SESS0162, SESS0163, SESS0164, SESS0181, SESS0182,
SESS0183, SESS0184, SESS0185, SESS0186, SESS0187, SESS0188,
SESS0189, SESS0190, SESS0191, SESS0192, SESS0193

- Test speakers: The remaining 6 speakers

SESS0006, SESS0011, SESS0012, SESS0015, SESS0020, SESS0021

- Number of the HMM states: 3
- Number of re-estimation of parameters: 4
- Vocabulary size: 1,310 words with 411 unique words
 - A word list from which the HTK looked for the best matched word with each segment of the test utterance was created from a scenario file which was the same for both training and recognition test. As a result, the list did not contain any extra words from the vocabulary of test speakers' utterances.
 - 411 words includes two words of silence models for a sentence beginning, '!ENTER', and ending, '!EXIT'.
 - However, '!ENTER' and '!EXIT' were ignored from statistics of recognition results.

4.2.2 Grammar Scale Factor (GSF)

GSF is an HTK parameter set in recognition test and decides the dependency on a word network defined by a lattice format (Young et al, 2001). The language model plays an important role in speech recognition (Becchetti and Ricotti, 1999; Rodman,

1999); however, it would unfairly improve the recognition accuracy in our experiments, as the same scenario was used for training and recognition tests and as the vocabulary size was not large. Such an influence would also give less clear results from the main HTK experiment for prosodic analysis. Moreover, acquiring higher recognition accuracy was not the purpose of the main experiment. Therefore, unless the recognition result became unreasonably low, it was better not to use any language model, that is, GSF should be fixed at 0.0.

Recognition tests were done against single density monophone and 1-, 2- and 3-mixture Gaussian component triphone HMMs setting GSF to 0.0 and 10.0. When GSF was 10.0, both 0-gram and 2-gram lattices were tested. WIP was fixed at -10.0 all the time. Table 4.1 shows recognition results.

Table 4.1 Affect of GSF to Recognition Accuracy

Training Models	GSF		
	0.0	10.0 (0-gram)	10.0 (2-gram)
Monophone	26.44%	22.82%	73.59%
1-mixture triphone	58.92%	56.26%	93.33%
2-mixture triphone	62.95%	59.92%	94.68%
3-mixture triphone	63.80%	61.21%	94.53%

According to results shown in Table 4.1, recognition accuracy was much higher against any training models by at least 30 %, when GSF was set to 10.0 and a 2-gram lattice was used, compared with when it was set to 0.0. However, as the latter setting still brought about slightly higher recognition accuracies than when GSF was set to 10.0 and a 0-gram lattice was used, setting GSF to 0.0 turned out to be reasonable enough. All the following experiences were examined fixing GSF at 0.0.

4.2.3 Training Models

In the previous experiment in Sub-Section 4.1.1, each GSF setting was tested against four different trained HMMs: single density monophone and 1-, 2- and 3-mixture

Gaussian mixture component triphone HMMs. Based on the results in Table 4.1, we selected two types of HMMs to train in later experiments.

Against triphone HMMs, recognition accuracy was higher as the number of Gaussian mixture was bigger, except one case in which a 2-gram lattice with GSF at 10.0 was used for tests against 2- and 3-mixture triphone HMMs. Although the HTK was able to train more than 3-mixture HMMs, considering the small improvement and increasing training time as more mixture would be used in the HMMs, following experiments were executed with 3-mixture Gaussian component triphone HMMs⁸.

Monophone HMMs brought about much lower results compared with triphone HMMs in all cases. However, to compare results against monophone and triphone HMMs from the main experiments for prosodic analysis, monophone models were still used in later experiments.

4.2.4 Word Insertion Penalty (WIP)

WIP is another HTK parameter for recognition tests and a value given to each token from the end of a word to the beginning of the next (Young et al, 2001). To investigate its influence and find out the best value, we tested against monophone and 3-mixture triphone HMMs with WIP from -50.0 to 10.0 increasing the value by 10.0 each time.

Table 4.2 and 4.3 show recognition results with various WIPs against triphone and monophone HMMs, respectively. They indicate the numbers of recognised, deleted, substituted and inserted words as well as recognition accuracies.

According to the two tables, WIP did not have a significant influence on recognition accuracies, but on the number of deleted and inserted words against both

⁸ In the following, ‘monophone’ and ‘triphone’ mean ‘single density monophone’ and ‘3-mixture Gaussian component triphone’, respectively, unless stated otherwise.

monophone and triphone HMMs. As speakers did not delete or insert these words in their actual recordings, it was reasonable to set WIP value which kept the balance of these two errors. When WIP was at -40.0, the number of the two errors was best balanced in triphone case, although -20.0 was the value for monophone case and recognition accuracy was the highest and almost the highest for monophone and triphone cases, respectively, when it was set to -20.0.

WIP was examined again after the number of training speakers were reduced to 6 as explained in Sub-Section 4.3.3.

Table 4.2 Affect of WIP to Recognition Accuracy against Triphone HMMs

WIP	Rate	Rec	Del	Sub	Ins
-50.0	62.77%	4934	717	2209	440
-40.0	63.60%	4999	568	2293	568
-30.0	64.08%	5037	428	2395	755
-20.0	64.07%	5036	323	2501	989
-10.0	63.80%	5015	226	2619	1394
0.0	63.09%	4959	177	2724	2005
10.0	61.45%	4830	127	2903	3003

Table 4.3 Affect of WIP to Recognition Accuracy against Monophone HMMs

WIP	Rate	Rec	Del	Sub	Ins
-50.0	24.35%	1914	1625	4321	233
-40.0	25.15%	1977	1357	4526	351
-30.0	25.73%	2022	1090	4748	564
-20.0	26.49%	2082	761	5017	922
-10.0	26.44%	2078	488	5294	1485
0.0	26.13%	2054	312	5494	2510
10.0	25.41%	1997	232	5631	4205

❖
$$\text{Rate} = \text{Rec} / (\text{Total number of tested words}) * 100 [\%]$$

➤ Total number of tested words was 7860 (1310 from 6 test speakers).

❖ Rec: Number of correctly recognised words.

❖ Del: Number of deleted words.

❖ Sub: Number of substituted words.

❖ Ins: Number of inserted words.

4.3 Reduction of Training Speakers

The second HTK experiment was to observe the influence of reducing training data to recognition accuracy. In the previous HTK experiment, the HMMs were trained with 17 randomly chosen training speakers, but the main experiment for prosodic analysis explained in the next section would use only 6 training speakers. If the amount of training data is insufficient, it might not be possible to estimate some of the model parameters robustly (Thambiratnam, 2001). As training and test speakers recorded the same scenario, this problem might not be occurred in our experiments; however we confirmed by an experiment that recognition accuracy would not be enormously lower as the number of training speakers was reduced.

After confirming that 6 training speakers still brought about a reasonable recognition accuracy, another experiment was undertaken to adjust the WIP parameter. When the HMMs were trained with 17 speakers, optimal WIP was between -20.0 and -40.0. The experiment found that -60.0 was best for the HMMs trained with 6 speakers.

Sub-Section 4.3.1 states the conditions of the HTK experiments in this section. The experiment of training speaker reduction is described in Sub-Section 4.3.2, and the WIP parameter is examined in Sub-Section 4.3.3.

4.3.1 Experimental Conditions

Experiments in this section were executed under the same conditions as those mentioned in Sub-Section 4.2.1, except:

- Number of training speakers was reduced;
- WIP was fixed at -20.0 and -40.0, unless stated otherwise.

4.3.2 Reduction of Training Speakers

The experiment was repeated reducing the number of training speakers from 17 to 12 and 6. The 12 and 6 speakers were randomly chosen from the 17 training speakers shown in the previous section. The experiment was done setting WIP to -20.0 and -40.0. Each of the trained HMMs was tested against the same 6 speakers as the previous section. Table 4.4 and 4.5 show results from the both experiments.

Table 4.4 Reduction of Training Data and

Table 4.5 Reduction of Training Data and

Recognition Accuracy: WIP = -20.0

Recognition Accuracy:
WIP = -40.0

Training Models	Number of Training Speakers		
	17	12	6
Monophone	26.49%	26.42%	22.51%
Triphone	64.07%	58.77%	47.01%

Training Models	Number of Training Speakers		
	17	12	6
Monophone	25.15%	25.23%	21.54%
Triphone	63.60%	58.49%	46.44%

- ❖ 12 training speakers: SESS0161, SESS0162, SESS0163, SESS0164, SESS0181, SESS0182, SESS0183, SESS0184, SESS0185, SESS0186, SESS0187, SESS0188
- ❖ 6 training speakers: SESS0161, SESS0162, SESS0163, SESS0164, SESS0181, SESS0182

According to the two tables, recognition accuracy gradually went down as the number of speakers was reduced for both WIP values: Results were almost same against monophone models trained with 17 and 12 speakers, but became about 4 % lower when the number of training speakers was reduced to 6; Reduction of recognition accuracy bigger against triphone models with about 5~6 % and 8~11 % when the number was reduced from 17 to 12 and from 12 to 6.

Although reduction of recognition accuracy was observed when the number of training speakers became smaller, it was a gradual change, and the result against

models from 6 training speakers was not significantly lower. As a result, we could confirm that 6 speakers would be enough to train the HMMs for the main HTK experiment in the next section.

4.3.3 Adjustment of WIP

This sub-section shows that the reduction of training data affected a WIP value which would balance the number of deleted and inserted words, and it describes another experiment executed to find out the best WIP for recognition test against the HMMs trained with 6 speakers.

Table 4.6 and 4.7 show reveal that there were more inserted words and less deleted word errors as the number of training speakers became smaller. Results in both tables were from triphone models. The balance of the two errors was better when WIP was set to -40.0 compared with -20.0.

We tested the same triphone model trained with 6 speakers setting WIP to -60. Table 4.8 shows that the balance of two errors became better, although there were more inserted words. The balance should be even better against the triphone model if WIP was set slightly lower. However, considering that the well balanced WIP value against monophone models was lower than that for triphone models, it was decided that WIP would be fixed at -60.0 against both monophone and triphone models in the main HTK experiment.

