SPOKEN LANGUAGE PROCESSING APPLIED TO NONNATIVE LANGUAGE PRONUNCIATION LEARNING

(音声言語情報処理を利用した非母語発音学習)

GOH KAWAI

河合　剛

SPOKEN LANGUAGE PROCESSING APPLIED TO NONNATIVE LANGUAGE PRONUNCIATION LEARNING

Goh Kawai

A dissertation submitted for the degree of Doctor of Philosophy

University of Tokyo

Department of Information and Communication Engineering

March 1999

音声言語情報処理を利用した非母語発音学習

河合　剛

博士（工学）学位論文

東京大学大学院　工学系研究科　電子情報工学専攻

１９９９年３月

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CALL	Computer-aided language learning
HMM	Hidden markov model
HTK	Hidden Markov Model Toolkit (a commercial software product for developing speech recognizers)
L1	Learner's first, native language
L2	Learner's second, target language
TJSL	Teaching Japanese as a second language

FOREWORD

This dissertation is coauthored. Coauthored dissertations are rare, which is why I did not list my wife's name on the cover. All the same, Noriko reads every word of all papers I write, points out flaws, and suggests improvement strategies. This dissertation is our joint work.

This dissertation is unfinished. All research papers are reports of work in progress — this is one of the lessons I learned during the past three years as a student. I am not apologizing for all my shortcomings; I offer my work to your perusal so that you may discover what I have learned and inform me of my mistakes.

This dissertation is fortunate. I am grateful to have worked with Mr. Noboru Takahashi (laboratory technician), Mr. Osamu Tsuchida (B.S. class of 1998), Miss Shirley Chingsiu Lim (B.S. class of 1999), and Mr. Carlos Toshinori Ishi (Ph.D. expected 2001). The collaborative research projects with Yamanashi University's Dr. Akira Ishida (lecturer), Mr. Tomokazu Amemiya (B.S. class of 1999), and Mr. Tomohiro Yamamoto (B.S. class of 1999) were fruitful. The insights I received from the rest of my laboratory colleagues have deeply affected me. I am indebted to my brother Kan for document processing.

This dissertation is a miracle. When I applied for graduate school three years ago, I knew I needed three miracles — to be accepted, to graduate, and to find a job. Professor Keikichi Hirose, my academic advisor, was pivotal in making these miracles happen. I shall endeavor to repay his faith in me through my future publications.

In the summer of 1986, I dedicated my master's thesis to my father. In the winter of 1999, I dedicate my doctoral dissertation to my mother.

St. Valentine's Day, 1999, Hongo campus

Goh Kawai

FOREWORD IN JAPANESE

巻頭言

この論文は共著である。著者が複数の学位論文は稀有だから表紙に氏名を記さなかったけれども、紀子がこの論文をすみずみまで読み、欠点を批判し、改良方法を提案した。この論文は妻との合作である。

この論文は未完である。論文は研究の進捗報告なり。いままでの三年間で学んだ教訓のひとつだ。未熟な成果で茶を濁すのではない。識者に自分の仕事ぶりを見せ、誤りを早期に見つけだすとともに、同志を触発せんことをこいねがう。

この論文は幸福である。畏兄高橋登技官を筆頭に、わが研究室の土田統君(1998年学部卒業)、林増秀嬢(1999年学部卒業)、石井寿憲カルロス君(2001年博士卒業予定)とともに仕事ができて好運だった。山梨大学工学部の石田朗講師、雨宮友和君(1999年学部卒業)、山本智宏君(1999年学部卒業)との共同研究も多実だった。他の同僚諸兄姉から受けた啓発も終生の宝となろう。文書処理は賢弟寛に負う。

この論文は奇跡である。入学できれば第一の奇跡、卒業できれば第二の奇跡、就職できれば第三の奇跡、こう考えての進学であった。三奇の要軸の広瀬啓吉教授にいかにして報恩すべきか。ひとつでも多く論文を著そう。

1986年の夏、修士論文を父に捧げた。1999年の冬、博士論文を母に捧げる。

1999年2月14日 本郷にて

河合　剛

ACKNOWLEDGMENTS

I gratefully acknowledge Dr. Kazuya Takeda (Nagoya University) and his colleagues for allowing me the use of Japanese HMMs [28]. Many thanks to Dr. Steve Young (Cambridge University) and his team for allowing me to use their American English HMMs [36]. This dissertation would have been impossible without the use of these two HMM sets. I thank the two gentlemen and their colleagues for their generosity.

1. ABSTRACT

1.1. EXECUTIVE SUMMARY IN ENGLISH

The problem addressed is assisting language teachers in improving their learners' acquisition and retention of pronunciation skills. This is an important issue because spoken language is the most commonly used form of communication. Self-study methods are required because there are not enough teachers to teach students individually.

Conventional self-study systems forced the learner to evaluate his pronunciation himself — an impossible task given that if the learner could judge pronunciation quality he would have no need to learn it to begin with.

The optimal solution is a computer-aided language learning system that train the pronunciation of non-native learners. Such a system must explain to the learner (1) what his mistake was, (2) the severity of the mistake, and (3) how to correct his mistake.

This dissertation proposes methods for teaching pronunciation in three areas — phone duration, pitch, and phone quality — that provide beginning learners with corrective feedback for essential pronunciation skills. Knowledge of speech recognition technology, phonological theory, and foreign language pedagogy are combined to model and measure non-native behavior. Focusing on measurable, identifiable pronunciation skills achieves high reliability and validity. The proposed methods can be extended to teach any language; this dissertation implements examples of systems for learning Japanese or American English.

The most significant direct contribution of this dissertation is in implementing the first computer-aided pronunciation learning system that provides feedback similar to human teachers. For phone duration and pitch accent in particular, pronunciation intelligibility is estimated by coupling speech recognition technology with quantitative knowledge of how native speakers perceive pronunciation differences.

Methods used in this dissertation may aid related research in spoken language processing by (1) improving speech recognition performance by using duration-based phone models for languages where phone durations are phonemic or lengthening indicates stress placement, (2) improving performance of language recognition, dialect detection and speaker adaptation by using non-native or nonstandard-dialect phone models, and (3) improving spoken language understanding performance by segmenting utterances into intonational phrases by using pitch models of lexical pitch accent and intonation patterns of fixed phrases.

1.2. EXTENDED SUMMARY IN JAPANESE

1.2.1. 概説

成人が母語でない言語の発音を学習する際、専門の教師と1対1で指導を受けるのが理想だが、授業時間の制約や教育費用の高騰などの理由から学習時間が不足しがちである。発音教育の指導効率を上げ、就学期間の短縮や学習量の増加を図り、発音能力の習得と保持を高めるために、習得が困難な発音技能を教師の手を借りずに教室外で単独で学べるようにしたい。学習者が自分の発音の適否を自分で判断できるくらいなら発音を学ぶ必要がないから、発声の練習には矯正フィードバックが不可欠である。自習するからには発音能力の測定ならびに評価の自動化が必須の条件である。

本研究は、教師が学習者に発音学習の概要を教室で指導したうえで、機械的な反復練習をコンピュータを用いて行ない、発音誤りを自動的に判別して矯正フィードバックを与えるシステムを提案した。言語音の主だった特徴である音長(音の持続時間)、韻律(音のピッチ)、音色(音のスペクトル)を測定する手法を音声言語情報処理技術を利用して開発した。一部の音声特徴は単に正誤判定をするだけでなく、母語話者の何パーセントに通じるかを知覚実験に基づいて推定し採点の尺度としたので、学習者の発音能力の向上が理解しやすいのが特長である。

システムの有効性を評価するために、音長の測定技術を「日本語の特殊拍の学習システム」として実装した。これは日本語非母語話者が日本語の拍感覚を身につけ、長音と短音を区別して発音できるように訓練するシステムである。同様に韻律は「日本語のピッチアクセントの学習システム」として実装した。これは日本語非母語話者が日本語のアクセント感覚を身につけ、単語や語句の高低アクセントを正しく発音できるように訓練するシステムである。そして音色は「英語の母音挿入の矯正システム」(日本語母語話者が英語を話すときに挿入してしまう母音を検出し、修正を促すシステム)、「英語の音の置換・挿入・脱落を検出するシステム」(日本語母語話者が英語を話すときに置換・挿入・脱落される音を検出するシステム)、「日本語の英語なまり測定システム」(英語母語話者が日本語を話すときの英語なまりの度合を測定するシステム)、として実装した。これらの実装例を通じて要素技術の可搬性や言語独立性を示すとともに、発音学習の自動化の有効性を明らかにした。

以下、本研究の目的、採用したアプローチの基本姿勢、音長・韻律・音色を測定する手法ならびに評価実験の結果について、要点を順に述べる。

1.2.2. 研究の目的

本研究の原動力になっている疑問は「どうやったら成人が外国語を話せるようになるか」である。(いま用いた「外国語」という表現は適切でない。正しくは「母語でない言語」と表現せねばならない。母語でない言語は必ずしも外国の言語と限らないからである。し

かしここでは正確だけれでも不自然な言い回しよりも、不正確ながらもわかりやすい表現を採用する。)

幼児はいかなる言語でもやすやすと身につけるのに、大人は驚異的な努力をもってせねば新しい言語を覚え切れない。学習の到達が低いと、母語で簡単に言い表せる内容を伝えるのにつたない表現しか選べず、会話も遅くもどかしい。語学学習は繰り返しが多く、素材も幼稚で、学習者と教師の両者にとって単調である。この語学学習の苦痛を和らげる方法を探るのが、本研究の動機である。

言語技能の中でもとりわけ発音は母語話者に近いレベルまで学ぶのがが難しい。母語話者なみの発音は往々にして成人非母語話者の手が届かない夢である。専門の教師の個人指導を受けられれば良いが、現実的には教師との時間が限られているので、教師の手をなるべく借りずに学びたい。教師も退屈な作業から解放されて、高度な言語活動の指導などの高付加価値の教育作業に従事できる。ところがテープレコーダや音声分析装置を用いた在来の自習方式は、学習者が自分の発音の適否を自分で判断せねばならないとか、発音誤りの修正方法がわからないとか、聞き取りの訓練に終止して発音をまるで練習しないといった本質的な欠点をはらんでいる。

発音の練習には権威ある矯正フィードバックが不可欠である。このためには次の3条件がそろわなければならない。

(1) 何を間違えたのかを明らかにする。(発音の間違いの特定)

(2) どのくらい通じるのかを明らかにする。(発音の了解度の測定)

(3) どう直せばいいのかを明らかにする。(発音の矯正方法の提示)

いままでの発音自習方式は、これらの3条件をまったく満たさなかった。たとえば、発話全体の発音の良否を測定する手法は提唱されているが、発話中の発音誤りを検出できない(したがって発音矯正の方法も指示できない)。発音の良否の尺度も、いわゆる「外国語なまり」の度合を主観的な順序尺度で表しただけで、発話がどのくらい通じるかを予測するものではなかった。発音自習方式の本質をさぐるために上記3条件を満たす発音自習方式の理論を明らかにし、かつ実装例を通じてシステム開発戦略の有効性を証明するのが本研究の狙いである。

1.2.3. 研究の理論的背景

発音能力の測定ならびに評価の自動化の理論的背景の基本は次の2点に要約できる。

(1) コンピュータによる音声言語情報処理技術を利用した学習者音声の分析。

(2) 人間の音声言語に対する知覚を調べるための心理学実験の結果にもとづく発音の了解度の測定。

上記の(1)にとって重要なのは、発話に含まれる話者性(その話者の生理的条件などに起因する、言語に依存しない話者の個人的側面)と言語性(その話者が話す言語が、その言語を話す者すべてに共通して与える影響)とを分離する技法である。言語性を無視すれば発音教育が成り立たないのは明らかである。話者性を保ったままの学習は、発音品位の評価に声の美醜といった発音能力と無関係な要因が関与する恐れがあり、また、学習者と同一の話者性をもつ模範話者を見つけない限り、模範話者の話者性を含んで学習するので声帯模写の訓練に陥る恐れがある。したがって発音を学習するためには話者性を除去し、言語性のみに注目しなければならない。本研究では、言語性を保ったまま話者性を取り除く方法として、不特定話者音声認識技術を採用した。言語性だけを取り出せば、外国語教授法の分野でいままでに得られた知見を用いて発音を矯正できる。