Table 4.6 Reduction of Training Data and Data and

Word Error Types: WIP = -20.0

Speaker	Rate	Rec	Del	Sub	Ins
17	64.07%	5036	323	2501	989
12	58.77%	4619	291	2950	1331
6	47.01%	3695	271	3894	2572

Table 4.7 Reduction of Training

Word Error Types: WIP = -40.0

Speaker	Rate	Rec	Del	Sub	Ins
17	63.60%	4999	568	2293	568
12	58.49%	4567	509	2754	804
6	46.44%	3647	490	3716	1744

Table 4.8 Balance of Word Error Types: WIP = -60.0

Speaker	Rate	Rec	Del	Sub	Ins
6	45.12%	3543	803	3507	1169

- ❖ Speaker: Number of training speakers.
- ❖ Rate = $\text{Rec} / (\text{Total number of tested words}) * 100 [\%]$
 - Total number of tested words was 7860 (1310 from 6 test speakers).
- ❖ Rec: Number of correctly recognised words.
- ❖ Del: Number of deleted words.
- ❖ Sub: Number of substituted words.
- ❖ Ins: Number of inserted words.

4.4 The HTK Experiments for Prosodic Analysis

This main HTK experiments were to analyse prosody of 23 samples of German spoken learners' English in the ISLE speech corpus. Then, 23 speakers were grouped by their intonation abilities in Chapter 3. Grouping was done three times: Grouping I based on Evaluator I; Grouping II based on Evaluator II; Grouping III based on agreement by both evaluators. Three independent experiments were done using one of the three groupings each time.

In Grouping I and II, there were 8 'good' and 8 'poor' intonation speakers, who divided into 4 'good' and 4 'poor' sub-groups with 2 speakers in each. In experiments with the two groupings, the HMMs were trained with 3 'good' and 3 'poor' sub-groups separately, and each model was tested with the other 1 sub-group from both intonation groups. This process was repeated 4 times for each grouping.

In Grouping III, there were 7 'good' and 7 'poor' intonation speakers. Similarly to Grouping I and II, the HMMs were trained with 6 'good' and 'poor' intonation speakers separately, and each model was tested with the other 1 'good' and 1 'poor' intonation speakers. This experiment was repeated 7 times.

The results reveal that recognition accuracy became higher when the HMMs are trained with sub-groups of the same intonation ability as test speakers. Cross-merging experiments confirm that these results are generally consistent.

Sub-Section 4.4.1 states experimental conditions in this section. Sub-Section 4.4.2 through 4.4.4 show the experiments based on Grouping I through III, respectively.

4.4.1 Experimental Conditions

Experiments in this section were executed under the same conditions as those mentioned in Sub-Section 4.2.1, except:

- Number of training speakers was 6.
- Number of test speakers was 2 or 1 from each intonation ability group.
- WIP was fixed at -60.0 unless there is a specification

4.4.2 Experiment I – Grouping I

This HTK experiment was done using training and test speakers from Grouping I based on Evaluator I, which was shown in Sub-Section 3.5.4. The experiment was repeated 4 times using different sub-groups of training and test speaker groups of Grouping I as follows.

- 1) a: Training speakers: Good B, C and D Test speakers: Good A and Poor A
 b: Training speakers: Poor B, C and D Test speakers: Good A and Poor A
- 2) a: Training speakers: Good A, C and D Test speakers: Good B and Poor B
 b: Training speakers: Poor A, C and D Test speakers: Good B and Poor B

- 3) a: Training speakers: Good A, B and D Test speakers: Good C and Poor C
 b: Training speakers: Poor A, B and D Test speakers: Good C and Poor C
- 4) a: Training speakers: Good A, B and C Test speakers: Good D and Poor D
 b: Training speakers: Poor A, B and C Test speakers: Good D and Poor D

Table 4.9 and 4.10 show recognition accuracies of ‘good’ and ‘poor’ intonation test speakers, respectively, against triphone HMMs. Table 4.11 and 4.12 are those against monophone HMMs.

Table 4.9 Result: ‘Good’ Intonation Speakers **Table 4.10** Result: ‘Poor’ Intonation Speakers

(Triphone Models) <1>

Test Speakers		Training Speakers		
		Good	Poor	Impr I
Good A	SESS0021	38.57%	50.69%	-12.12%
	SESS0186	56.71%	42.07%	14.64%
Good B	SESS0193	55.49%	48.02%	7.47%
	SESS0188	56.94%	43.22%	13.72%
Good C	SESS0020	49.39%	30.56%	18.83%
	SESS0161	36.63%	20.24%	16.39%
Good D	SESS0011	46.65%	33.92%	12.73%
	SESS0185	61.66%	52.29%	9.37%
Average		50.26%	40.13%	10.13%

(Triphone Models) <1>

Test Speakers		Training Speakers		
		Good	Poor	Impr II
Poor A	SESS0184	25.00%	42.91%	17.91%
	SESS0183	19.05%	41.92%	22.87%
Poor B	SESS0192	50.20%	52.44%	2.24%
	SESS0006	31.77%	43.00%	11.23%
Poor C	SESS0190	36.97%	48.40%	11.43%
	SESS0189	56.71%	38.19%	-18.52%
Poor D	SESS0181	34.38%	54.34%	19.96%
	SESS0015	51.91%	38.87%	-13.04%
Average		38.25%	45.01%	6.76%

Table 4.11 Result: ‘Good’ Intonation Speakers **Table 4.12** Result: ‘Poor’ Intonation Speakers

(Monophone Models) <1>

Test Speakers		Training Speakers		
		Good	Poor	Impr I
Good A	SESS0021	26.60%	25.23%	1.37%
	SESS0186	38.26%	22.18%	16.08%
Good B	SESS0193	34.68%	23.09%	11.59%
	SESS0188	38.95%	22.94%	16.01%
Good C	SESS0020	34.22%	15.40%	18.82%
	SESS0161	25.49%	8.31%	17.18%
Good D	SESS0011	31.86%	19.74%	12.12%
	SESS0185	36.43%	24.01%	12.42%
Average		33.31%	20.11%	13.20%

(Monophone Models) <1>

Test Speakers		Training Speakers		
		Good	Poor	Impr II
Poor A	SESS0184	21.95%	20.43%	-1.52%
	SESS0183	8.92%	17.30%	8.38%
Poor B	SESS0192	26.98%	24.09%	-2.89%
	SESS0006	17.46%	17.08%	-0.38%
Poor C	SESS0190	26.37%	21.42%	-4.95%
	SESS0189	33.92%	18.45%	-15.47%
Poor D	SESS0181	19.59%	25.08%	5.49%
	SESS0015	27.82%	16.39%	-11.43%
Average		22.88%	20.03%	-2.85%

- ❖ Impr I = Good – Poor
- ❖ Impr II = Poor – Good

According to the results of triphone HMMs, recognition accuracies were generally higher when intonation abilities of training speaker groups and test speaker were matched, except for one ‘good’ intonation speaker, SESS0021, and two ‘poor’ intonation speakers, SESS0015 and SESS0189. Average percentages of recognition accuracies were about 10 % and 7 % among ‘good’ and ‘poor’ intonation test speakers, respectively. If speakers with negative improvement were excluded, the average was almost same for both speaker groups at 13.31 % and 14.27 % for ‘good’ and ‘poor’ intonation groups, respectively.

In monophone case, the improvement was observed from all ‘good’ intonation test speakers and average of their improvements was 13.20 %. However, only two ‘poor’ intonation test speakers, SESS0181 and SESS0183, who had the highest improvement in triphone case among 8 ‘poor’ intonation speakers, showed the same tendency in monophone case too.

4.4.3 Experiment II – Grouping II

This experiment was done using training and test speakers from Grouping II based on Evaluator II, which was shown in Sub-Section 3.5.5. The experiment was also repeated 4 times in the same way as the previous experiment.

Table 4.13 and 4.14 show recognition accuracies of ‘good’ and ‘poor’ intonation test speakers, respectively, against triphone HMMs. Table 4.15 and 4.16 are those against monophone HMMs.

Table 4.13 Result: ‘Good’ Intonation Speakers **Table 4.14** Result: ‘Poor’ Intonation Speakers

		(Triphone Models) <2>			(Triphone Models) <2>				
		Training Speakers			Training Speakers				
Test Speakers		Good	Poor	Impr I	Test Speakers		Good	Poor	Impr II
Good A	SESS0015	54.80%	19.21%	35.59%	Poor A	SESS0006	26.84%	46.77%	19.93%
	SESS0193	59.60%	36.97%	22.63%		SESS0189	59.45%	22.71%	-36.74%
Good B	SESS0188	52.90%	43.22%	9.68%	Poor B	SESS0190	25.84%	45.66%	19.82%
	SESS0182	49.09%	25.91%	23.18%		SESS0184	18.67%	43.29%	24.62%
Good C	SESS0162	57.55%	37.42%	20.13%	Poor C	SESS0192	37.88%	59.45%	21.57%
	SESS0185	60.59%	51.68%	8.91%		SESS0183	15.93%	44.21%	28.28%
Good D	SESS0011	44.74%	33.92%	10.82%	Poor D	SESS0181	25.91%	54.34%	28.43%
	SESS0186	57.62%	34.98%	22.64%		SESS0021	36.20%	54.27%	18.07%
Average		54.61%	35.41%	19.20%	Average		30.84%	46.34%	15.50%

Table 4.15 Result: ‘Good’ Intonation Speakers **Table 4.16** Result: ‘Poor’ Intonation Speakers

		(Monophone Models) <2>			(Monophone Models) <2>				
		Training Speakers			Training Speakers				
Test Speakers		Good	Poor	Impr I	Test Speakers		Good	Poor	Impr II
Good A	SESS0015	30.72%	10.37%	20.35%	Poor A	SESS0006	15.08%	18.31%	3.23%
	SESS0193	34.30%	19.44%	14.86%		SESS0189	36.66%	13.11%	-23.55%
Good B	SESS0188	35.37%	24.24%	11.13%	Poor B	SESS0190	22.03%	23.70%	1.67%
	SESS0182	41.84%	18.22%	23.62%		SESS0184	17.23%	21.04%	3.81%
Good C	SESS0162	35.98%	21.49%	14.49%	Poor C	SESS0192	19.21%	26.98%	7.77%
	SESS0185	34.60%	25.23%	9.37%		SESS0183	8.38%	18.83%	10.45%
Good D	SESS0011	31.33%	19.74%	11.59%	Poor D	SESS0181	18.45%	25.08%	6.63%
	SESS0186	34.60%	16.54%	18.06%		SESS0021	21.27%	24.62%	3.35%
Average		34.84%	19.41%	15.43%	Average		19.79%	21.46%	1.67%

❖ $\text{Impr I} = \text{Good} - \text{Poor}$, $\text{Impr II} = \text{Poor} - \text{Good}$

Improvement of recognition accuracy by matching of intonation abilities between training and test speakers was observed from every speaker in both triphone and monophone cases except one ‘poor’ intonation speaker, ‘SESS0189’. Especially in the triphone case, the improvement was double that with Experiment I; 10.13 % to 19.20 % and 6.76 % to 15.50 % in average among ‘good’ and ‘poor’ intonation speakers. SESS0189 was also chosen as a ‘poor’ intonation speaker in Grouping I and showed negative improvements in Experiment I too. If SESS0189 was excluded, the average of the improvement among ‘poor’ intonation speakers was 22.96 %, which was even higher than 19.20 % of ‘good’ intonation speakers. In the monophone case,

the tendency was also observed as much as Experiment I among ‘good’ intonation speakers. The tendency was seen from ‘poor’ intonation speakers, but the improvement was very low.

4.4.4 Experiment III – Grouping III

This experiment was done using training and test speakers from Grouping III based on agreement of Evaluator I and II, which was shown in Sub-Section 3.5.6. As there were only 7 ‘good’ and 7 ‘poor’ intonation speakers in this grouping, and as each HMM was still trained with 6 speakers, 1 speaker, instead of 2, from both groups was used for recognition test. The experiment was repeated 7 times taking a different pair of test speakers each time.

Table 4.17 and 4.18 show recognition accuracies of ‘good’ and ‘poor’ intonation test speakers, respectively, against triphone HMMs. Table 4.19 and 4.20 are those against monophone HMMs.

Table 4.17 Result: ‘Good’ Intonation Speakers **Table 4.18** Result: ‘Poor’ Intonation Speakers

		(Triphone Models) <3>					(Triphone Models) <3>		
		Training Speakers					Training Speakers		
Test Speakers		Good	Poor	Impr I	Test Speakers	Good	Poor	Impr II	
Good A	SESS0011	46.04%	33.31%	12.73%	Poor A	SESS0006	27.92%	45.77%	17.85%
Good B	SESS0161	36.16%	18.35%	17.81%	Poor B	SESS0181	24.31%	54.34%	30.03%
Good C	SESS0162	52.97%	38.26%	14.71%	Poor C	SESS0183	19.13%	44.74%	25.61%
Good D	SESS0185	62.42%	48.93%	13.49%	Poor D	SESS0184	23.93%	43.90%	19.97%
Good E	SESS0186	54.80%	20.27%	34.53%	Poor E	SESS0189	56.25%	19.60%	-36.65%
Good F	SESS0188	53.05%	43.06%	9.99%	Poor F	SESS0190	26.98%	49.09%	22.11%
Good G	SESS0193	60.23%	43.75%	16.48%	Poor G	SESS0192	40.26%	53.35%	13.09%
	Average	52.24%	35.13%	17.11%		Average	31.25%	44.40%	13.14%

Table 4.19 Result: ‘Good’ Intonation Speakers **Table 4.20** Result: ‘Poor’ Intonation Speakers

(Monophone Models) <3>

Test Speakers		Training Speakers		
		Good	Poor	Impr I
Good A	SESS0011	31.86%	21.27%	10.59%
Good B	SESS0161	25.25%	9.02%	16.23%
Good C	SESS0162	33.84%	19.74%	14.10%
Good D	SESS0185	37.73%	23.93%	13.80%
Good E	SESS0186	36.28%	10.52%	25.76%
Good F	SESS0188	36.43%	22.33%	14.10%
Good G	SESS0193	40.09%	19.82%	20.27%
Average		34.50%	18.09%	16.41%

(Monophone Models) <3>

Test Speakers		Training Speakers		
		Good	Poor	Impr II
Poor A	SESS0006	16.62%	17.46%	0.84%
Poor B	SESS0181	20.12%	25.08%	4.96%
Poor C	SESS0183	8.77%	17.97%	9.20%
Poor D	SESS0184	17.99%	20.73%	2.74%
Poor E	SESS0189	35.67%	11.97%	-23.70%
Poor F	SESS0190	22.64%	21.65%	-0.99%
Poor G	SESS0192	22.41%	25.99%	3.58%
Average		20.60%	20.12%	-0.48%

❖ $\text{Impr I} = \text{Good} - \text{Poor}$

❖ $\text{Impr II} = \text{Poor} - \text{Good}$

Similar to the previous two experiments, agreement of training and test speakers’ intonation abilities brought about higher recognition accuracies almost as much as in the previous two experiments, except SESS0189 in both triphone and monophone cases and SESS0190 in monophone case. The average of the improvements among ‘poor’ intonation speakers in the triphone case was about 4 % lower than that of ‘good’ intonation speakers; however, if SESS0189 was excluded from the former group, the average became 21.44 %, which was more than 4 % higher than that of the latter group.

The average improvement of ‘poor’ intonation speakers in the monophone case was low again; even if SESS0189 and SESS0190 were excluded, the average was still 4 % among the other 5 test speakers.

SESS0189 showed the negative improvement in both triphone and monophone cases of Experiment I and II too. SESS0190 also had the negative improvement in monophone case of Experiment I, although this speaker showed positive in monophone case of Experiment II. However the positive improvement was only 1.67%

4.5 Dominance of Prosodic Factors and Irrelevance of Pronunciation Abilities

This section confirms the independence of prosodic factors and irrelevance of pronunciation abilities in the results from the main HTK experiments of the previous section. These explanations are given taking Grouping II, whose experiment showed the most significant improvement of recognition accuracy by the agreement of training and test speakers' intonation abilities.

Sub-Section 4.5.1 shows the average improvement of recognition accuracy among 'keywords' for prosody was more significant than that of overall shown in Sub-Section 4.4.3. Sub-Section 4.5.2 explains another HTK experiment using 2 'worst' pronunciation speakers from a 'good' intonation group and 2 'best' pronunciation speakers from a 'poor' intonation group.

4.5.1 Recognition of Prosodic 'Keywords'

This sub-section shows that the improvement of recognition accuracy by matching ability of training and test speakers observed in the main HTK experiments was especially significant among 'keywords' for intonation.

Research on speech timing by White (2002) found that the locus of accentual lengthening was shown to be the word, with the greatest lengthening tending to be at word edges. Although other features of prosody such as loudness and tempo might have influence over multiple words, we focused on the word containing the last accented syllable of each tone group with a respect to the accentual lengthening, and called it a 'keyword'.

As mentioned in Sub-Section 4.1.2, the HTK outputs a recognised word-level transcript for each recognition test. We counted the number of correctly recognised

keywords against triphone models of the first 10 sentences of Block A, A_01 through A_10, whose written text was prosodically annotated and tone types were modified by comparing with 2 native speakers' recordings. The 10 sentences contained 15 'fall-rise' and 10 'fall' patterns as can be seen in Section 2.2. In practice, the last accented syllable was in the last word of each tone group throughout the 10 sentences.

Figure 4.1 is an example of a 'recognised words' file of a sentence A_04 from a 'poor' intonation speaker, 'SESS0006', against triphone HMMs trained by 6 'good' intonation speakers in Sub-Section 4.4.3. Its original sentence and tone types were 'It is in fact a story of many **years** <FR>, in which many **men**<FR> tried to climb that **mountain** <F>.' By comparison, keywords 'men' for a 'fall-rise' pattern and 'mountain' for a 'fall' pattern were correctly recognised, while the other keyword 'years' was not recognised.

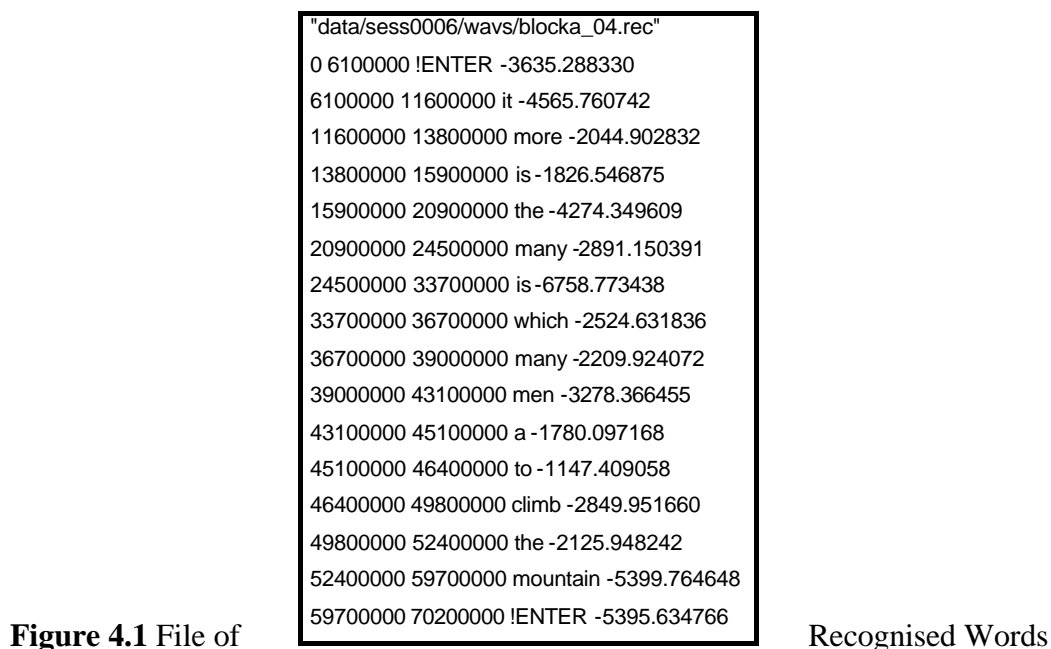


Table 4.21 and Table 4.22 show the percentages of recognised words among 15 keywords for 'fall-rise', 10 keywords for 'fall' and their total of 25 keywords for 'good' and 'poor' test speakers, respectively.