上記の(2)にとって重要なのは、人間の音声知覚に与える要因の特定である。たとえばピッチの知覚は、単に基本周波数の高低で決まるのではなく、音の長さやパワーの影響も受ける。これらの要因の相互関係を知るために、段階的に調節した音声を音声合成技術を用いて作り、合成音を母語話者に聞かせる心理学実験を行なった。測定対象の変数だけを人工的に調節して複数の母語話者に適否を判断させれば、学習者がもしそのような音声を発音した場合、母語話者がどのように発音を判断するかを予測できる。この予測値を本研究は「了解度」と呼んで発音品位の尺度としている。了解度は、言語を実際に運用する場面において発音がどの程度通じるかを表す尺度であるから、学習者の発音能力がどの程度高まったのかが理解しやすい。

上記の(1)(2)からわかるように、本研究が採用したアプローチの基軸は、学習者の発音から言語性だけを抽出し、発音誤りを検出し、了解度と修正方法を学習者に伝えるという手順である。以下、日本語と英語を指導するための具体的な実装例を通して、発音自習のための理論がシステム開発にどのように適用されるのかを説明する。

1.2.4. 音長の測定と評価

音長(言語音の持続時間)は音声の基本的特徴量のひとつである。音長を用いた発音学習の可能性を示すために、音長が音韻的に意味を持つ言語における音長の発音学習のありかたを検討した。

日本語では長音と短音の音韻対立があり、日本語教育の分野ではまとめて「特殊拍」と呼ばれ、日本語を外国語として学ぶ人の学習上の障害となっている。そこで音声合成器を用いてさまざまな音長を合成し、日本語母語話者に聞かせて長短のいずれに聞こえるかを判断させた。つづいて音声認識器を用いて入力音声の音長を測定し、測定された音長から音長の適否を判断するアルゴリズムを開発した。このアルゴリズムを対話型学習システムとして実装し、学習効果測定実験を行なった。実装にまつわる一連の作業なら

びに学習効果測定の結果から、音長を用いた発音学習の自動化が可能であり、かつ語学学習の観点からも意義が高いことがわかった。

1.2.5. 韻律の測定と評価

韻律(言語音のピッチの上がり下がり)も音声の基本的特徴量のひとつである。韻律を用いた発音学習の可能性を示すために、韻律が音韻的に意味を持つ言語における韻律の発音学習のありかたを検討した。

日本語では拍によって音の高低があり、日本語学でピッチアクセントと呼ばれる単語の韻律パターンを持っている。「雨」と「飴」、「箸」と「橋」の対立がこの例である。音長を指導するシステムと基本的に同様の方法で知覚実験を行なった。音声合成器を用いてさまざまなピッチを合成し、日本語母語話者に聞かせて高低のいずれに聞こえるかを判断させた。つづいて音声認識器を用いて入力音声のピッチを測定し、測定されたピッチからピッチの適否を判断するアルゴリズムを開発した。このアルゴリズムを対話型学習システムに実装し、学習効果測定実験を行なった。韻律を用いた発音学習の自動化が可能であり、かつ語学学習の観点からも意義が高いことがわかった。

1.2.6. 音色の測定と評価

音色(言語音のスペクトル情報)も音声の基本的特徴量のひとつである。音色が意味を持たない言語は存在しない。スペクトルには多くの情報が含まれるので、話者性を除去し、言語性だけを抽出するためのアルゴリズムが必要である。外国語学習者は学習者の母語と学習対象の言語の両方の発音を混ぜて発話すると考えられる。このような場合、音声信号からだけでは調音器官の動きを特定できない。本研究では音の弁別的認識を用いた発音学習のありかたを検討した。

はじめに学習者の母語と学習対象の言語のそれぞれの母語話者音声から学習させた音響モデルを用いた単音音声認識器を用意した。語学教師から得た知見をもとに、目的の正しい音がどのような音に代替されて発音されるのか(音の置換誤り)、あるいは全く発音されないのか(音の脱落誤り)、あるいは余計な音が発音されるのか(音の挿入誤り)を調べ、これらの現象を外国語学習者の発音に見られる音韻規則として記述した。学習者の音声がシステムに入力されると、システムはどこでどの音が生じたかを認識する。音の種類ごとに認識するので話者性が相殺される。発音学習の具体例として、「英語の母音挿入の矯正システム」(日本語母語話者が英語を話すときに挿入してしまう母音を検出し、修正を促すシステム)、「英語の音の置換・挿入・脱落を検出するシステム」(日本語母語話者が英語を話すときに置換・挿入・脱落される音を検出するシステム)、「日本語の英語なまり測定システム」(英語母語話者が日本語を話すときの英語なまりの度合を測定するシステム)を実装した。音色を用いた発音学習の自動化が可能であり、語学学習の観点からも意義が高く、言語に依存しない高い汎用性の方式であることがわかった。

音色の了解度を測る知覚実験は実施が困難であり、おそらく原理的に不可能ではないかと思われる。しかし「外国語なまり」の度合いを示す尺度として「単音の発音誤りの種類の数」に基づく定量化を試みた。英語母語話者が日本語の母音を発音するときのように、多数の音をもつ言語の話者が少数の音を持つ言語を発音する場合は、提案した尺度が外国語なまりの指標として役立つ。

1.2.7. まとめ

本研究は発音能力の自動評価の理論的背景を明らかにし、実装例を通じて理論の有効性を示した。本研究が提唱する技法は発音教育に限らず、言語認識(母語認識や方言認識を含む)や話者適応などの音声認識の頑健さを高める効果も期待できる。また、いままでは音声認識技術の基本性能に含まれなかった音長(日本語の特殊拍や、英語の単語強勢など)や韻律(日本語のピッチアクセントや、各国語のイントネーションパターンなど)の認識にも応用できる。知覚実験に基づく音の範疇型判断を音声認識アルゴリズムに取り込めば、データ駆動型の統計処理一辺倒の研究現況に好影響をおよぼしうる。

2. THE PROBLEM

The problem addressed in this paper is automatically detecting, measuring and correcting non-native pronunciation characteristics (so-called "foreign accents") in foreign language speech. Acquiring nativelike pronunciation ranks first in desirability among foreign language learners. Unfortunately most adult learners develop fossilized pronunciations that are distinctly non-native. Research suggests that closer-to-native pronunciation might be obtained through intense pronunciation training early in the learner's study program [2].

In particular, the past decade has seen an influx of non-native speakers entering Japan. Jobs, education, childcare, shopping — the language barrier hinders communication in all aspects of the newcomers daily lives. Learning to speak Japanese fluently behooves good hearing and listening skills — skills not effectively learned by studying written language. Nonnatives (especially Asians) speaking with an accent are occasionally branded as inferior undesirables [30]. Good pronunciation skills are essential for succeeding in Japan [18].

Yet only token attention has been paid to pronunciation teaching. The time spent in classrooms practicing pronunciation is minimal. So is the quality of teaching material. According to a survey conducted on 158 TJSL (teaching Japanese as a second language) teachers in the nation, most teachers teach pronunciation in some form or another, but limit classroom activities to less than 10 hours total, typically concentrated at the very beginning of entry-level courses [29]. The lack of instruction time is mainly due to the shortage of hours both learners and teachers can afford. Learners are eager to learn useful Japanese as quickly as possible — an understandable desire given they are already in the country. Teachers are under pressure to prepare learners for Japanese language proficiency tests [13]. These tests primarily measure written skills and listening comprehension. Oral production is not tested because measuring speaking skills is unreliable without absurdly intense effort [34]. A common misconception that reading and writing Japanese is hard but speaking it is easy does not help. As a consequence, classroom instruction has concentrated on orthography, vocabulary and syntax. It is not an exaggeration to state that students' abilities are measured by how many kanji they can read and write. Students are lucky if their teacher uses a tape recorder or hiragana chart for pronunciation practice. Most teachers do not use textbooks or teaching aids at all [29]. TJSL practitioners both here and abroad are clamoring for a systematic syllabus for Japanese pronunciation training [21].

It is ironic that, after months of hard work and a well-deserved Japanese language proficiency certificate in his hand, a non-native speaker seeking a job would be turned down at his first interview. His poor oral skills suggest he knows less Japanese than he really does. Even if he is hired, his native speaker colleagues may not let him into the communication loop — "He won't understand," they may decide among themselves. Cast away from the social fabric, the non-native speaker is doomed. The final insult is that by this time it is too late for remedies. His pronunciation mistakes have become solidified, incapable of change.

This perhaps overly melodramatic tragedy can be avoided by using self-study methods for pronunciation skills. The acute shortage of classroom time and the appalling lack of teaching material can be rectified while simultaneously strengthening the acquisition and retention of correct pronunciation.

Implementing such a self-study system is challenging in several ways. First, we need to target pronunciation mistakes that either occur frequently and/or cause communication confusion. Second, we need to measure these mistakes in consistent, meaningful terms. Third, we need to unambiguously instruct the learner how to correct these mistakes. These three conditions are necessary for the learning system's reliability and validity. As detailed pronunciation curricula for TJSL do not exist, one must be built.

3. THE LITERATURE

3.1. OVERVIEW

This chapter reviews systems for spoken language self-study. We start with a classic example of comparing native and non-native speech by ear. We then discuss systems that process the learner's speech in some way or another to produce information other than audio playback of the learner's utterance. There are two examples of systems that display the speech waveform, pitch track, and/or formants of both the native's model and the non-native's rendition. These systems use graphic information to assist the learner in identifying pronunciation mistakes. Next, we review two examples that use speech recognition technology to understand the learner's speech. These systems engage the learner in a simulated dialogue, and force the learner to speak. Last, we review an example that uses speech recognition technology to measure the quality of the learner's pronunciation. This system quantifies pronunciation quality at the phoneme level using metrics derived from acoustic distances.

All systems reviewed except the last are for TJSL, and are representative of the state of the art including systems for the training of non-Japanese languages.

3.2. COMPARISON BY EAR

The classical pronunciation self-study method is using a tape recorder [31]. The learner listens to native speech, repeats and records it, and compares his utterance with the native's. Advantages of this method are that the learner can practice listening skills and that the equipment can be portable. Disadvantages are that the learner is not forced to say anything, and that the learner receives no corrective feedback from the system. The learner must be highly motivated and have a good ear for foreign languages.

Learners rarely have these qualifications. Indeed, a common shortcoming of all existing self-study methods including the tape recorder method is that none tell learners whether their speech is intelligible or what they can do to improve their pronunciation. If learners could judge the appropriateness of their renditions by themselves, they would have no need to learn pronunciation in the first place. Learners need to know what their mistakes are, how serious their mistakes are, and how to correct their mistakes. Self-study systems that impose learners to judge the accuracy of their productions are fundamentally flawed.

3.3. COMPARISON ASSISTED BY SPEECH PROCESSING (1)

Imaizumi et al proposed a system that displays the speech waveform, pitch track, formant trajectories, and other features derived from the learner's utterance [12]. These features are graphically displayed on a computer screen along with previously processed and stored images of the native's model. The learner changes his articulation so that his features match that the native model's. There is no specific instruction on how this matching might be accomplished.

Imaizumi's system suggests various possibilities to automated pronunciation training, but the effectiveness of the system is unclear. This is because the system was not intended primarily as a pronunciation teaching aid. It was proposed as an extension to Imaizumi's existing speech analysis program for F0, F1 and F2 feature extraction; as such, Imaizumi did not run educational experiments. The lack of instruction on how to alter the learner's speech with the native's leads one to imagine that the system may not be effective.

Another concern is that the native model is a single utterance of a particular individual. If the learner's speech waveform, formant trajectories and so forth were to overlap completely with the native model's, then the learner must sound exactly like that particular native speaker, mirroring the native's idiosyncratic speech mannerisms and voice quality. Imaizumi's system risks turning into a method for voice actor training unless a native speaker is chosen carefully, perhaps by using multiple native speakers over the course of study, or by pairing natives with non-natives having similar physical characteristics.