Table 4.21 Result of Recognised Keywords: ‘Good’ Intonation Test Speakers

Test Speakers		Fall-Rise			Fall			Total		
		Training speakers		Impr I (F-R)	Training speakers		Impr I (F)	Training Speakers		Impr I (Total)
		Good	Poor		Good	Poor		Good	Poor	
Good	SESS0015	66.67%	13.33%	53.34%	80.00%	20.00%	60.00%	72.00%	16.00%	56.00%
A	SESS0193	73.33%	46.67%	26.66%	80.00%	70.00%	10.00%	76.00%	56.00%	20.00%
Good	SESS0188	46.67%	33.33%	13.34%	60.00%	40.00%	20.00%	52.00%	36.00%	16.00%
B	SESS0182	73.33%	33.33%	40.00%	60.00%	40.00%	20.00%	68.00%	36.00%	32.00%
Good	SESS0162	73.33%	46.67%	26.66%	60.00%	20.00%	40.00%	68.00%	36.00%	32.00%
C	SESS0185	73.33%	53.33%	20.00%	40.00%	80.00%	-40.00%	60.00%	64.00%	-4.00%
Good	SESS0011	73.33%	40.00%	33.33%	80.00%	70.00%	10.00%	76.00%	52.00%	24.00%
D	SESS0186	73.30%	40.00%	33.30%	70.00%	40.00%	30.00%	71.98%	40.00%	31.98%
Average		69.16%	38.33%	30.83%	66.25%	47.50%	18.75%	68.00%	42.00%	26.00%

Table 4.22 Result of Recognised Keywords: ‘Poor’ Intonation Test Speakers

Test Speakers		Fall-Rise			Fall			Total		
		Training speakers		Impr II (F-R)	Training speakers		Impr II (F)	Training Speakers		Impr II (Total)
		Good	Poor		Good	Poor		Good	Poor	
Poor	SESS0006	40.00%	73.33%	33.33%	10.00%	70.00%	60.00%	28.00%	72.00%	44.00%
A	SESS0189	73.33%	26.67%	-46.66%	60.00%	50.00%	-10.00%	68.00%	36.00%	-32.00%
Poor	SESS0190	26.67%	66.67%	40.00%	20.00%	60.00%	40.00%	24.00%	64.00%	40.00%
B	SESS0184	13.33%	33.33%	20.00%	20.00%	60.00%	40.00%	16.00%	44.00%	28.00%
Poor	SESS0192	46.67%	86.67%	40.00%	30.00%	60.00%	30.00%	40.00%	76.00%	36.00%
C	SESS0183	6.67%	60.00%	53.33%	10.00%	50.00%	40.00%	8.00%	56.00%	48.00%
Poor	SESS0181	20.00%	66.70%	46.70%	40.00%	50.00%	10.00%	28.00%	60.02%	32.02%
D	SESS0021	73.33%	66.67%	-6.66%	40.00%	50.00%	10.00%	60.00%	60.00%	0.00%
Average		37.50%	60.01%	22.51%	28.75%	56.25%	27.50%	34.00%	58.50%	24.50%

❖ Impr I = Good – Poor, Impr II = Poor - Good

Firstly, average improvements on keywords were 26.00 % and 24.50 % among 8 ‘good’ and 8 ‘poor’ intonation speakers, both of which were higher than those on overall recognition, 19.20 % and 15.50 % as shown on Table 4.13 and 4.14, respectively. This result showed that trained models were clearly distinguished by prosodic features.

Secondary, the average improvement on keywords for ‘fall-rise’ patterns among ‘good’ intonation speakers was much higher at 30.83 % than that for ‘fall’ patterns at 18.75 %., while the reverse result was obtained from ‘poor’ intonation speakers. This indicates that the ‘poor’ intonation speakers did not have clear ‘fall-

rise' patters as perceived by Evaluator II. It might be also because there was a diversity of errors for 'fall-rise' patterns among 'poor' intonation speakers.

Thirdly, negative improvements were observed from several speakers including SESS0189 again. SESS0185 also showed the overall ne gative improvement. However it, was only 4.00 %, and this is too small to be significant given from the sample size of 25 keywords.

4.5.2 Experiment to Prove Irrelevance of Pronunciation Abilities

This HTK experiment explains the irrelevancy of pronunciation abilities from the results of the main experiments. There was a tendency that 'good' intonation speakers had slightly better pronunciation scores than 'poor' intonation speakers, although speakers with exceptionally 'poor' pronunciation speakers had been excluded before the experiments. In Grouping II, average number of word-level pronunciation errors out of 1,083 words was 5.25 among 'good' intonation speakers and 10.50 among 'poor' intonation speakers. See Appendix 3.2 for individual speakers' pronunciation scores.

In this experiment, two 'worst' pronunciation speakers, SESS0182 (9) and SESS0186 (13), among 'good' intonation speakers, and two 'best' pronunciation speakers, SESS0181 (4) and SESS0192 (3), from 'poor' intonation speakers of Grouping II were used as test speakers. The number of pronunciation error words of each test speaker was given in parenthesis. Models were separately trained by the other 6 'good' and 6 'poor' intonation speakers. The other experimental conditions were the same as mentioned in Sub-Section 4.4.1.

Table 4.23 and 4.24 show recognition accuracies of 'good' and 'poor' intonation test speakers, respectively, against triphone HMMs. Table 4.25 and 4.26 are those against monophone HMMs.

Table 4.23 Result: ‘Good’ Intonation Speakers **Table 4.24** Result: ‘Poor’ Intonation Speakers

(Triphone Models) <4>

Test Speakers		Training Speakers		
		Good	Poor	Impr I
Good	SESS0182	50.69%	29.73%	20.96%
	SESS0186	58.16%	35.75%	22.41%

(Triphone Models) <4>

Test Speakers		Training Speakers		
		Good	Poor	Impr II
Poor	SESS0181	21.11%	53.66%	32.55%
	SESS0192	38.57%	57.09%	18.52%

Table 4.25 Result: ‘Good’ Intonation Speakers **Table 4.26** Result: ‘Poor’ Intonation Speakers

(Monophone Models) <4>

Test Speakers		Training Speakers		
		Good	Poor	Impr I
Good	SESS0182	40.55%	29.73%	20.43%
	SESS0186	37.35%	35.75%	18.29%

(Monophone Models) <4>

Test Speakers		Training Speakers		
		Good	Poor	Impr II
Poor	SESS0181	20.35%	23.78%	3.43%
	SESS0192	26.68%	27.74%	1.06%

- ❖ Impr I = Good – Poor
- ❖ Impr II = Poor – Good

Higher recognition accuracy was observed in this experiment with a slightly higher tendency than the previous result based on Grouping II. The improvement was barely seen from ‘poor’ intonation speakers in monophone case as also seen in the previous result. This result successfully confirmed that the improvement was regardless of test speakers’ intonation abilities.

4.6 Analysis of Experiment Results

Experiment I through III observed relationships between German spoken learners’ English intonation abilities and recognition accuracies with the HTK. Similar results were generally obtained from the three experiments.

- Matching of English intonation abilities between training speakers and test speakers generally brought about higher recognition accuracies. (See Table 4.11 through 4.24)
- Improvement of the accuracies was higher in Experiment II (Table 4.13 through 4.20) compared with Experiment I (Table 4.9 through 4.12).
- One ‘poor’ intonation speaker, SESS0189, always showed the negative improvements in recognition tests.
- Improvement was relatively small from ‘poor’ intonation speakers in monophone case compared with their results in triphone case and those from ‘good’ intonation speakers in both triphone and monophone cases.

The fact that agreement of speakers’ intonation abilities generally brought about higher recognition accuracies implies these two conclusions: the HTK speech recogniser was clearly affected by intonation; and the human evaluator was consistent in his judgments.

The HTK was able to distinguish ‘good’ and ‘poor’ intonation abilities. According to Werner and Keller (1994), the main acoustic parameters bearing in prosody are fundamental frequency, intensity and duration in speech recognition. Young (2001) mentions that each wave form is converted to a sequence of parameter vectors equally spaced by the specific duration. In linguistic terms, duration is one the three main prosodic features along with pitch and loudness (Cruttenden, 1997) As tone groups were linked to the number of accented syllables in prosodic annotation of Knowles (1996), intonation is correlated with accented syllables. Accented syllables generally have longer duration than unaccented syllables (Bolinger, 1965; Werner and Keller 1994). In our experiments, each monophone was treated as a single HMM, and the model was trained by comparing speaker’s utterance with its monophone-level transcription. This means that if the speaker has a ‘good’ intonation pattern, then the keywords will be correctly stressed and hence lengthened. Longer duration will secure a better match with ‘good’ models.

The HTK was able to train such distinguished HMMs because the evaluator was consistent in his judgments. Experiment I and II used Grouping I based on Evaluator I and Grouping II based on Evaluator II, respectively. There were several speakers whose intonation ability groups were differently categorized by the two evaluators. However, as their decision norms were always consistent, these grouped speakers generally had similar intonation characteristics. This led to train the HMMs trained with speakers with similar duration of utterance for each phone, and eventually to almost consistent results of higher recognition accuracies by matching intonation abilities of these speakers.

Experiment II showed higher improvement of recognition accuracies compared with Experiment I. One explanation is that Evaluator II has been more trained to listen to foreign speakers' English utterances as an EFL researcher compared with Evaluator I, a computational linguistics researcher. This might result in that Grouping II was more finely categorized with speakers' English abilities.

More importantly, there was a difference in norms of the two evaluators' judgments. As mentioned in Section 3.4, Evaluator II commented that German speakers tended not to have a clear 'fall' in a 'fall-rise' pattern. This might affect groupings of speakers, such as SESS0020 and SESS0021, who were categorized into a 'good' intonation group by Evaluator I, but an 'intermediate' or a 'poor' intonation group by Evaluator II. As a result, a 'good' intonation group in Grouping I should be slightly different from that in Grouping II. However, recognition accuracies were higher in both experiments when 'good' intonation test speakers used the HMMs trained by a 'good' intonation group, as each evaluator was consistent throughout his judgement. That is, two 'good' intonation groups were different in Grouping I and II, but each group had its consistent characteristic of intonation.

Two speakers, SESS0189 and SESS0190 were chosen as 'poor' intonation speakers in three groupings, but their recognition accuracies became generally lower against the HMMs trained by the rest of the 'poor' intonation groups, except SESS0190 in Experiment II. As both evaluators categorized them into 'poor' intonation groups, their English intonations were truly different from a 'native speaker

target'. However, they might have different types of intonation errors from the rest of 'poor' intonation speakers, which led to the negative results of the two speakers.

Types of intonation errors were varied among the rest of 'poor' intonation speakers as there was a relatively low improvement of recognition accuracies in their monophone case. However, the diversity should not be so significant that the average improvement of 'poor' intonation speakers except SESS0189 and SESS0190 was as high as that of 'good' intonation speakers in triphone case.

To summarize, the HTK was clearly affected by German speakers' English intonation abilities. Recognition results were generally higher when intonation abilities of training and test speakers were matched, and this was confirmed via cross-validation in all three experiments with the independency from pronunciation factors. The human evaluator's consistency in his judgment was confirmed by the experimental results, although there was a slight difference in judgment norms between the two evaluators. It was observed that there was a variation in error types of English intonation among German speakers.

Chapter 5

Conclusion

This chapter summarizes results of this research: prosodic annotation, human evaluation of German speakers' intonation abilities, and the HTK experiments to analyse their prosody. Then, we explain contributions of the work and discuss future directions of the research. Paper based this research (Oba and Atwell, 2003) will be presented to the International Conference on Corpus Linguistics (CL-2003).

Section 5.1 summarizes the results, and Section 5.2 and 5.3 explain the contributions and the future work, respectively.

5.1 Summary

We exploited the HTK speech recogniser to analyse German spoken learners' English prosody. The HTK experiments succeeded in distinguishing training models, which would bring about higher recognition accuracy, by matching the abilities of training and test speakers' intonation abilities of English.

Sub-section 5.1 and 5.2 summarize prosodic annotation and human evaluation of 23 German speakers' English intonation abilities for grouping the speakers, respectively, which were undertaken as preparation for the HTK experiments explained in Sub-Section 5.3.

5.1.1 Prosodic Annotation

Chapter 2 explained prosodic annotation of a written English transcription of 27 spoken sentences recorded from 23 German speakers in the ISLE corpus. This annotation was done by following the set of instructions of Knowles (1996), to predict ‘model’ prosodic patterns.

The 27 sentences consisted of 429 words and were divided into 84 tone groups: 1 ‘low rise’, 3 ‘high rise’, 52 ‘fall-rise’ and 28 ‘fall’ patterns. The prosodic annotation produced many ‘fall-rise’ patterns as predicted in the instructions. Tone types of the first 10 sentences, which were used for evaluating German speakers’ intonation abilities, were modified by comparing against 2 native speakers’ recordings. The 10 sentences consisted 157 words retained 15 ‘fall-rise’ and 10 ‘fall’ patterns after canceling 1 ‘low rise’, 2 ‘high rise’ and 4 ‘fall-rise’ patterns by the modification.

The modification was undertaken, as we tended to create more tone groups than actual native speakers’ utterances. This is due to an ambiguity of the annotation instructions; when it was not clear from the instructions if the two successive blocks should be merged to form a single tone group, we left them as individual tone groups in most such cases. However, the instructions generally required simple mappings such as grammatical tags to the degrees of accentual types, which could be handled by non-linguists such as ourselves.

5.1.2 Human Evaluation and Grouping

Chapter 3 described human evaluation of 23 German speakers’ English intonation abilities for grouping them into ‘good’ and ‘poor’ intonation groups. Two evaluators, one computational linguistics researcher (Evaluator I) and one ELT researcher (Evaluator II), listened to the 10 utterances from each speaker, and compared their prosodic patterns against ‘model’ tone types from the annotation. If the evaluator perceived that all the tone types of each utterance were the same as the model patterns, the utterance was marked as ‘correct’; otherwise ‘error’. We separately counted the number of ‘errors’ for every speaker marked by each evaluator.

The agreement of two evaluators' judgments was at about 63 %. Evaluator II's judgment was stricter; this evaluator marked 109 errors out of 230 judgments (10 utterances from 23 speakers), while Evaluator I marked 78 errors. This was probably due to the difference of their judgments norms. Evaluator II mentioned that German speakers tended not to have a clear 'fall' in a 'fall-rise' pattern. This should result in marking more errors due to 'fall-rise'; however it can not be seen from the score sheet, as one judgment was made for one utterance.

As evaluators' judgments did not agree in some cases, three different groupings were done to 23 German speakers: Grouping I based on Evaluator I, Grouping II based on Evaluator II and Grouping III based on the agreement of the two evaluators. Before the groupings, 3 exceptionally 'poor' speakers were eliminated, so that the following HTK experiments should not be affected by pronunciation factors. In Grouping I and II, top 8 and bottom 8 speakers were categorized into 'good' and 'poor' intonation ability groups leaving 4 intermediate speakers each time. 5 'good' and 7 'poor' speakers were in the same groups in both groupings. In Grouping III, 7 'good' and 7 'poor' speakers were grouped by adding 2 speakers, who were categorized as 'good' by one evaluator and 'intermediate' by the other, into the 5 agreed 'good' speakers. Despite the high rate of disagreement from two evaluators, many speakers were categorized into the same intonation ability groups. Therefore, it can be said that human evaluation was successful enough.

5.1.3 The HTK Experiments

Chapter 5 explained the HTK speech recognition experiments for analysing German speakers' English prosody. For the main experiments, following terms were decided by preparation HTK experiments.

- Monophone and 3-mixture word internal triphone HMMs would be trained.

- No language models would be used for recognition tests.
- Recognition accuracy was still reasonable, even when the number of training speakers were reduced from 17 to 12 and 6.
- WIP would be set to -60.0.

The three main experiments were undertaken using one of the three groupings: Experiment I through III taking Grouping I through III, respectively. The HMMs were trained with 6 ‘good’ and 6 ‘poor’ speakers separately, and each model was tested by the rest of speakers from both groups. This was repeated by taking different sets of test speakers each time.

In all of the three experiments, recognition accuracy was generally higher except ‘poor’ test speakers against monophone models when training and test speakers’ intonation abilities were the same. Experiment II showed the most significant improvement, whose grouping was based on Evaluator II, who had a stricter norm on ‘fall-rise’ patterns. Average improvement of recognition accuracies by the agreement in triphone cases (with monophone cases) were:

- Experiment I: 10.13 % (13.20 %) among ‘good’ test speakers
6.76 % (-2.85 %) among ‘poor’ test speakers
- Experiment II: 19.20 % (15.43 %) among ‘good’ test speakers
15.50 % (1.67 %) among ‘poor’ test speakers
- Experiment III: 17.11 % (16.41 %) among ‘good’ test speakers
13.14 % (-0.48 %) among ‘poor’ test speakers

The improvement was lower from ‘poor’ test speakers. This was because of one speaker, SESS0189, who was categorized into a ‘poor’ intonation group by both evaluators, but always had much higher recognition accuracy against models trained by ‘good’ intonation speakers. This must be because this speaker had different

intonation error types from the other ‘poor’ speakers, while the rest of ‘poor’ speakers created similar intonation errors.

Although exceptionally ‘poor’ pronunciation speakers were excluded from the groupings, the following two support experiments gave supporting evidence that the above results were obtained by intonation abilities. These two experiments were done taking Grouping II, which showed the most significant improvement in previous experiments.

We counted correctly recognised ‘keywords’ for tone types. White (2002) found that the locus of accentual lengthening was shown to be the word, with the greatest lengthening tending to be at word edges. We called the word containing the last accented syllable of each tone group a ‘keyword’. Improvement of recognition accuracy among the ‘keywords’, especially for ‘fall-rise’ patterns, was higher than that among all the words. This result showed that trained models were clearly distinguished by prosodic features, and ‘poor’ intonation speakers tended to show the difficulties at ‘fall-rise’ patterns as perceived by Evaluator II.

The other experiment was done taking two ‘worst’ and ‘best’ pronunciation speakers from ‘good’ and ‘poor’ intonation groups, as the former group tended to have slightly better scores on ‘pronunciation’ abilities. This result also showed the improvement when training and test speakers’ intonation abilities agreed. This confirmed the result is not relevant to pronunciation factors.

Overall, we can conclude that the HTK was able to train clearly different HMMs according to training speakers’ intonation abilities. We found that it was better to use models trained by speakers with the same intonation ability as the test speakers in order to achieve higher recognition accuracy, and that German speakers who showed ‘poor’ English intonation abilities, generally had similar errors.

5.2 Contributions

5.2.1 Contribution to Speech Recognition Research

Our research focussed on analysing non-native speakers' prosody using the HTK speech recogniser. As mentioned in chapter 1, the main focus of speech recognition researchers is generally on pronunciation factors, or if prosody is taken into account, they tend to deal with native speakers', that is, this research was unique, and important; we showed that a test German speaker should choose a model trained by other German speakers with the same intonation abilities as the test speaker, in order to obtain higher recognition accuracy. Therefore, it is worth considering intonation abilities for speech recognisers for non-native speakers.

5.2.2 Contribution to HTK Research

There has not been research which used the HTK speech recogniser to directly analyse prosody. Our work proved that the HTK was able to deal with prosodic factors. The HTK trained two clearly different HMMs: those trained with similar prosodic patterns to native speakers' 'model' patterns; and those trained with common intonation error patterns among German speakers. Speech recognition researchers can deal with prosody using the HTK.

5.2.3 Contribution to Foreign Language Learning

Our research suggests that foreign language learning software should be able to detect learners' intonation abilities unlike any existing educational software. The learning tool should contain different models separately trained by 'good' and 'poor' intonation speakers. By comparing recognition accuracies of 'keywords' for prosody

against the two models, it should be possible to detect the accuracy of the learner's intonation and to point out intonation patterns where the learner especially showed weakness.

Fox (1984) and Grabe (1998) compared English and German intonations, and revealed that German language rarely had a similar 'fall-rise' pattern to that of English. One of our experiments implied that German speakers with 'poor' English intonation tended to have errors at 'fall-rise' patterns. German speakers of learning English require intensive practice of the 'fall-rise' pattern.

5.3 Future Work

This research successfully showed that the agreement of training and test speakers' intonation abilities, 'good' or 'poor', brought about higher recognition accuracy. The intonation abilities were judged at only 'fall-rise' and 'fall' patterns; however, there are also other tone types, such as 'rise' and 'level'. This suggests that further investigations are required:

- Whether the same grouping would be given when all the tone types were taken into account in human evaluation of German speakers' English intonation abilities;
- If not, whether the different grouping would also shows the improvement of recognition accuracy by the agreement of intonation abilities.

We also need to consider the diversity of intonation errors. In the HTK experiments, one 'poor' intonation speaker showed the opposite result; recognition accuracy was better when models trained by 'good' intonation speakers were tested against this speaker. This was probably because this speaker had different types of intonation errors from the rest of 'poor' intonation speakers. Speech recognition should deal with this kind of exceptional speaker when it takes account of intonation abilities.

In this research, only German speakers were considered. It is worth investigating whether the same results would be obtained from other nationalities, and the possibility of adjusting the idea into multilingual speech recognition, in which there should be diversity even within the same intonation group because of influences from different mother languages.