3.4. COMPARISON ASSISTED BY SPEECH PROCESSING (2)

While Imaizumi's system is an application of his speech analysis tool, Saida et al designed a system specifically for TJSL [24][25]. Saida's system shares many features with Imaizumi's; both record, play back and process the learner's speech, both display the time waveform and pitch track, and both expect the learner to identify mistakes on his own without teaching him how. Saida intentionally left out this last step because she believes explicit corrective feedback embarrasses the learner. She prefers to leave learners alone, saying learners should actively seek out pronunciation errors, fix them, and quit practice when the learner is satisfied with his skills.

Saida's approach has several serious problems. First, a learner is unlikely to be able to point out causes of pronunciation errors from a multitude of factors that comprise what we might call the intelligibility or clarity of pronunciation. Second, a learner may terminate pronunciation practice for the wrong reasons; he may be tired, or feel he is not making progress. Third, even if discovery

learning as Saida suggests were feasible, repeating trial and error sessions with native speakers seems excessively arduous for people who are in a rush to learn.

Imaizumi and Saida's systems can be expanded to include more speech processing information. But the learning system should not demand learners to acquire new skills unrelated to pronunciation practice. For instance, displaying spectrograms should be avoided, because learning to read spectrograms is unessential for learning pronunciation. (Many TJSL learners lack technical backgrounds.) The learning system should concentrate on analyzing pronunciation errors and suggesting remedies. The learner should follow the system's suggestions. In this regard, even relatively simple processing of speech (such as time waveforms) may be unsuitable.

3.5. INTERACTIVE DIALOGUE (1)

Syracuse Language Systems combined an interactive TJSL application with a Japanese speech recognition component developed by Dragon Systems [27]. This commercial product allows some user-system interactions to be completely oral, thereby forcing the learner to speak. Speech recognition accuracy is insufficient for native speakers — the system often refuses to recognize native speech — but this does not seem to bother non-native learners, who eagerly spend hours talking to the system.

As the system is a commercial product, details of the speech recognizer are proprietary. What is known from using the system is that the learner is asked to choose and read a sentence from a set of two to four; the recognizer selects the sentence having the closest acoustic signal to the learner's utterance, and the conversation branches off into different directions according to what was recognized. The recognizer handles rejection.

The system succeeds in engaging the learner in a purposeful, oral conversation. The learner needs to understand the conversation's context, find the correct answer, and say it more or less correctly. The problem is, as far as pronunciation learning is concerned, that there is no incentive to speak more clearly or intelligibly. The system does not grade pronunciation. Insofar as the learner reads one of the offered sentences aloud, the system will accept his pronunciation. The acceptance threshold seems to have little if anything to do with the intelligibility of the utterance as perceived by native speakers. No provision exists for training pronunciation per se. Similar problems exist in other systems (e.g., [1], [4], [8]). The current state of the art of speech recognition applications to spoken language learning is improving but still imperfect [3].

3.6. INTERACTIVE DIALOGUE (2)

The system designed by Ehsani et al is similar to the one by Syracuse Language Systems [4]. Both systems simulate situation-based, system-learner dialogues. Learners reply to the system's verbal prompts by saying whole sentences. Ehsani's dialogues has exactly one correct path the student

should follow. Errors elicit corrective feedback for content; for instance, when the student says "Pleased to meet you," the system might suggest "This isn't the first time you meet." Pronunciation errors are not corrected.

Ehsani's system is essentially a sentence recognizer. The instructor scripts a dialogue consisting of system prompts and student responses. The instructor anticipates correct and incorrect student utterances based on the instructor's knowledge of the learner's linguistic skills. The student does not choose from a set of sentences shown on the computer screen; he is free to respond to the system's verbal prompts in any way. The strong advantage of this approach is that the learner must actively generate a correct response instead of passively choosing one from a system-provided list. The equally strong disadvantage is that if the dialogue is not scripted correctly, the student can say something unforeseen by the system, causing the interaction to fail. The system's lack of responsiveness was cause for concern during field trials. In general, giving feedback in realtime is important in pronunciation training because the learner must receive corrections immediately after speaking or he will forget how he articulated.

3.7. PRONUNCIATION EVALUATION

Witt's system was designed for British English but studying it is useful because the basic algorithm is language-independent [33]. Witt defines the goodness for each phone in the utterance as the posterior probabilities that the learner uttered phone p given the acoustics $\boldsymbol{O}$ and the set of all phones $\boldsymbol{Q}$. She assumes that (1) all phones are equally likely, and (2) the total likelihood of all phones in $\boldsymbol{Q}$ yielding the acoustics $\boldsymbol{O}$ can be approximated by the maximum likelihood of any single phone yielding $\boldsymbol{O}$. (Strictly speaking, neither of these assumptions is true in any language, but is chosen as a first approximation.) Witt's goodness metric can quantify the pronunciation quality of each phone individually or as a group; for instance, a particular token of q can be analyzed on its own, or all phones of type q occurring in the utterance can be treated at once.

Witt's method is one of the most systematic approaches to automated pronunciation grading to date. Its weakness is relying on a single probability likelihood measure, which in turn is based on multidimensional acoustic distance measures. Analyzing the acoustic signal of a phone or an entire utterance, either by calculating the distance in acoustic feature space or by evaluating the probability likelihood of non-native speech being produced given native HMMs (hidden Markov models), is incorrect because the learner's speech is a mixture of characteristics specific to the individual (such as voice quality) and the learner's native language (namely its phonology) [22].

To improve the speech recognizer's performance, Witt uses speaker adaptation [16]. This method is risky because it does not necessarily guarantee separation of speaker-specific characteristics from that of the training-data population. In the case of well-trained native speaker HMMs, a particular native speaker (to whom we wish to adapt HMMs) differs only on speaker-specific characteristics; the rest are language-common features shared with the HMMs. This is not true for non-native

speakers, again because their speech is heavily influenced by their native language's phonology as well as their particular personal features [9][23]. An improvement is to use both native and target phone models to quantify the proximity of the learner's speech to his native and target languages.

3.8. DISCUSSION

How do adults learn to speak non-native languages? Are all adults capable of learning second languages, provided appropriate learning techniques are used over sufficient amounts of time? If adults can learn to speak non-native languages at levels of competence at or close to native speakers, then how can we automate the learning procedure? In particular, how can we automate pronunciation learning? How can we use spoken language understanding technology to measure the goodness of pronunciation?

Children acquire languages easily, but adults must commit tremendous effort to learn a new language. Adults learners often chafe at their restrictive language skills during the early stages of non-native language learning. Messages that would be conveyed clearly in the learners' native languages are transmitted slowly, vaguely and imprecisely. Language practice tends to be repetitive, mechanical, monotonous and void of intellectual content. If we could alleviate the learners' and instructors' boredom, learners might become more motivated to learn (thereby acquiring and retaining more skills) and instructors might focus on teaching tasks that require the instructors expertise.

In particular, although various proposals have been made for TJSL pronunciation teaching, none provide the learner with useful feedback such as "Your pronunciation is intelligible," "Native speakers will not understand you," "You should do such and such to improve your pronunciation," and so forth. The system ought to determine when non-native accents cannot be removed further given that most adults never attain complete nativeness, it is pragmatically useful for the system to say "Your pronunciation is at a level where native speakers will understand you. However, your skills are not likely to improve beyond this point. I suggest you stop practicing at this time." A system that explains how to improve and indicates when to stop training is a system that teaches similarly to human instructors.

To gain insight to these issues, a CALL (computer-aided language learning) system is implemented for assisting entry-level non-native adult learners. The system helps learners acquire pronunciation skills within a reasonably short learning period. By choosing pronunciation problems that occur frequently and/or disrupt communication seriously, and by using classroom-proven techniques that teach how to avoid such mistakes, automated pronunciation learning is shown to be feasible, reliable and valid.

The dissertation covers three pronunciation areas. The first area is phone duration, which deals with phones phonemically distinct based on duration. The second area is phone quality, which

deals with how phones in the learner's native language affects the articulation of phones in his target language. The third area is pitch, which deals with with how pitch is used to characterize lexical items and fixed phrases.

Skills specifically not covered in this dissertation are sentence-level and discourse-level pronunciation skills such as intonation patterns of interrogative sentences or back-channeling behavior. These skills are directly relevant to pronunciation teaching but at this time teachers do not have well-established teaching strategies that can be used with computerized self-study. Also not covered are higher-level linguistic skills such as choosing words, building sentences, interacting with the listener and so on. These skills are important for acquiring language but are beyond the scope of this study as they are not specific to pronunciation production.

The three pronunciation skills — duration, pitch and phone quality — are concentrated on pronunciation errors at the phone or word level rather than at the sentence level. High reliability and validity is achieved by targeting subskills, such as particular phoneme sets, rather than judging an entire sentence as an amalgam. By combining durational and spectral differences within phones and pitch changes across phones, we cover all pronunciation skills necessary for producing correct speech given well-formed target language words or phrases. Japanese and English were chosen as the target languages in this study to demonstrate the language-independent nature of the proposed methods.

4. PHONE DURATION

4.1. INTRODUCTION

Acquiring the pronunciation of individual phones is a crucial requirement in non-native language training because the pronunciation of entire utterances depends heavily upon the correct pronunciation of each phone. Unfortunately current methods of teaching are inefficient because they rely on human instructors who in many cases cannot spend much time with each student. Using CALL to teach many students individually in parallel may increase skill acquisition and retention by shortening the time required to detect and correct pronunciation errors early in the student's learning careers.

The pronunciation of Japanese tokushuhaku (long vowels, the mora nasal and mora obstruents) is phonemically distinct based on phone duration. Long vowels and short vowels are spectrally almost identical but their phone durations differ significantly. Similar conditions exist between mora nasals and non-mora nasals, and between mora and non-mora obstruents. We built a system that measures tokushuhaku phone duration using speech recognition technology. The system tells the learner the likelihood of native speakers understanding the learner's utterance as the learner intended.

The system's contribution to the field is that this is the first pronunciation learning system that provides feedback similar to human teachers. Two key technologies are involved: (1) using speech recognition to accurately measure phone duration, and (2) pedagogically evaluating the learner's phone durations. The know-how incorporated into the system might be applied to various areas, including (1) teaching other languages where duration is phonemic (e.g., Swedish), (2) teaching languages where duration denotes lexical stress (e.g., English), and (3) improving speech recognition performance by using phone duration information.

This study proposes a method that automatically measures the duration of tokushuhaku phones produced by non-native talkers by using a speech recognizer incorporating native phone models of L2. The system instructs the learner to read aloud the words appearing on the computer screen. The speech recognizer's output includes the duration of each phone in milliseconds. The durations of phones that are phonemically distinct based on duration are used to estimate the intelligibility of those phones. Intelligibility estimates are based on perception experiments run on native speakers of L2. Since duration is single-dimensional, corrective feedback to the learner consists of easy-to-comprehend instructions such as "lengthen your [o]," or "shorten your [a]."

For instance, let us assume a native speaker of Chinese who is beginning to speak Japanese. We know from TJSL research that such speakers tend to shorten double-mora vowels — a long [aa] becomes a short [a] for example. By using a pronunciation network that looks for the phone [a] of any length, the speech recognizer identifies the start-point and end-point locations of [a] with re-

spect to the length of the utterance, and from start and end points calculate the duration of [a]. This duration is compared with results from a perception experiment run on native speakers of Japanese, where subjects were asked to discriminate short and long tokushuhaku that had been synthesized with various durations. The perception experiment is the basis of intelligibility scores that are fed back to the learner. The learner repeats saying the reading material until his intelligibility scores rise to his goals.

This system expects strong support from language teachers. Tokushuhaku is a basic pronunciation skill that needs to be taught at the very beginning of TJSL study for two reasons: (1) the functional load of tokushuhaku is significant, with the ratio between short and long phones being roughly 3:1, and (2) tokushuhaku pronunciations are reflected in the kana orthography, which is taught in the first several weeks of study. Most learners have difficulty with tokushuhaku because practically all learners' native languages lack phonemic distinction based on phone duration.