A significant challenge is to use these results in real language-teaching systems. A lesson from the ISLE project is that theoretical results and practical of these results are quite different achievements!

References

Atwell, E., Dementriou, G., Hughes, J., Schiffrin, A., Souter, C., and Wilcock, S. (2000a) A comparative evaluation of modern English corpus grammatical annotation schemes. In: *International Computer Archive of Modern and Medieval English (ICAME) Journal*, vol.24, pp.7-23.

Atwell, E., Herron, D., Howarth, P., Morton, R. and Wick, H. (1999) *Pronunciation Training: requirements and solutions*, Project Report, Interactive Spoken Language Education Project LE4-8353, Deliverable D1.4. Cambridge: Entropic.

Atwell, E., Howarth, P., and Souter, C. (2003) The ISLE Corpus: Italian and German Spoken Learners' English In: *International Computer Archive of Modern and Medieval English (ICAME) Journal*, vol.27.

Atwell, E., Howarth, P., Souter, C., Baldo, P., Bisiani, R., Pezzotta, D., Bonaventura, P., Menzel, W., Herron, D., Morton, R., and Schmidt, J. (2000b) User-Guided System Development in Interactive Spoken Language Education. In: *Natural Language Engineering journal, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*, vol.6 (3-4), pp.229-241.

Becchetti, C. and Ricotti, L.P. (1999) *Speech Recognition: Theory and C++ Implementation*. Chichester: John Wiley & Sons Ltd.

BEEP dictionary (1996) [Online]. Available from World Wide Web:
<<http://www-svr.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html>>.

Berkling, K., Zissman, M., Vonwiller, J., and Cleirigh, C. (1998) Improving Accent Identification through Knowledge of English Syllable Structure In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, vol.2, 30 November-4 December 1998, Sydney. pp.89-92.

Brown, G. (1977) *Listening to Spoken English*. London: Longman.

Chen, W. and Warren, M. (1999) Facilitating a description of intercultural conversations: the Hong Kong Corpus of Conversational English In: *International Computer Archive of Modern and Medieval English (ICAME) Journal*, vol.23. pp.5-20.

Chomsky, N. (1957) *Syntactic Structures*. The Hague: Mouton

Chomsky, N. (1965) *Aspects of the Theory of the Syntax*. Cambridge: M.I.T. Press.

Cruttenden, A. (1997) *Intonation*, 2nd edition. Cambridge: Cambridge University Press.

Eskenazi, M. (1996) Detection of foreign speakers' pronunciation errors for second language training In: *Proceedings of the 4th International Conference on Spoken*

Language Processing (ICSLP-1996), vol.3, 3-6 October 1996, Philadelphia. pp.1465-1468.

Eskenazi, M. (1999) Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype In: *Language & Technology (LLT) Journal*, vol.2 (2), pp.62-76.

Evermann, G. (2002) *HTK History* [Online]. Available from World Wide Web: <<http://htk.eng.cam.ac.uk/history.shtml>>.

Fox, A. (1984) *German Intonation*. Oxford: Clarendon Press.

Grabe, E. (1998) *Comparative Intonational Phonology: English and German*, PhD thesis. Max-Planck-Institute for Psycholinguistic and University of Nijmegen.

Hansen, J.H.L. and Arslan, L.M. (1995) Foreign Accent Classification Using Source Generator Based Prosodic Features In: *Proceedings of the 1995 International Conference on Acoustic, Speech, and Signal Processing (ICASSP-1995)*, vol.1, 9-12 May 1995, Detroit. pp.836-839.

HTK (2000) [Online]. Available from World Wide Web: <<http://htk.eng.cam.ac.uk/index.shtml>>.

Hunt, J. (1996) *The Ascent of Everest*. Stuttgart: Ernst Klett Verlag, English Readers Series.

Interactive Spoken Language Education Non-Native Speech Data (1999) [CD-ROM]. Cambridge: Entropic.

Johansson, S., Atwell, E., Garside, R., and Leech, G. (1986) *The Tagged LOB Corpus – User Manual*. Bergen: Norwegian Computing Centre for the Humanities.

Jurafsky, D and Martin, J.H. (2000) *Speech and Language Processing*. New Jersey: Pearson Higher Education.

Jurafsky, D., Wooters, C., Tajchman, G., Segel, J., Stolcke, A., Folser, E., and Morgan, N. (1994) The Berkeley Restaurant Project In: *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP-1994)*, September 1994, Yokohama. pp.2139-2142.

Knowles, G. (1987) *Patterns of Spoken English*. London: Longman.

Knowles, G. (1996) From text structure to prosodic structure. In: Knowles, G., Wichman, A. and Alderson, P., (editors). *Working with Speech*. Harlow: Addison Wesley Longman Limited. pp. 146-167.

Leech, G. (1992) Corpus Linguistics and Theories of Linguistic Performance. In: Svartvik, J., (editor). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: Mouton de Gruyter. pp. 105-122.

- MATE: Project Overview* (1998) [Online]. Available from World Wide Web: <<http://mate.nis.sdu.dk/about/summery.html>>.
- Matsuda, T., (general editor). (1999) *Kenkyusha's English-Japanese Dictionary for the General Reader*, 2nd edition. [CD-ROM] Tokyo: Kenkyusha.
- McEnery, T. and Wilson (1999) *Corpus Linguistics*, 2nd edition. Edinburgh: Edinburgh University Press
- Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., and Souter, C. (2000) The ISLE Corpus of non-native spoken English. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (editors). *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, vol.2, 31 May-2 June 2000, Athens. pp.957-964.
- Morton, R. (1999) *Recognition of Learner Speech*, Project Report, Interactive Spoken Language Education Project LE4-8353, Deliverable D3.3. Cambridge: Entropic.
- Oba, T. and Atwell, E. (2003) Using the HTK Speech Recogniser to analyse prosody in a Corpus of German Spoken Learners' English. In: *Proceedings of the 2003 International Conference on Corpus Linguistics (CL- 2003)*, 28-31 March 2003, Lancaster.
- O'Connor, J.D. and Arnold, G.F. (1970) *Intonation of Colloquial English*, 7th edition. London: Longman.
- Rodman, R. D. (1999) *Computer Speech Technology*. Norwood: Artech House Inc.
- Souter, C., Howarth, P., and Atwell, E. (1999) *Speech Data Collection and Annotation*, Project Report, Interactive Spoken Language Education Project LE4-8353, Deliverable D3.1. Cambridge: Entropic.
- Stemmer, G., Nöth, E., and Niemann, H. (2001) Acoustic Modeling of Foreign Words in a German Speech Recognition System In: Dalsgaard, P., Lindberg, B., and Benner, H., (editors). *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, vol.4, 3-7 September 2001, Aalborg. pp.2745-2748.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992) ToBI: A Standard for labeling English Prosody In: *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-1992)*, 12-16 October 1992, Banff. pp.867-870.
- Svartvik, J., (editor). (1990) *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.
- Taylor, J. and Knowles, G. (1998) *Manual of Information to accompany the SEC Corpus*. Available from World Wide Web: <<http://khnt.hit.uib.no/icame/manuals/sec>>

Taylor, P., King, S., Isard, S. and Wright, H. (1998) Intonation and Dialog Context as Constraints for Speech Recognition. In: *Language and Speech*, vol.41 (3-4), pp.493-512.

Taylor, P., King, S., Isard, S., Wright, H., and Kowtko, J. (1997) Using Intonation to Constrain Language Models in Speech Recognition. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-1997)*, vol.5, 22-25 September 1997, Rhodes. pp.2763-2766.

Teixeira, C., Trancoso, I., and Sarralheiro, A. (1996) Accent Identification. In: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-1996)*, vol.3, 3-6 October 1996, Philadelphia. pp.577-580.

Tench, P. (1996) *The Intonation Systems of English*. London: Cassell.

Thambiratnam, D. (2001) *[HTK-Users] Problem with HERest* [Online].

Available from World Wide Web:

<<http://htk.eng.cam.ac.uk/pipermail/htk-users/2001-August/001145.html>>.

Uebler, U., Schüßler, M., and Niemann, H. (1998) Bilingual and Dialectal Adaptation and Retraining In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, vol.15, 30 November-4 December 1998, Sydney. pp.1815-1818.

Warwick, C. (1997) *What is the BNC?* [Online]. Available from World Wide Web: <http://www.hcu.ox.ac.uk/BNC>>.

Werner, S and Keller, E. (1994) Prosodic Aspects of Speech. In Keller, E. (editor) *Fundamentals of Speech Synthesis and Speech Recognition*. Chichester: John Wiley & Sons Ltd. pp.23-40.

White, L. (2000) *English speech timing: a domain and locus approach*, PhD thesis. University of Edinburgh.

White, L. and King, S. (2003) *EUSTACE Corpus* [Online]. Available from World Wide Web: <<http://www.cstrr.ed.ac.uk/projects/eustace>>.

Witt, S. and Young, S. (1997) Language Learning Based on Non-Native Speech Recognition In: *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-1997)*, vol.2, 22-25 September 1997, Rhodes. pp.633-636.

Woodland, P. (2000) *HTK History* [Online]. Available from World Wide Web:

<<http://htk.eng.cam.ac.uk/history.shtml>>.

Yan, Q. and Vaseghi, S. (2002) A Comparative Analysis of UK and US English Accents in Recognition and Synthesis In: *Proceedings of the 2002 International Conference on Acoustic, Speech, and Signal Processing (ICASSP-2002)*, vol.1, 13-17 May 2002, Florida. pp.413-416.

Young, S., Everman, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland P. (2001) *The HTK Book 3.1*. Cambridge: Entropic.

Appendix 2

Final Version of Prosodic Annotation

This appendix shows the final version of prosodic annotation to the English text of the ISLE speech corpus, which was explained in Chapter 2. The annotation was done to all 27 sentences of Block A of the corpus; however, only the first 10 sentences, which were used for judging German speakers' intonation abilities in Chapter 3, are given in this appendix.