4.2. SYSTEM STRUCTURE

The overall system-user interaction is shown in figure 4.1. First, known reading material is presented to the learner. Next, the learner's speech is forced-aligned by the speech recognizer (i.e., phone boundary locations are obtained with respect to the beginning of the utterance given a correct transcription or similarly tightly constrained language model of the utterance), and tokushuhaku phone durations are measured (figure 4.2). Each duration is compared with results from perception experiments ran on native speakers (this procedure is explained in section 4.6). Feedback to the learner consists of (a) an intelligibility score showing the percentage of native speakers who will understand the learner's pronunciation, (b) instructions on whether to lengthen or shorten the tokushuhaku, and optionally, (c) the tokushuhaku duration in milliseconds. An example of a feedback display is given in figure 4.3. The process flow of the system is shown in figure 4.4.

For speech recognition, HTK v2.1 [36] was used with gender-dependent phone models based on [28]. Prior knowledge of the reading material is used to determine whether a phone was a tokushuhaku or not. Audio input is 16-bit linear sampled at 16 kHz, using a desktop electret condenser microphone. The entire system runs on a Sun workstation in realtime. Each practice turn takes 6 seconds (3 seconds record, 3 seconds playback).

The reading material of this system is comprised of minimal pairs of actual words, for example "kado" (corner) and "kaado" (card). The learner may choose at any time to listen to a native speaker's recording of the reading material. Doing so tends to sway the learner's speech rate towards the native model's. After the learner reads the word pairs, he immediately receives an intelligibility score and instructions on how to correct his pronunciation. For instance, the feedback might be "Your kado can be understood by 100 percent of native speakers, but your kaado can be understood by only 10 percent. Say kaado longer." Depending on the phone's duration, the system instructs the

learner to "say it longer" or "say it shorter." This kind of feedback is straightforward regardless of the learner's educational background.

4.3. SYSTEM EVALUATION

4.3.1. RELIABILITY AND VALIDITY

The system's pedagogical reliability is based on its accuracy of forced alignment. In order to determine how accurately the system measures phone duration, we compared hand-labeled and system-measured segment durations. Comparing 1651 phones obtained from 3 non-natives showed that all durations of phones occurring within a word were measured at differences at or below 10 ms (10 ms being one frame width for the speech recognizer). Given that the appropriate durations of short and long tokushuhaku differ at magnitudes significantly larger than 10 ms, measurement differences of 10 ms are well within the acceptable threshold.

The system's pedagogical validity is based on the appropriateness of corrective feedback, which in turn is based on estimates of tokushuhaku intelligibility. To clarify what durations are unambiguously perceived by native speakers as tokushuhaku for particular cases, we ran perception experiments of native speakers judging artificially altered tokushuhaku durations.

Minimal pairs of actual Japanese words differing solely on the presence or absence of tokushuhaku were chosen. Next, each word was synthesized in isolation with 13 varying tokushuhaku durations. Vowel durations were adjusted at roughly constant ratios. Preplosive closures were varied at 20 ms steps. For nasals, a moraic nasal of varying length was combined with a non-moraic nasal of fixed length (45 ms). A shallow downstep pitch based on a prosodic model was added to all words, but no lexical pitch accent was used [6][10]. Table 4.1 shows the minimal pairs along with the minimum and maximum phone durations created by the terminal analog speech synthesizer [11][26]. Double phones such as [aa] denote tokushuhaku; single phones such as [a] are non-tokushuhaku. Durations for each synthesized word were measured by hand.

All 13 different varieties of each word were played twice in random order. Twelve native speakers of Japanese were asked to categorize the words as a word containing tokushuhaku, not containing tokushuhaku, or neither of the above.

Some frequency plots of the subjects responses are shown in figure 4.5. There was almost perfect agreement among subjects with regard to short and long durations. As expected, mid-range durations were judged as ambiguous, as were overly short or long durations. Discrimination curves closely matched normal distributions (table 4.2). Tokushuhaku and non-tokushuhaku are clearly distinguishable.

Assuming the subjects' responses were normally distributed, we calculated best-fit normal distributions. Estimated best-fit cumulative normal distribution curves, overlaid on figure 4.5, show that the subjects' responses closely match normal distributions. Table 4.2 shows the means and standard deviations of these estimated normal distributions.

The percentages of curves shown in figure 4.5 can be interpreted as intelligibility indices based on tokushuhaku duration. For instance, the tokushuhaku curve in figure 4.5(e) can be interpreted as the percentage of native speakers understanding a hypothetical learner's rendition of "toru" or "tooru" produced with varying tokushuhaku lengths. The level of agreement among native speakers as a particular phone being short or long tokushuhaku indicates the appropriateness of that phone. By knowing the learner's intention beforehand, we can provide corrective feedback that quantifies the likelihood of the learner being understood correctly.

The estimated means of subjects' responses reflect short and long tokushuhaku durations having a 50 percent likelihood of being understood correctly. For non-native speakers to be understood correctly significantly better than chance, the training system should aim for learners to say tokushuhaku phones at intelligibility rates higher than 80 percent. Thus the high intelligibility range for short tokushuhaku translates to durations of no more than 0.84 standard deviations below the mean (for long tokushuhaku, no less than 0.84 standard deviations above the mean). These 80-percent-intelligibility durations are also listed in table 4.2. The appropriate durations for short and long tokushuhaku differ considerably; long tokushuhaku should be between 1.3 and 3.0 times longer than short tokushuhaku, depending on the phone.

The bottom section of table 4.2 shows that the 19 tokushuhaku durations were concentrated within a reasonably small range. Thus when designing pronunciation lessons for minimal pairs not studied in the perception experiment, average means and standard deviations may be substituted; intelligibility distributions for short tokushuhaku can be approximated by N(105,25), and for long tokushuhaku by N(126,25).

4.3.2. LEARNING EXPERIMENTS

We ran an evaluation experiment in order to study the effectiveness of the proposed system. The objective was to measure (a) the improvement of pronunciation of particular tokushuhaku minimal pairs after using our system to practice them, (b) the improvement of pronunciation of tokushuhaku minimal pairs that were not taught explicitly, and (c) the length of time needed to acquire sufficient skills. The first measurement indicates the system's responsiveness to the learner's mistakes. The second measurement indicates the degree of skill transfer. The third measurement indicates the learner's workload when acquiring skills.

The experiment consisted of three consecutive stages (figure 4.6). All three stages were conducted in one seating without rest periods.

In stage one, subjects were asked to read all 19 tokushuhaku minimal pairs in table 4.2 one pair at a time in short-long order. The subjects' utterances were recorded and immediately played back for aural review by the subject. Each subject was asked to repeat reading each pair until he determined he had done the best he could. The last rendition was used for analysis. Subjects did not listen to native speaker renditions during stage one. Some of the minimal pairs were unfamiliar to the subjects.

In stage two, subjects used the proposed system to practice pronouncing three of the 19 word pairs read in stage one. The word pairs "hata/hatta", "kado/kaado" and "kona/konna" were chosen to represent plosives, vowels and nasals. Subjects had the option to, but were not forced to, listen to a native speaker reading the word pairs at any time. Subjects were instructed to continue practicing each word pair until they received "excellent" ratings (intelligibility at or above 80 percent) for both short and long tokushuhaku at least once.

In stage three, subjects repeated the task in stage one. As in stage one, the subject's last rendition was analyzed.

Table 4.3 lists the five non-native speakers who were chosen as subjects. The subjects were all male graduate students in engineering, and had stayed in Japan for periods ranging from 4 to 42 months. Table 4.3 also shows the total amount of elapsed time spent on all three stages of the experiment.

Figures 4.7(a) and (b) show scatter plots of intelligibility scores of the tokushuhaku minimal pairs that subject number 1 read at the ends of stages one and three.

Of the 19 tokushuhaku minimal pairs the subjects read, figure 4.7(a) shows data for the 3 pairs which were trained in stage two. Data points on the upper-left half of the chart (above the line denoted by y=x) indicate pronunciation improvement caused by automated training during stage two. (Each data point does not necessarily reflect a single word pair because some data points overlap.) We see that his short tokushuhaku improved greatly (his long tokushuhaku were correct to start with).

Figure 4.7(b) shows data for the 16 minimal pairs that were not trained in stage two. Intelligibility scores for most word pairs improved even though they were not explicitly taught. This result suggests that tokushuhaku skills transfer from trained phones to untrained phones, which in turn suggests that learners may acquire wide competence in tokushuhaku production after practicing on a relatively small number of tokushuhaku pairs.

Figures 4.8(a) and (b) are analogous charts for subject 2, and show similar tendencies.

Table 4.4 shows the intelligibility of short and long tokushuhaku before and after training. The average intelligibility of untrained short tokushuhaku rose from 49.6 percent to 61.3 percent (n=80), indicating the transfer of tokushuhaku skills. As subjects did very well on long tokushuhaku (approximately 97 percent intelligibility both before and after training), learning patterns for short

tokushuhaku only were further analyzed (table 4.5). We found that obstruents are apparently the easiest category of tokushuhaku (average intelligibility 70.4 percent prior to training, and 76.9 percent afterwards), while vowels and nasals were more difficult (average intelligibility for both is less than 40 percent before training and about 50 percent afterwards). Further subdivision of tokushuhaku phones did not yield specific patterns regarding tokushuhaku difficulty. The poor intelligibility scores for the untrained moraic nasal was probably an error caused by the small sample size.

Table 4.6 shows the percentage of short tokushuhaku (vowels, nasals and obstruents) whose intelligibility scores increased by more than 0 after training. Tokushuhaku skills that were trained we trained until subjects obtained intelligibility scores of 80 percent or more; the reason why the average scores of trained short tokushuhaku was 60 percent is probably because some subjects forgot how to produce short tokushuhaku correctly. This result suggests that longer and/or repeated training sessions are needed for tokushuhaku skill retention.

Figure 4.9 shows intelligibility scores of all subjects as they used the system to practice during stage two. Each set of connected data points represents a training session for a particular word pair. Although scores occasionally fluctuate (possibly because the subjects were exploring various pronunciation strategies), scores generally improve with each practice turn, eventually reaching high levels. The system's usefulness is suggested by the fact that subjects successfully completed training for each word pair after at most 20 turns.

The reason why subjects had excellent long tokushuhaku but not necessarily good short tokushuhaku may be that, on the one hand, the subjects had considerable exposure to spoken Japanese and were well-aware of the difficulties of long tokushuhaku. Advanced subjects (subjects 4 and 5) had apparently acquired tokushuhaku thoroughly, as indicated by short learning times in figure 4.9. On the other hand, subjects at the intermediate level (such as subjects 1, 2 and 3) may have not paid attention to short tokushuhaku; these subjects may have erroneously adopted a pronunciation strategy to lengthen all tokushuhaku rather than correctly discriminate short and long.

Further study is needed to verify the effectiveness of the system in actual learning settings. Given that each experiment seating took less than half an hour, the workload of the system appears to be within the expectations of pronunciation self-study. The proposed system may complement the teaching of reading and writing skills in the TJSL classroom.

4.4. DISCUSSION

The proposed system uses speech recognition algorithms to accurately measure the duration of tokushuhaku, a set of phonemically distinct phones most non-native learners have difficulty with. Perception experiments showed that while native speakers tolerate a wide range of tokushuhaku duration, subtle changes in duration can differentiate short and long tokushuhaku. The level of

agreement among native speakers as a particular phone being short or long tokushuhaku indicates the appropriateness of that phone.

The learner receives authoritative, corrective feedback that allows efficient acquisition of skills. The system explains to the learner what his mistake was, the severity of the error, and how to correct his mistake. Presenting learners with an intelligibility score that shows the percentage of native speakers who will understand the learner's pronunciation is significant improvement over conventional techniques, which at most merely return a good/bad categorical result. The new method informs the learner how far he has progressed in easy-to-grasp terms. This is the first CALL system for Japanese pronunciation learning that provides feedback similar to professional instructors.

Another benefit is that the learner can terminate training when his communicative performance has met his expectations. For instance, when a learner hits a learning plateau, intelligibility indices can help him decide whether further learning effort is worthwhile. Given that most adult learners never attain complete nativeness, it is of practical use to be told when non-native accents cannot be removed further.

Several issues remain. First, further experiments are required to assess the system's usefulness in the language classroom; in particular, we need a large-scale study based on a group non-natives who share the same native language and have spent minimal time studying Japanese. Second, perception experiments on phone quality are needed to quantify the naturalness of each phone's pronunciation; this information may help learners remove non-native accents. Third, pitch information might be overlaid on phones to teach Japanese pitch accent. Finally, efforts for evaluating pronunciation subskills should be scoped within an integrated curriculum for teaching oral production.

Table 4.1.
Minimal pairs and their synthesized tokushuhaku duration ranges.

phone	tokushuhaku minimal pair		phone duration min [ms]	max [ms]
	short	long		
i	biru	biiru	60	288
i	chizu	chiizu	64	197
e	kaite	kaitee	45	335
e	seki	seeki	44	345
a	kado	kaado	40	328
a	nasu	naasu	37	360
o	koi	kooi	45	288
o	toru	tooru	50	301
u	kuro	kuuro	44	250
u	kutsu	kutsuu	12	191
p	supai	suppai	20	260
t	hata	hatta	20	260
t	ita	itta	20	260
k	haka	hakka	20	260
k	kokee	kokkee	20	260
ch	ichi	icchi	20	260
ch	sachi	sacchi	20	260
n	kona	konna	30	200
n	hone	honne	30	200

Table 4.2.

(a) Means and standard deviations of estimated normal distributions of subjects' responses.

(b) Estimated tokushuhaku durations with 80-percent intelligibility.

(c) Standard deviations of estimated means and standard deviations.

(d) Estimated normal distribution averaged over all phones.

phone	short tokushuhaku word	mean, SD [ms]	80% intelligibility duration [ms]	long tokushuhaku word	mean, SD [ms]	80% intelligibility duration [ms]
i	biru	*N*(135,20)	118.2	biiru	*N*(150,25)	171.0
i	chizu	*N*(105,18)	89.9	chiizu	*N*(125,27)	147.7
e	kaite	*N*(140,15)	127.4	kaitee	*N*(155,20)	171.8
e	seki	*N*(95,20)	78.2	seeki	*N*(110,22)	128.5
a	kado	*N*(100,20)	83.2	kaado	*N*(120,30)	145.2
a	nasu	*N*(105,27)	82.3	naasu	*N*(125,27)	147.7
o	koi	*N*(115,27)	92.3	kooi	*N*(135,27)	157.7
o	toru	*N*(115,27)	92.3	tooru	*N*(137,27)	159.7
u	kuro	*N*(110,20)	93.2	kuuro	*N*(122,19)	138.0
u	kutsu	*N*(100,30)	74.8	kutsuu	*N*(110,25)	131.0
p	supai	*N*(90,25)	69.0	suppai	*N*(120,27)	142.7
t	hata	*N*(110,30)	84.8	hatta	*N*(150,25)	171.0
t	ita	*N*(110,27)	87.3	itta	*N*(140,27)	162.7
k	haka	*N*(100,30)	74.8	hakka	*N*(120,20)	136.8
k	kokee	*N*(100,25)	79.0	kokkee	*N*(120,22)	138.5
ch	ichi	*N*(100,35)	70.6	icchi	*N*(130,35)	159.4
ch	sachi	*N*(110,30)	84.8	sacchi	*N*(140,25)	161.0
n	hone	*N*(50,25)	29.0	honne	*N*(65,25)	86.0
n	kona	*N*(97,17)	82.7	konna	*N*(115,22)	133.5
		SD of means = 17.3 SD of SDs = 5.1			SD of means = 18.9 SD of SDs = 3.6	
average mean and SD		*N*(104.6,24.6)			*N*(125.7,25.1)	

Table 4.3.

(a) Subjects' linguistic background.

(b) Total time spent on all three stages of the experiment.

subject number	native language	length of stay in Japan [months]	total elapsed experiment time [minutes]
1	Chinese	4	32
2	Korean	4	14
3	Chinese	16	21
4	Chinese	34	16
5	Hebrew	42	11

Table 4.4

Intelligibility scores of short and long tokushuhaku before and after training.

Gain is the difference between *before* and *after* scores, divided by the *before* score, and multiplied by 100.

training	number of tokushuhaku	short tokushuhaku intelligibility scores			long tokushuhaku intelligibility scores		
		before	after	gain	before	after	gain
trained	15	57.7%	75.3%	30.6%	99.7%	100.0%	0.3%
untrained	80	49.6%	61.3%	23.7%	96.9%	97.0%	0.1%
both	95	50.8%	63.5%	24.9%	97.3%	97.5%	0.2%

Table 4.5.

Intelligibility scores of short tokushuhaku (vowels, nasals and obstruents) before and after training.*Number* is the number of scores obtained from the subjects. *Gain* is the difference between *before* and *after* scores, divided by the *before* score, and multiplied by 100.

training	short vowel intelligibility scores				short nasal intelligibility scores				short obstruent intelligibility scores			
	number	before	after	gain	number	before	after	gain	number	before	after	gain
trained	5	16.0%	52.0%	225.0%	5	67.0%	83.0%	23.4%	5	90.0%	91.0%	1.1%
untrained	45	40.7%	55.6%	36.6%	5	8.0%	16.0%	100.0%	30	69.8%	77.5%	11.0%
both	50	38.2%	53.9%	41.1%	10	37.5%	49.5%	32.0%	35	70.4%	76.9%	9.1%

Table 4.6. Short tokushuhaku (vowels, nasals and obstruents) whose intelligibilities improved after training.

Percentages reflect the number of short tokushuhaku whose intelligibility scores increased by more than 0 percent after training. *Number* is the number of scores obtained from the subjects.

training	vowels	nasals	obstruents	all
trained	80% (*n*=5)	60% (*n*=5)	40% (*n*=5)	60% (*n*=15)
untrained	58% (*n*=45)	80% (*n*=5)	53% (*n*=30)	58% (*n*=80)
both	60% (*n*=50)	70% (*n*=10)	51% (*n*=35)	57% (*n*=95)

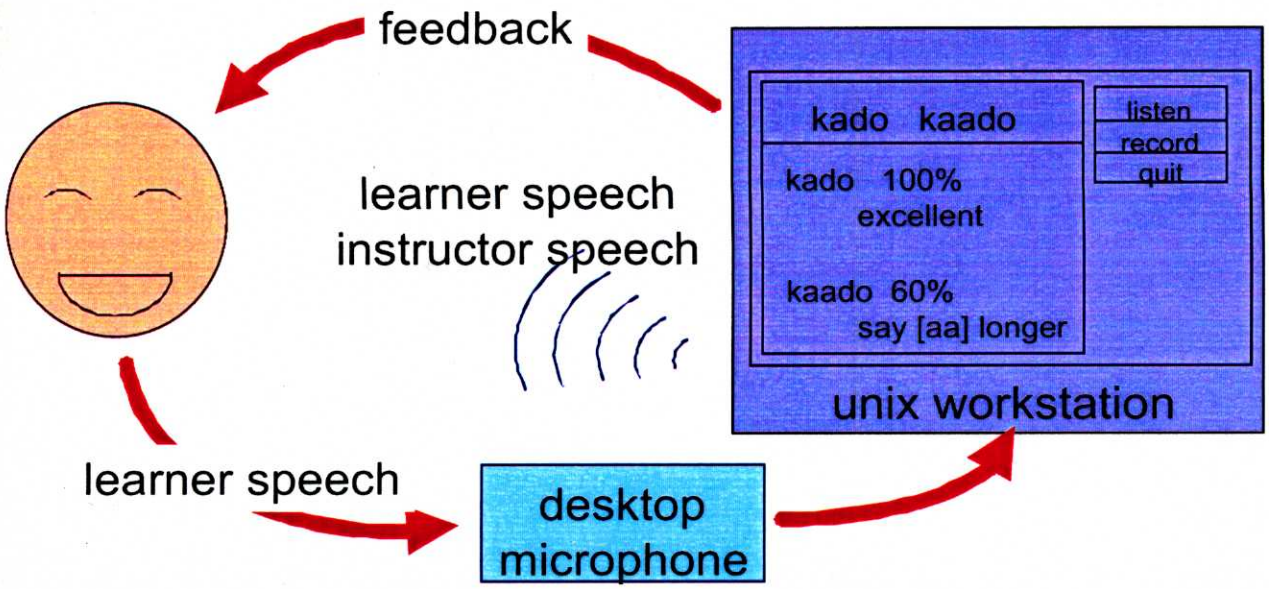


Figure 4.1.

System-user interaction of phone-duration system. Learner receives intelligibility scores on tokushuhaku durations. This figure shows the screen in English only. The actual system uses both kana and English, and with more detailed feedback.

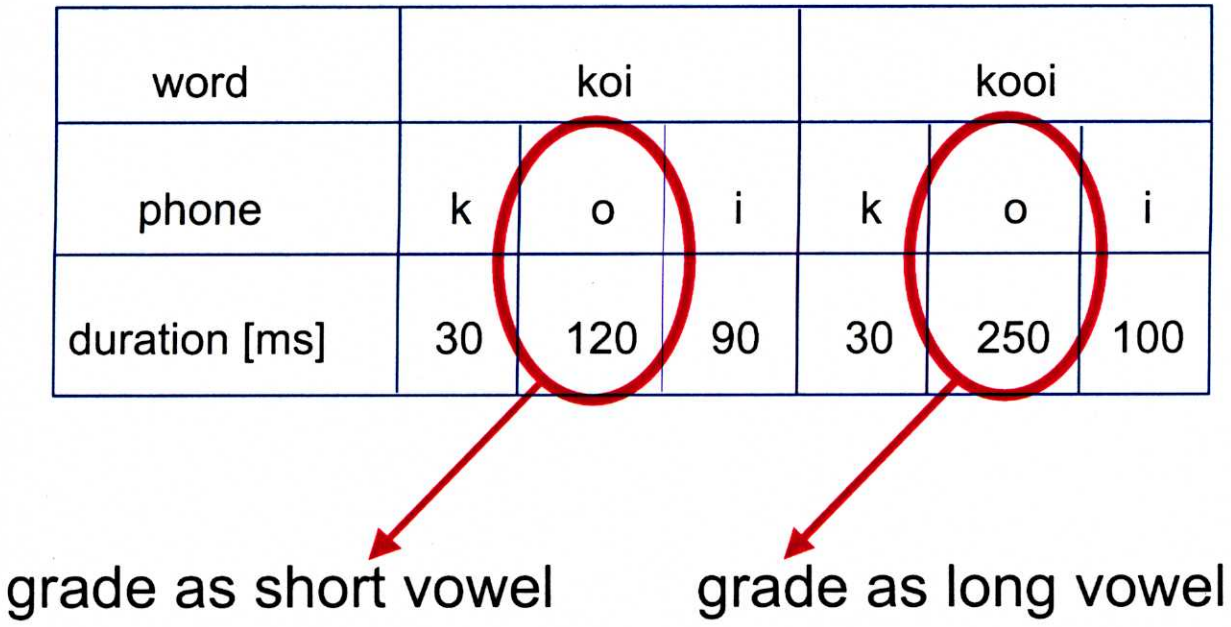


Figure 4.2.

Phone duration measurements. The locations of phones within the utterance are determined using a tightly constrained pronunciation network. Durations are measured by the number of frames (1 frame = 10 ms). Knowledge of the reading material is used to decide whether a phone is a short or long tokushuhaku. The phone labeled [sil] in this figure denotes silence, as found at the beginning and end of the utterance, between words, or before stop bursts.