(A_01)

This		is		the		story		of		how	
NIS	:	VMF	,	D	_	NCO	;	I		W	;
S		W		W		A		W		U	
A		D		D		AD		D		D	
A		D		D		AD		D		D	
							HR				

two		men		reached		the		top		of	
J	_	NCS	:	VMF	,	D	_	NCO	;	I	
A		A		S		W		A		W	
A		A		A		D		A		D	
A		A		A		D		D		D	
			FR								

Everest		on		the		twenty		ninth		of	
NPO	:	I		D	_	K	_	K	;	I	
A		U		W		A		A		W	
ADD		D		D		AD		AD		D	
ADD		D		D		AD		D		D	
	FR										

May		nineteen		fifty		three		and		came	
NPO		K	_	K	_	K	:	C	?	VMF	_
A		A		A		A		W		S	
A		D		D		A		D		A	
D		D		D		A		D		A	

							FR				
back		safely		to		their		friends		below.	
A	;	A	;	I		NIS	_	NCO	:	A	.
A		A		W		U		A		A	
A		AD		D		D		A		DA	
D		AD		D		D		A		DA	
			HR								F

- ‘reached the top of Everest’: ‘the top’ and ‘of Everest’ were merged first because of the meaning of the context, although a transition marker between ‘reached’ and ‘the top’ was closer than that of between the former two blocks. Then, by the first merging rule, ‘reached’ was merged with ‘top of the Everest’ by suppressing nucleus of the left block, ‘reached’, as its original accentual type, ‘S’, was weaker than that of onset of the right block, which was ‘A’ of ‘top’.

(A_02)

Yet		this		will		not		be		the	
C	?	NIS	:	VAI		X		VMI	,	D	_
S		S		%		A		W		D	
A		A		D		A		D		D	
D		D		D		A		D		D	

whole		story.	
J	_	NCO	.
A		A	
A		AD	
D		AD	
			F

- ‘will not be’ and ‘the whole story’ were merged first as the transition marker was the closest one. According to the second block merger rule, either nucleus of the former block, ‘not’, or onset of the latter, ‘whole’ should be suppressed. As the both accentual types were the same, considering the meaning of the sentence, ‘not’ retained its accent. Then, the block was merged with ‘this’ and ‘yet’. In each case, the left block was suppressed, as its accentual type was weaker than that of the right block.

- Conjunctions were not clear from the instructions at several points: whether a conjunction should be given an independent block; which transition marker should be given. However, as accentual types of conjunctions were up to ‘S’; most of the cases ‘W’ or ‘U’, conjunctions were generally suppressed and merged with neighboring block(s).

(A_03)

The		ascent		of		Everest		was		not	
D	_	NCS	;	I		NPO	:	VMF	,	X	;
W		A		W		A		W		A	
D		DA		D		ADD		D		A	
D		DA		D		ADD		D		A	
							FR				

the		work		of		one		day,		nor	
D	_	NCO	;	I		J	_	NCO	:	C	?
W		A		W		A		A		S	
D		A		D		A		A		A	
D		D		D		D		A		D	
									FR		

even		of		those		few		unforgettable		weeks	
A	;	I		J	_	J	_	J	_	NCO	:
A		W		A		A		A		A	
A		D		A		D		D		A	
A		D		D		D		D		A	
											FR

in		which		we		prepared		and		climbed	,
I		W	:	NIS	:	VMF	,	C	?	VMF	,
W		U		W		S		W		S	
D		D		D		DA		D		A	
D		D		D		D		D		D	

that		summer.	
J	_	NCO	.
A		A	
A		AD	
A		AD	
			F

(A_04)

It		is		in		fact		a		story	
NIS	:	VMF	,	I		NCO	:	D	_	NCO	;
W		W		W		A		W		A	
D		D		D		A		D		A	
D		D		D		A		D		D	
							FR				

of		many		years,		in		which		many	
I		J	_	NCO	:	I		W	:	J	_
W		A		A		W		U		A	
D		A		A		D		D		A	
D		A		A		D		D		A	
					FR						

men		tried		to		climb		that		mountain.	
NCS	:	VMF		I		VMI	,	J	_	NCO	.
A		S		W		S		A		A	
A		A		D		A		A		AD	
A		D		D		D		A		AD	
	FR										F

(A_05)

The		First		important		expedition		was		sent	
D	_	J	_	J	_	NCS	:	VAF	_	VMP	,
W		A		A		A		W		S	
D		A		D		DDAD		D		A	
D		A		D		DDAD		D		A	
							FR				

to		Everest		in		nineteen		twenty		one.	
I		NPO	:	I		K	_	K	_	K	.
W		A		W		A		A		A	
D		ADD		D		DA		D		A	
D		ADD		D		DA		D		A	
			FR								F

(A_06)

Then		followed		eleven		large		expeditions,		mostly	
A	;	VMF	,	J	_	J	_	NCS	:	A	;
A		S		A		A		A		A	
A		A		AD		D		DDAD		AD	
A		D		D		D		DDAD		AD	
									FR		

from		Britain,		America		and		Switzerland.			
I		NPO	:	NPO	:	C	?	NPO	.		
W		A		A		W		A			
D		AD		DADD		D		ADD			
D		D		D		D		ADD			
									F		

(A_07)

In		Nineteen		twenty		four		and		nineteen	
I		K	_	K	_	K	:	C	?	K	_
W		A		A		A		W		A	
D		AD		D		A		D		AD	
D		AD		D		A		D		AD	
							FR				

thirty		three,		British		climbers		nearly		reached	
K	_	K	:	J	_	NCS	:	A	;	VMF	,
A		A		A		A		A		S	
D		A		AD		AD		AD		A	
D		A	 	AD		D		D		D	
			FR								

the		top.	
D	_	NCO	.
W		A	
D		A	
D		A	
			F

(A_08)

In		all		these		attempts,		several		people	
I		J	_	J	_	NCO	:	J	_	NCS	:
W		A		A		A		A		A	
D		A		D		DA		ADD		AD	
D		A		D		DA	 	ADD		D	
							FR				

died.	
VMF	,
S	
A	
A	
	F

(A_09)

Most		of		these		expeditions		had		tried	
NIS	:	I		J	_	NCO	:	VAF	_	VMP	
S		W		A		A		W		S	
A		D		A		DDAD		D		A	
D		D		A		DDAD	 	D		A	
							FR				

to		climb		the		mountain		from		the	
I		VMI	,	D	_	NCO	:	I		D	_
W		S		W		A		W		W	
D		A		D		AD		D		D	
D		D		D		AD	 	D		D	
							FR				

north.	
NCO	.
A	
A	
A	
	F

(A_10)

Then,		In		nineteen		forty		nine		for	
A	:	I		K	_	K	_	K	:	I	
A		W		A		A		A		W	
A		D		AD		D		A		D	
A		D		AD		D		D		D	
	FR										

the		first		time,		foreigners		were		allowed	
D	_	J	_	NCO	:	NCS	:	VAF	_	VMP	
W		A		A		A		W		S	
D		A		A		A		D		DA	
D		D		A		A		D		D	
					FR						

to		enter		the		Kingdom		of		Nepal.	
I		VMI	,	D	_	NPO	;	I		NPO	.
W		S		W		A		W		A	
D		AD		D		AD		D		DA	
D		AD		D		AD		D		DA	
			LR								F

Appendix 3.1

Score Sheets of Human Evaluation

Chapter 3 described human evaluation of 23 German speakers' English intonations. Two human evaluators (Evaluator I and Evaluator II) compared their recorded utterances of the first 10 sentences of Block A in the ISLE corpus, A_01 through A_10, with a 'model' script produced by prosodic annotation in Chapter 2. Then, they judged if each utterance contained intonation error(s). Details of the human evaluation process were explained in Section 3.2 and results of the evaluation were described in Section 3.3.

Table 3.1 showed the number of sentences for each speaker in which each evaluator perceived intonation error(s). Table A3.1 and A3.2 are score sheets of Evaluator I and II, which show all 230 judgments (10 utterances from 23 speakers) of these evaluators.

Table A3.1 Score Sheet of Evaluator I

Speaker	Sentence Number										Err	Opt	Pst
	A_01	A_02	A_03	A_04	A_05	A_06	A_07	A_08	A_09	A_10			
sess0006	1=	1=	1=	1=	1	0=	1	0=	0=	0	6	6	6
sess0011	0	0=	0	0	0=	0=	0=	0=	0=	1?=	1	0	2
sess0012	1	0=	1=	0?	0=	1?	1	1=	0?=	1	6	5	8
sess0015	1=	0=	1	1	0=	0=	0=	0?=	1	1?	5	4	6
sess0020	0	0=	0	0	0=	0?=	0	0	0?=	0=	0	0	2
sess0021	0	0=	0	0	0=	0	0	0	0=	0	0	0	0
sess0161	1?=	0=	0?	0?	0	0=	0	0	0=	0=	1	0	3
sess0162	1?	0=	1=	0	0=	0?=	0=	0=	0?=	1=	3	2	5
sess0163	1	0=	1=	0	0=	1?=	1?=	1	1=	0=	6	4	6
sess0164	1=	0=	1?=	1=	1	0	0=	0=	0=	0=	4	3	4
sess0181	1	1	1?=	1=	1?=	0	0=	0	0?=	0	5	3	6
sess0182	0?	0=	1?=	1?=	1?	0=	0?=	0=	0=	0	3	0	5
sess0183	1=	0=	1=	1=	1=	1?=	1	0?	1	0	7	6	8
sess0184	0?	0=	1=	0	1=	1?=	0	0	1=	0	4	3	5
sess0185	0?	0=	0	0	0	0=	0=	0=	0=	1?	1	0	2
sess0186	1?=	0=	1?=	0?=	0	0	0=	0=	0?=	1?=	3	0	5
sess0187	1=	0=	1=	1?=	1=	0=	0=	0=	0=	0	4	3	4
sess0188	0?=	0=	1?=	0	0=	0=	0=	0=	1?=	0=	2	0	3
sess0189	1=	1?=	1?=	1=	1=	0	0?	0?	0	1=	6	4	8
sess0190	1=	0=	0?=	0	0=	0=	0=	0?	1?=	1	3	2	5
sess0191	1	0?=	1?=	0	1	0=	0=	0	0	1?	4	2	5
sess0192	1=	0=	1=	0=	0?	1?=	0=	0=	1=	0?=	4	3	6
sess0193	0	0?=	0=	0	0=	0=	0	0=	0=	0?	0	0	2
Sum	15	3	16	8	9	5	4	2	7	9	78	50	106