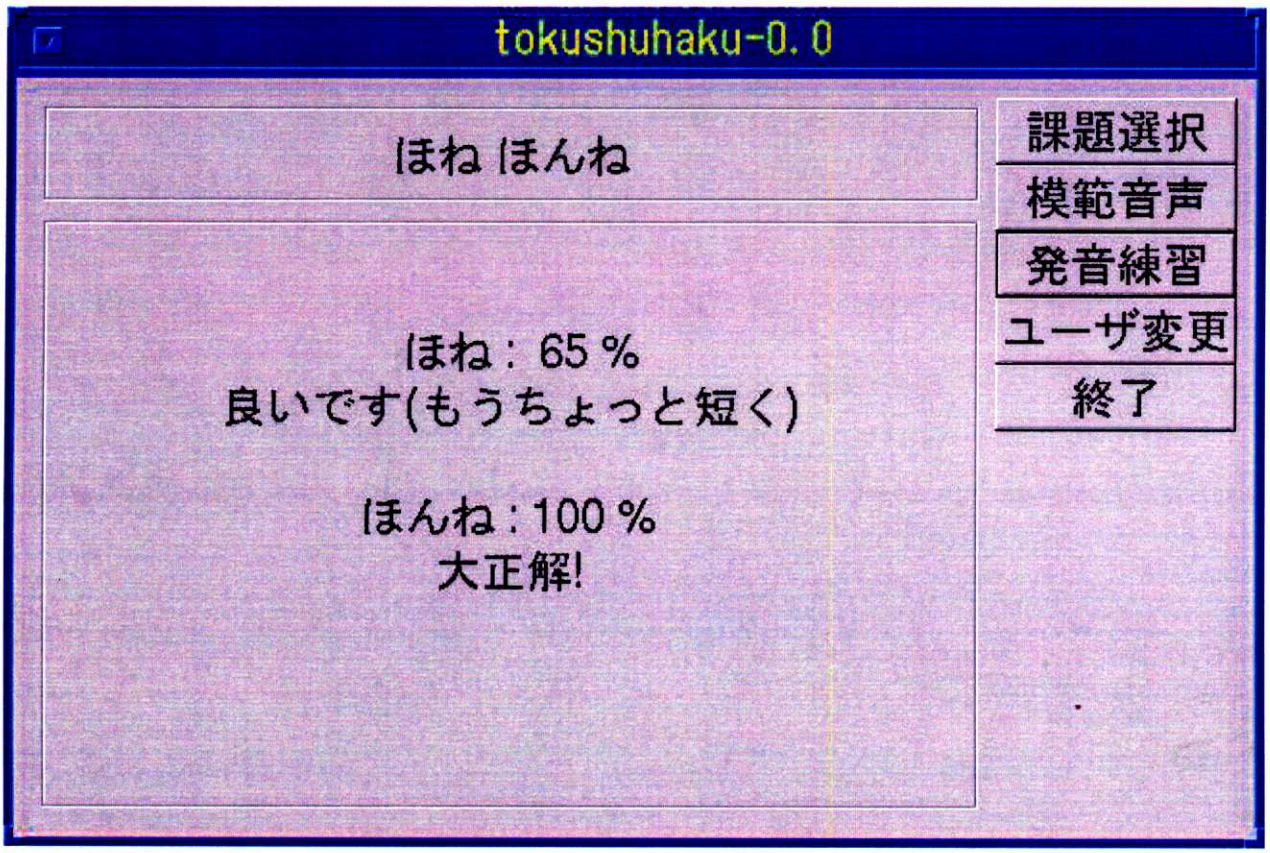


Figure 4.3.

Phone duration (tokushuhaku) training system screen example. Here the system prompted the learner to say "hone" (bone) and "hon-ne" (true thought). The learner's speech was graded as having 65 percent intelligibility for "hone" and 100 percent perfect score for "hon-ne." The learner is advised to shorten his "hone" a little bit.

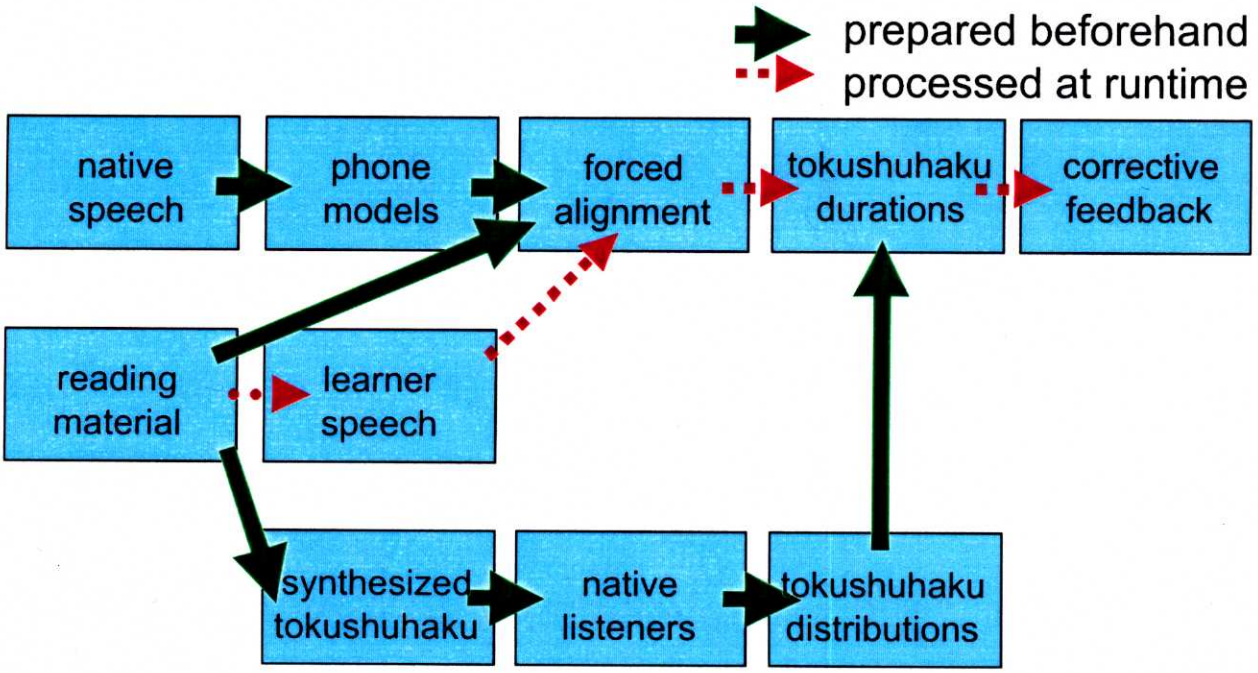


Figure 4.4.

Process flow of phone-duration system. Native listeners judged the appropriateness of tokushuhaku phones based on synthesized tokushuhaku durations. The judgements form the basis of intelligibility scores. Speech recognition accuracy is augmented by using forced-alingment techniques.

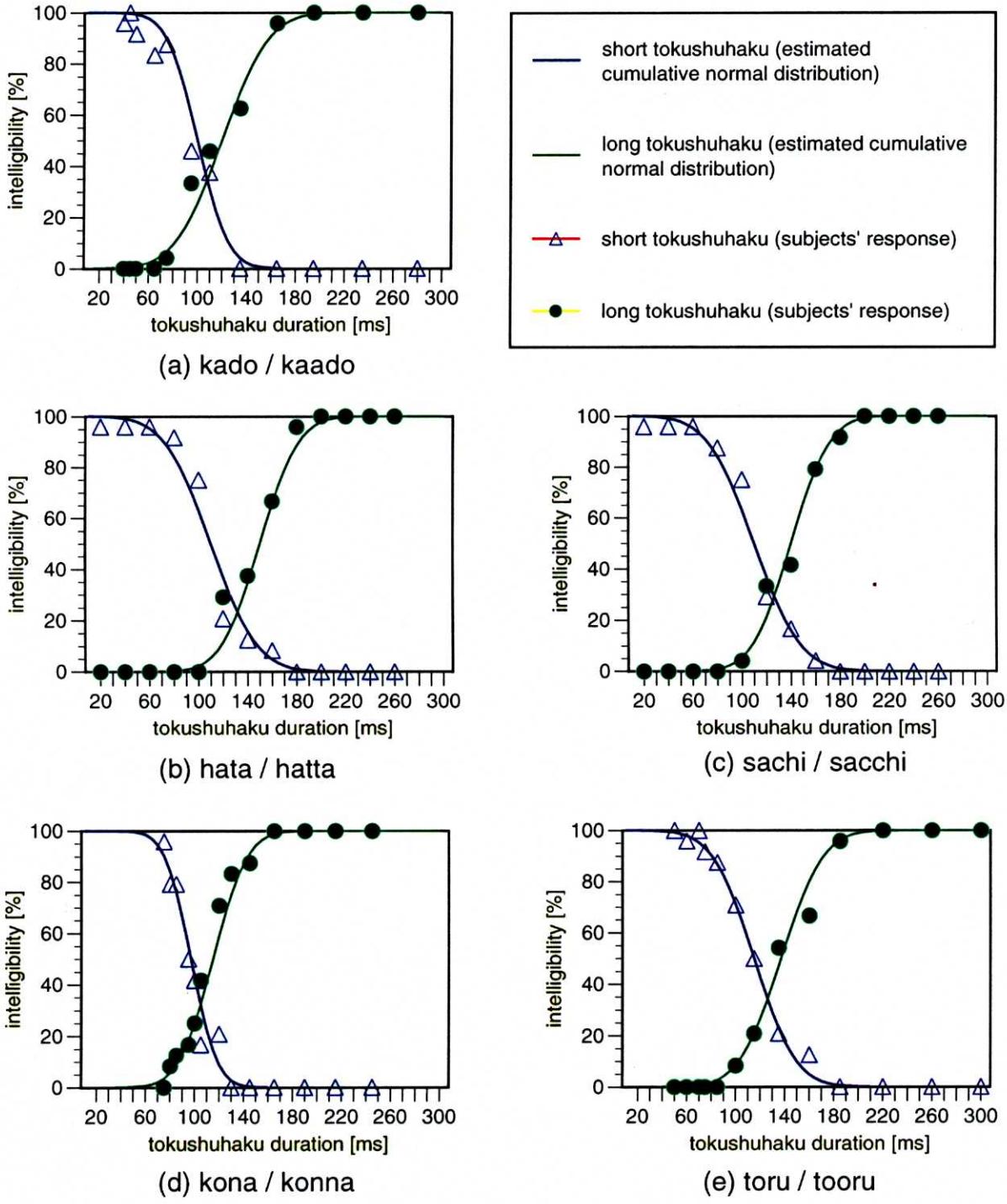


Figure 4.5.

Results of perception experiment using native listeners and synthesized tokushuhaku durations.

(a) Native speaker subjects' responses in perception experiment where tokushuhaku durations were altered in 13 steps using a speech synthesizer. Vertical axes show the percentage of subjects who judged a that particular duration as short or long.

(b) Not shown: third classification choice offered to subjects: ambiguous between short and long.

(c) Estimated best-fit cumulative normal distributions overlaid on subjects' responses as given in (a).

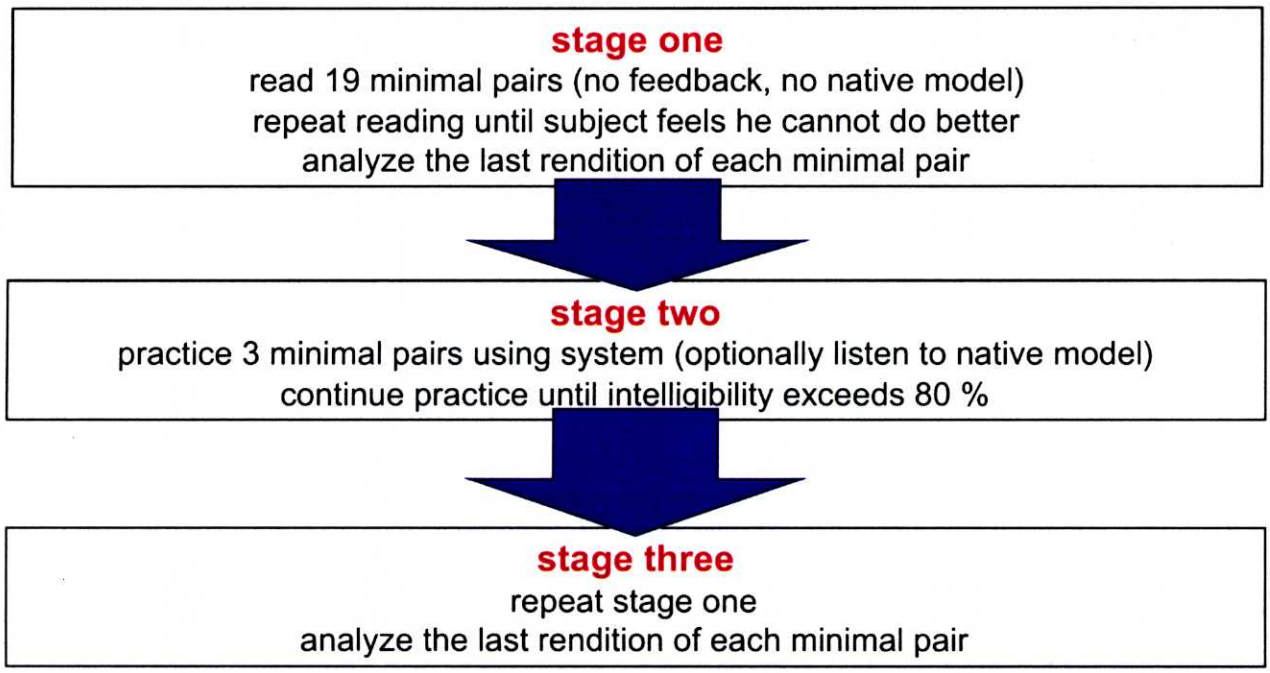


Figure 4.6.

Three stages of the evaluation experiment. The system was used to train three tokushuhaku minimal pairs during Stage Two. Tokushuhaku intelligibilities were measured for nineteen tokushuhaku pairs in Stage One and Stage Three (the before and after tests), which included the three tokushuhaku pairs trained in Stage Two.

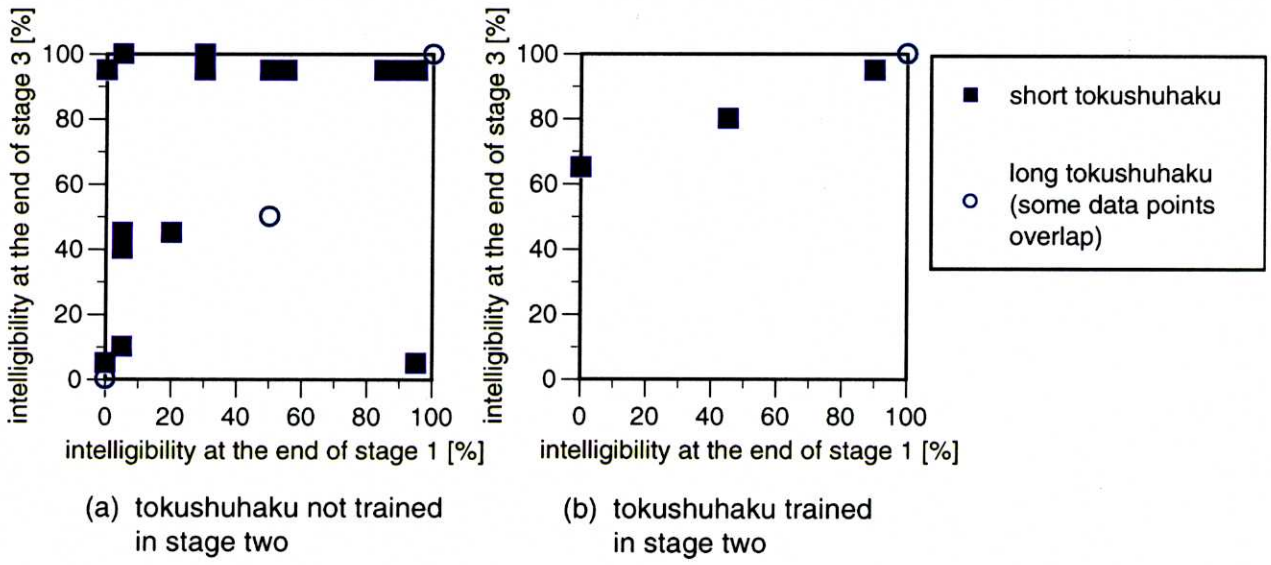


Figure 4.7.

Subject 1's intelligibility scores at the end of Stage One and Stage Three. Horizontal axis shows scores prior to training. Vertical axis shows scores after training. Figure (a) shows 16 tokushuhaku minimal pairs not trained by the system. Figure (b) shows 3 tokushuhaku pairs trained by the system during Stage Two.

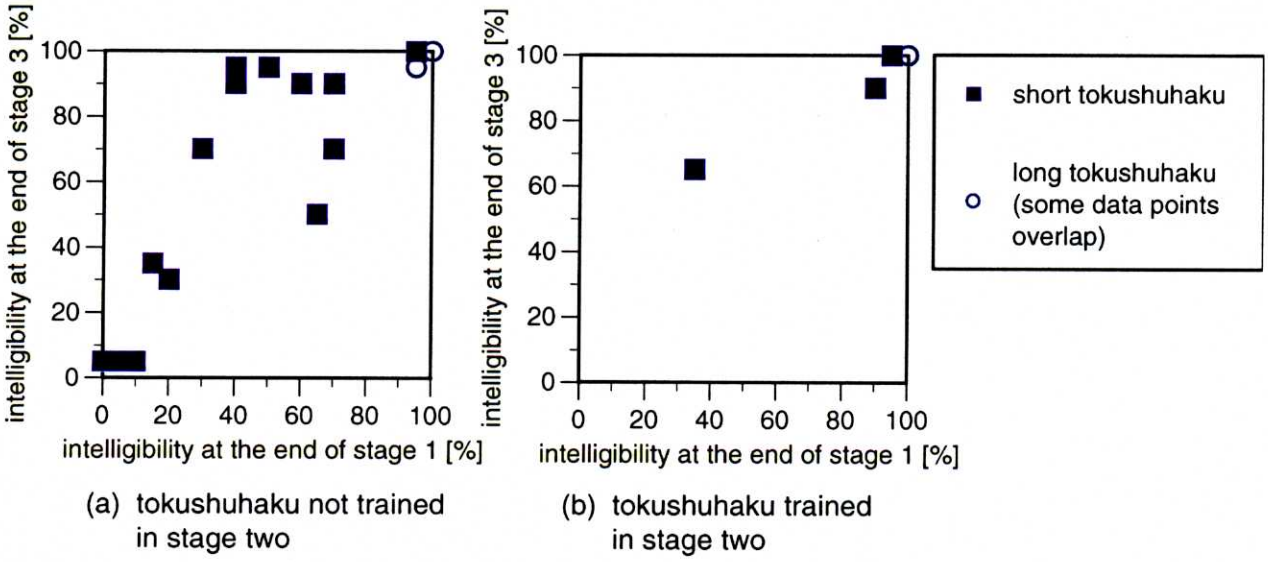


Figure 4.8.

Subject 2's intelligibility scores at the end of Stage One and Stage Three. Horizontal axis shows scores prior to training. Vertical axis shows scores after training. Figure (a) shows 16 tokushuhaku minimal pairs not trained by the system. Figure (b) shows 3 tokushuhaku pairs trained by the system during Stage Two.

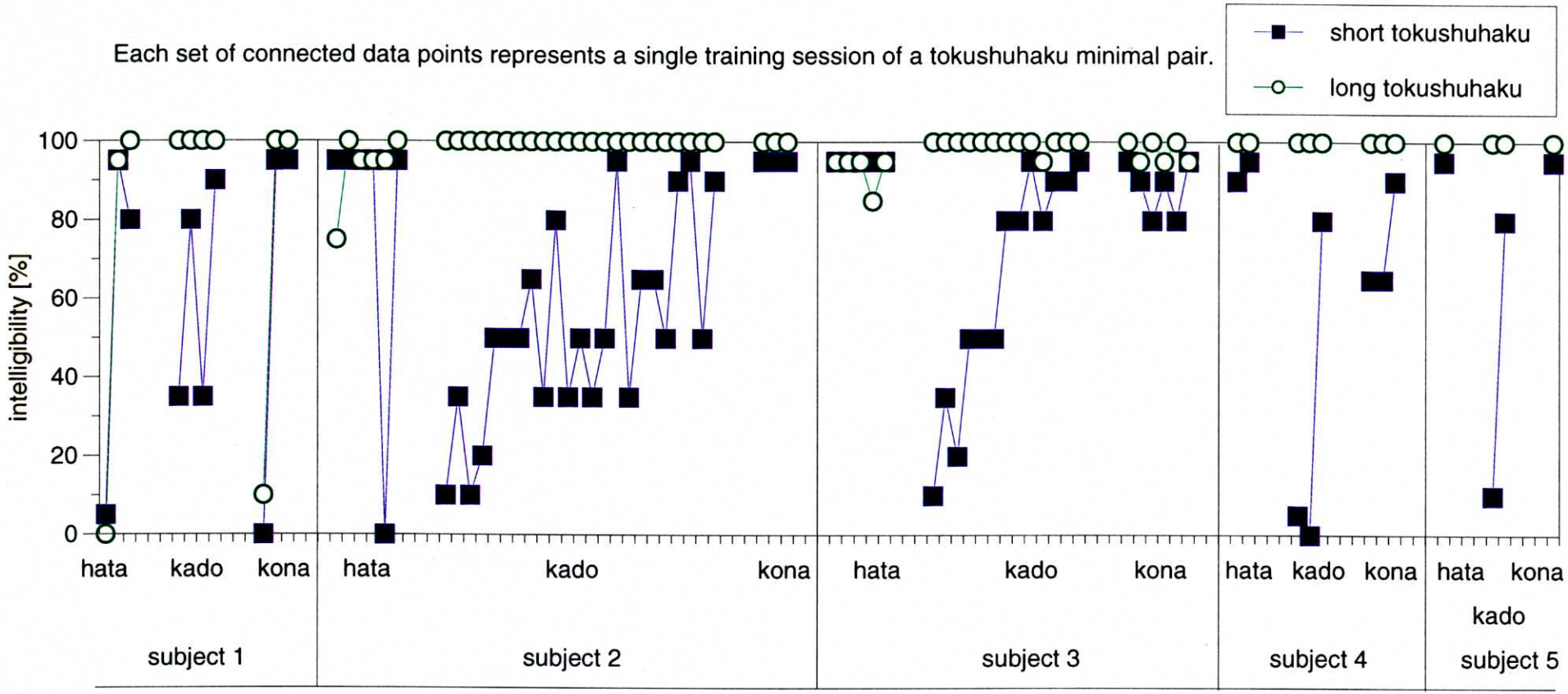


Figure 4.9.

Intelligibility scores of each of the five nonnative subjects during Stage Two. Horizontal axis shows each subject practicing three word pairs. Each tick mark is one practice turn. Each successive practice turn appears on the right of the preceding turn. Vertical axis shows intelligibility scores for each practice turn.

5. PITCH

5.1. INTRODUCTION

In this chapter we describe a method to teach pronunciation skills related to pitch. Pitch is the psychophysical correlate of fundamental frequency (F0), the latter being directly measurable for voiced speech segments, while the former is the perceived based on physical stimuli that may be intermittently present. These two are indirectly associated with phonetic properties including pitch accent, tone, and intonation patterns.

The distinction between physical, psychophysical and phonetic qualities is found also in the auditory strength of speech sounds (intensity is physical, loudness is psychophysical, and stress is phonetic) and in length of speech sounds (duration is physical, quantity is psychophysical, and the mora for instance is phonetic).

The physical factors of F0, phone intensity, spectral envelope, phone duration and so on combine to form salient phonetic features. For example, stressed syllables in English have distinct spectral envelopes, higher fundamental frequencies, and greater phone durations. The relative importance these factors play in conveying the presence of stressed syllables (and if the syllables are stressed, by how much) is not always evident. The automatic detection of stress has been elusive for this reason.

Dialects of the same language can use pitch patterns differently. For instance, a high-low-high pitch pattern at the end of questions is a neutral pattern among British English speakers but sounds condescending to American English speakers. Words and phrases in context often have more than one possible pitch pattern. Pitch production varies widely among individuals. Given this variability of pitch manifestations between languages, dialects, pragmatics, and individuals, it seems unavoidable that cross-linguistic differences of intonation patterns are not well understood, and that the teaching of pitch has proved difficult.

Up till now, computer-aided pronunciation learning systems that teach prosody by measuring F0 have not mapped F0 values to perceptual thresholds of native speakers. Current learning systems often instruct their students to fit F0 contours within a certain band that the system designates, but the band's boundaries do not to correspond to meaningful intelligibility scores.

In the previous chapter on teaching phone duration, we proposed how phone duration measurements could be mapped to perceptual thresholds of native speakers. This mapping predicted the percentage of native speakers who would understand the learner's speech the way the learner intended to be understood.

In this chapter, we propose a similar method that uses F0 to teach phonemic differences in pitch. We developed a computer-aided pronunciation learning system that automatically measures the intelligibility of pitch accents Japanese [14][19].