Table A3.2 Score Sheet of Evaluator II

Speaker	Sentence Number										Err	Opt	Pst
	A_01	A_02	A_03	A_04	A_05	A_06	A_07	A_08	A_09	A_10			
sess0006	1=	1=	1=	1=	0	0=	0	0=	0=	1	5	5	5
sess0011	1	0=	1	1	0=	0=	0=	0=	0=	1=	4	4	4
sess0012	0	0=	1=	1	0=	0	0	1?=	0=	0	3	2	3
sess0015	1=	0=	0	0	0=	0=	0=	0=	0	0?	1	1	2
sess0020	1?	0=	1	1	0=	0=	1	1	0=	0=	5	4	5
sess0021	1	0=	1	1	0=	1	1	1	0=	1	7	7	7
sess0161	1=	0=	1	1	1	0=	0	1?	0=	0=	5	4	5
sess0162	0	0=	1=	1	0=	0=	0=	0=	0=	1=	3	3	3
sess0163	0	0=	1=	1	0=	0	0	0	1=	0	3	3	3
sess0164	1=	0	1	1=	0?	1	0?=	0	0	0=	4	4	6
sess0181	0	0	1=	1=	1=	1	0=	1	0=	1	6	6	6
sess0182	1	0=	1=	1=	0	0=	0=	0=	0=	1	4	4	4
sess0183	1=	0=	1=	1=	1=	1=	0	1	0	1	7	7	7
sess0184	1	0=	1=	1	1=	1=	1	1	1=	1	9	9	9
sess0185	1	0=	1	1	1	0=	0=	0=	0=	0	4	4	4
sess0186	1=	0=	0	1=	0	1	0=	0=	0=	1=	4	4	4
sess0187	1=	0=	1=	1=	1=	0=	0=	0=	0=	1?	5	4	5
sess0188	0=	0=	1=	1	0=	0=	0=	0=	0	0=	2	2	2
sess0189	1=	1=	1=	1=	1=	1	1	1	1	1=	10	10	10
sess0190	1=	0=	0=	1	0=	0=	0=	1	1=	1?=	5	4	5
sess0191	0	0=	1=	1	0	0=	0=	1	1	0	4	4	4
sess0192	1=	0=	1=	0?=	1	1=	0=	0=	1=	0=	5	5	6
sess0193	1	0=	0=	1	0?=	0=	1	0=	0=	1	4	4	5
Sum	17	2	19	21	8	8	5	10	6	13	109	104	114

- ❖ 0: marked for a correct intonation sentence.
- ❖ 1: marked for an intonation error sentence.
- ❖ ?: marked for an indefinite judgment.
- ❖ =: marked for agreement of two evaluators regardless of indefinite judgment(s).
- ❖ Err: Number of sentences, in which the evaluator judged intonation error(s).
- ❖ Opt = Err – (number of indefinite error sentences)
- ❖ Pst = Err + (number of indefinite correct sentences)

Appendix 3.2

Pronunciation Abilities of 23 German Spoken Learners' English

This appendix shows pronunciation ability of 23 German speakers. In the ISLE project, human annotators marked up these speakers' pronunciation errors at a word-level. Pronunciation mark-up added to Block D through Block G of the corpus with a total of 1083 words, although it was not done to Section A of the corpus, which was used for judging intonation abilities. Block D through Block G was designed as exercise sessions for EFL learners. We counted the number of the marked-up errors in these blocks. There were following two reasons that pronunciation abilities of the speakers were required.

Before the speakers were categorized into 'good' and 'poor' intonation groups in Section 3.5, some speakers with relatively 'poor' pronunciation speakers were excluded from the groupings in order to minimize influences of pronunciation factors in the HTK speech recognition experiments, in which prosody of German spoken learners' English should be analysed. Therefore, we referred to their pronunciation data and pinpointed speakers, who made specifically large number of pronunciation errors compared with other speakers.

Pronunciation abilities were also considered in groupings and sub-groupings. When the other conditions of groupings and sub-groupings could not decide the order of multiple speakers, pronunciation abilities became decisive factors.

Table A3.3 shows the number of pronunciation errors 23 German speakers. According to the table, three speakers, SESS0012, SESS0163 and SESS0191, whose

numbers of errors are in bold, had pronunciation errors at least twice as many as the rest of speakers.

Table A3.3 Number of Pronunciation Errors

Speaker	Errors	Speaker	Errors	Speaker	Errors
SESS0006	15	SESS0163	43	SESS0187	4
SESS0011	0	SESS0164	6	SESS0188	2
SESS0012	82	SESS0181	4	SESS0189	10
SESS0015	5	SESS0182	9	SESS0190	5
SESS0020	5	SESS0183	17	SESS0191	35
SESS0021	13	SESS0184	8	SESS0192	3
SESS0161	5	SESS0185	4	SESS0193	1
SESS0162	8	SESS0186	13	Average	12.91

Appendix 3.3

Listing of German Speakers in ‘Good’ English Intonation Order

This appendix explains how 20 German speakers, in which 3 relatively poor pronunciation speakers had been already excluded, were listed in ‘good’ intonation order. The speakers were listed separately for Grouping I and II based on Evaluator I and II, respectively, following rules described in Sub-section 3.5.3.

3.3.1 Grouping I

This is a list of the speakers listed in ‘good’ pronunciation order with reasons of the ordering based on Evaluator I.

- 1) SESS0021 (0-0-0, 13) [Good A]
 Number of pessimistic intonation errors
- 2) SESS0193 (0-0-2, 1) [Good B]
 Number of pronunciation errors
- 3) SESS0020 (0-0-2, 5) [Good C]
 Number of intonation errors
- 4) SESS0011 (1-0-2, 0) [Good D]
 Number of pronunciation errors
- 5) SESS0185 (1-0-2, 4) [Good D]
 Number of pessimistic intonation errors
- 6) SESS0161 (1-0-3, 5) [Good C]
 Number of intonation errors
- 7) SESS0188 (2-0-3, 2) [Good B]
 Number of intonation errors
- 8) SESS0186 (3-0-5, 13) [Good A]
 Balance of pronunciation abilities in ‘good’ and ‘poor’ intonation groups
 - SESS0186 and SESS0162 should be in ‘good’ and ‘intermediate’ intonation groups, respectively, as the former group tended to have fewer

pronunciation errors than a ‘poor’ intonation group, and as SESS0186 had more pronunciation errors than SESS0162.

9) SESS0162 (3-0-5, 8) [Intermediate]

Number of pronunciation errors

10) SESS0182 (3-0-5, 9) [Intermediate]

Number of intonation errors

11) SESS0187 (4-3-4, 4) [Intermediate]

Number of pronunciation errors

12) SESS0164 (4-3-4, 6) [Intermediate]

Number of pessimistic intonation errors

13) SESS0184 (4-3-5, 8) [Poor A]

Number of pessimistic intonation errors

14) SESS0192 (4-3-6, 3) [Poor B]

Number of pronunciation errors

15) SESS0190 (4-3-6, 5) [Poor C]

Number of intonation errors

16) SESS0181 (5-3-6, 4) [Poor D]

Number of optimistic intonation errors

17) SESS0015 (5-4-6, 5) [Poor D]

Number of intonation errors

18) SESS0189 (6-4-8, 10) [Poor C]

Balance of pronunciation abilities in sub-grouping Poor B and C

- SESS0189 and SESS0006 should be No18 and No19, respectively, to form sub-groups with SESS0190 and SESS0192, respectively, as the former two speakers in ‘good’ pronunciation order, and the latter two speakers in ‘poor’ pronunciation order.

19) SESS0006 (6-6-6, 15) [Poor B]

Number of intonation errors

20) SESS0183 (7-6-8, 17) [Poor A]

❖ Speaker’s data are shown as

(intonation errors-optimistic intonation errors- pessimistic intonation errors)
[group].

3.3.2 Grouping II

This is a list of the speakers listed in ‘good’ pronunciation order with reasons of the ordering based on Evaluator II.

- 1) SESS0015 (1-1-2, 5) [Good A]
Number of intonation errors
- 2) SESS0188 (2-2-2, 2) [Good B]
Number of intonation errors
- 3) SESS0162 (3-3-3, 8) [Good C]
Number of intonation errors
- 4) SESS0011 (4-4-4, 0) [Good D]
Number of pronunciation errors
- 5) SESS0186 (4-4-4, 13) [Good D]
Balance of pronunciation abilities in sub-grouping Good B, C, and D
 - SESS0186, SESS0182 and SESS0185 should be No5, No7 and No6, respectively, to form sub-groups with SESS0011, SESS0188 and SESS0162, respectively, as the former three speakers in ‘poor’ pronunciation order, and the latter three speakers in ‘good’ pronunciation order.
- 6) SESS0185 (4-4-4, 4) [Good C]
Balance of pronunciation abilities in sub-grouping Good B, C, and D
 - Same as above.
- 7) SESS0182 (4-4-4, 9) [Good B]
Number of pessimistic intonation errors
- 8) SESS0193 (4-4-5, 1) [Good A]
Number of pessimistic intonation errors
- 9) SESS0164 (4-4-6, 6) [Intermediate]
Number of intonation errors
- 10) SESS0187 (5-4-5, 4) [Intermediate]
Number of pronunciation errors
- 11) SESS0020 (5-4-5, 5) [Intermediate]
No special order

12) SESS0161 (5-4-5, 5) [Intermediate]

Number of optimistic intonation errors

13) SESS0006 (5-5-5, 15) [Poor A]

Number of pessimistic intonation errors

14) SESS0190 (5-5-6, 5) [Poor B]

Balance of pronunciation abilities in sub-grouping Poor B and C

- SESS0190 and SESS0192 should be No14 and No15, respectively, to form sub-groups with SESS0184 and SESS0183, respectively, as the former two speakers in 'poor' pronunciation order, and the latter two speakers in 'good' pronunciation order.

15) SESS0192 (5-5-6, 3) [Poor C]

Number of intonation errors

16) SESS0181 (6-6-6, 4) [Poor D]

Number of intonation errors

17) SESS0021 (7-7-7, 13) [Poor D]

Balance of Balance of pronunciation abilities in sub-grouping Poor C and D

- SESS0021 and SESS0183 should be No17 and No18, respectively, to form sub-groups with SESS0181 and SESS0192, respectively, as the former two speakers in 'good' pronunciation order, and the latter two speakers in 'poor' pronunciation order.

18) SESS0183 (7-7-7, 17) [Poor C]

Number of intonation errors

19) SESS0184 (9-9-9, 8) [Poor B]

Number of intonation errors

20) SESS0189 (10-10-10, 10) [Poor A]

❖ Speaker's data are shown as:

(intonation errors-optimistic intonation errors- pessimistic intonation errors)

[group]