We take Tokyo-dialect Japanese lexical pitch accent as an example of how the proposed method can be implemented. Japanese pitch accent is a type of word accent where the mora (a subsyllabic rhythmic unit that is phonemic in Japanese) of each word adheres to specific high (H) - low (L) pitch sequences. Examples of minimal pairs used in our study include "umi (H-L)" (sea) vs "umi (L-H)" (pus), and "ikoo (H-L-L)" (after) vs "ikoo (L-H-L)" (go) vs "ikoo (L-H-H)" (shift).

Errors in pitch accent clearly signal non-nativeness, and listeners may doubt the speaker's communicative skills. Among the various factors that contribute to pitch accent perception, F0 is the most important for non-native learners. My method measures pitch by estimating the F0 of segments in the speech signal. The F0 of Japanese words are matched with native-speaker perceptions of pitch accents. Learners receive an intelligibility score along with instructions on how to speak better. Intelligibility scores motivate learners and allow them to stop practicing when their pronunciation reaches a certain point even if their pronunciation is not completely native.

Our technique might be expanded to teach the intonation of phrases and sentences that have fixed intonation patterns. Similar systems might be built for teaching pronunciation skills in any language involving pitch patterns that native speakers deem unanimously correct, such as the pronunciation of fixed expressions (i.e., phrases with fixed lexical composition and pronunciation patterns, such as "How do you do?"). Our method might also help detect prominence in speech recognition applications.

5.2. PERCEPTION EXPERIMENT

5.2.1. EXPERIMENT CONDITIONS

We ran perception experiments using resynthesized words to quantify the effects of F0 on pitch accent perception. The objectives of the experiment were to:

(a) Verify the feasibility of creating pitch-accent minimal pairs by using artificially modified pitch patterns to resynthesize actual speech signals.
(b) Qualify and quantify the relationship between F0 and moras across pitch accent boundaries (i.e., where H-L movements occur).

Two Japanese speakers (1 male, 1 female) recorded the 10 pitch-accent minimal pairs shown on table 5.1. We used resynthesis techniques to adjust the F0 of the mora immediately following the pitch accent boundary. The F0 value of the mora immediately left to the pitch boundary ($F0_i$) was

held constant at approximately 100 Hz or 200 Hz depending on whether a male or female voice was being resynthesized using a prosodic analysis-by-synthesis tool [17]. The F0 value of the mora immediately right to the mora boundary ($F0_{i+1}$) was adjusted in 9 steps, starting from a low value through $F0_i$ to a high value, so that $F0_i >> F0_{i+1}$, $F0_i > F0_{i+1}$, $F0_i = F0_{i+1}$, $F0_i < F0_{i+1}$, $F0_i << F0_{i+1}$.

Both recordings of a single speaker's minimal pair were resynthesized, so that there were 4 resynthesized sets for each pair (2 speakers x 2 words). We defined the pitch difference between two adjacent moras as

$$\text{ACCENT} = \log_e F0_{i+1} - \log_e F0_i$$

Each step in the 9-step F0 adjustment mentioned above differed at ACCENT = 0.1. Within each mora, F0 contours were essentially flat, with smoothed transitions at mora boundaries.

Although an F0 model that decomposes F0 patterns into phrasal and accent components generates better-quality speech, we did not use the model for two reasons: first, automatically calculating the parameters for phrasal and accent components from the acoustic signal is unreliable, and second, even if accurate information were available, learners will probably be unable to comprehend or apply it. A simpler method for measuring F0 makes it easier both to implement pronunciation learning systems, and also understand the corrective feedback provided by the system.

Three native speakers of Japanese listened to a randomized list of the resynthesized words in a quiet room. For each resynthesized word that was played, subjects were asked to choose between the minimal pairs and fill in a response sheet.

5.2.2. EXPERIMENT RESULTS

The subjects' responses showed there were no significant differences between the results on the stimuli created from recordings of either minimal pair. This means that experiment stimuli can be resynthesized from either H-L or L-H patterns, regardless of changes in pitch accent possibly affecting segmental features.

No significant difference was found between resynthesized stimuli created from male and female utterances. This means that experiment stimuli can be resynthesized from either male or female speech. Together with the results described in the previous paragraph, we found that perception experiments of pitch-accent minimal pair are possible using resynthesized stimuli.

Figure 5.1 shows the subjects' responses depending on ACCENT values and pitch accent patterns. We found that agreement among native speakers was practically unanimous for most ACCENT values, showing that an ACCENT value of 0.1 is needed to perceive a H-L fall in 2 or 3-mora words. (An ACCENT value of 0.1 corresponds to about a 10-Hz difference when the base F0 is 100 Hz, as was the case for words resynthesized from male speech.) For 4-mora words, an ACCENT value of 0.2 was necessary. The fact that the ACCENT value increases as a function of the number

of moras in the word suggests that even though ACCENT is the pitch difference between two adjacent mora, the difference must be exaggerated for longer words. The ACCENT value may have been larger for longer words because we did not use declination in our F0 modeling.

5.3. SYSTEM STRUCTURE

The overall system-user interaction is shown in figure 5.2. The process flow of the system is shown in figure 5.3. An example of a feedback display is shown in figure 5.4. The words taught by the system are shown in table 5.2.

The main steps of pronunciation practice are as follows. First, the learner selects reading material from a system-provided list. The learner may choose at any time to listen to a native speaker's recording of the reading material. The learner speaks into either a desktop electret condenser microphone or a close-talking electret microphone. Audio input is encoded in linear format with 16-bit sampling rate at 16-kHz sampling frequency. The learner's speech is sent to two independent processes: forced-alignment using a speech recognizer, and F0 estimation using an autocorrelation algorithm. Both processes use 10-millisecond-wide frames.

Forced alignment is performed at the phone level using HTK v2.1.1 [36] with Japanese HMMs [28]. The HMMs are gender-dependent, tied-mixture triphones using 12th-order melcepstra, their deltas and delta-deltas, and delta and delta-delta power. The system asks the learner's gender at the beginning of a pronunciation practice session. Based on knowledge of the reading material, phones comprising the same mora are joined together.

The speaker's gender information is used also by the F0 estimator to avoid half-period and double-period errors. Thus F0 estimates have maximum and minimum values depending on the speaker's gender. The F0 estimator gives a confidence measure indicating voicing probability for each 10-ms frame.

When the forced-alignment and F0 estimation processes complete, the mora labels and F0 estimates are aligned with respect to time. Average F0 values of each mora are calculated by summing all non-zero F0 values of frames within the mora, and dividing the sum by the number of the non-zero F0 frames. The natural logarithm of a mora's average F0 value is the ACCENT value of that mora. ACCENT differences between adjacent moras are matched against results from the perception experiment to obtain the percentage of native speakers who will understand the pitch accent as how the learner intended.

When a mora consists mainly or totally of voiceless sounds (such as voiceless fricatives, moraic plosive closures, and devoiced or deleted vowels), the mora's average F0 value may be impossible to calculate. Interpolating the average F0 value of a voiceless mora from its neighboring moras is incorrect because interpolation assumes the speaker is predictably adjusting his pitch across the

moras. In order to reliably estimate F0 values from the speech of non-native learners, we use reading material that is predominantly voiced.

After the learner reads the words or phrase, he receives instructions on how to correct his pronunciation. Feedback to the learner consists of (a) the reading material with mispronounced portions highlighted, (b) instructions on where to raise or lower pitch, and optionally, (c) phonetic transcriptions of the reading material and the learner's rendition. For instance, the feedback might be "Your 'ame' (rain) sounds good, but your 'ame' (candy) sounds like 'ame' (rain). Lower the 'a' and raise the 'me'". Depending on the pitch pattern, the system instructs the learner to raise or lower particular portions of the reading material. This kind of feedback is straightforward regardless of the learner's educational background.

5.4. EVALUATION EXPERIMENT

Evaluation experiments were done under simulated conditions. 19 native speakers of Japanese (14 males, 5 females) and 3 semi-natives (2 males, 1 female) with at least 10 years experience using Japanese each recorded 157 words containing 79 minimal pairs. (One minimal pair, "ikoo", was a minimal trio; H-L-L vs L-H-H vs L-H-L.) Table 5.3 shows the minimal pairs recorded in this experiment. See the appendix for a complete list of minimal pairs with Japanese transcriptions, English translations and citation-form pitch accent patterns.

The pitch heights of moras in the recordings were labeled separately by hand and machine. According to phonological rules for pitch accent, the hand-labeling and machine-labeling procedures for determining pitch accent type forced the pitch heights of the first and second mora to be opposite of each other; the only possible combinations between the first and second pitch heights were H-L and L-H. In addition, only one pitch-fall was allowed within the word; once the pitch height dropped from H to L, all subsequent pitch heights were L. Applying these two phonological rules means that the chance probability of detecting pitch accent correctly was 1/n, where n is the number of moras in the word. Table 5.4 shows results of the comparison between hand-labeled and machine-labeled pitch accent patterns.

We found that pitch accent detection within long vowels (i.e., bimoraic vowels) was approximately 10 percent less accurate than within other segments. Long vowels are spectrally identical throughout except that pitch changes can occur between its two moras. The speech recognizer tended to assign long durations to the long vowel's first mora, meaning that the recognized first mora tended to include much of the second. Thus the average F0 of the first mora became closer to the second, and no pitch change could be detected. Forced alignment between spectrally distinct segmental boundaries is more accurate. Durational modeling may improve recognition performance for long vowels. Results for words without long vowels are included in table 5.4.

Minimizing the number of moras in the word helped recognition (see results for 2-mora words in table 5.4). The chance probability rose to 50 percent, because the system was choosing between H-L and L-H patterns. Recognition performance rose significantly for both native and semi-native speakers. These results, especially for 2-mora words, suggest that the system is a useful component technology for pitch accent training.

5.5. DISCUSSION

This chapter described a method that (a) measures pitch patterns produced by non-native speakers, (b) compares non-native and native speech, obtain a intelligibility score, and (c) instructs the non-native on how to correct his pronunciation.

The proposed system uses speech recognition algorithms and prosody analysis algorithms to accurately measure pitch and align it with the location of each phone in the learner's speech. This technology might allow teaching of pitch accent and intonation contours to non-native learners. Perception experiments showed that while native speakers tolerate a wide range of pitch height within a syllable, subtle changes in pitch across syllables can differentiate high and low pitch accents. The level of agreement among native speakers as particular adjacent syllables being H-L or L-H indicates the appropriateness of the pitch accent pattern for that syllable pair.

Evaluation results suggest that technology has advanced to a point where self-study systems might help learning foreign language pitch production if the appropriateness of a given pitch pattern can be unequivocally agreed upon by native speakers.

Table 5.1.
Pitch-accent minimal pairs that were resynthesized for the perception experiment. Recordings of both words in each minimal pair were resynthesized using artificially altered F0 values.

number of mora	pitch accent pattern	word (english translation corresponding to pitch accent pattern)
2	H-L / L-H	kiru (cut/wear), aka (red/dirt), umi (sea/pus), awa (millet/bubble)
3	H-L-L / L-H-H	ikoo (after/transition), jidou (child/automatic), korera (cholera/these)
4	L-H-L / L-H-H	ikoo (go/transition), aoi (blue/hollyhock)
4	H-L-L-L /L-H-H-H	kendoo (swordsmanship/regional road), shoogai (lifetime/public relations)

Table 5.2.
Words and phrases taught by the pitch accent trainer.

Minimal pairs.

1.1. kiru (wear/cut)
1.2. aka (filth/red)
1.3. umi (give birth/sea)
1.4. awa (bubble/millet)
1.5. ikoo (transition/after)
1.6. jidoo (child/automatic)
1.7. aoi (hollyhock/blue)
1.8. ikoo (transition/go)
1.9. kendoo (regional road/swordsmanship)
1.10. shoogai (barrier/lifetime)

Sentences containing minimal pairs.

2.1. ame no hi ni ame o kau (Buy candy on a rainy day.)
2.2. sore ijoo wa ijoo da (Exceeding that is absurd.)
2.3. hiroi kendoo no soba de kendoo no renshuu o sita (We practiced swordsmanship along a wide regional road.)
2.4. kono hurui wa moo hurui (Here's an old shaker.)
2.5. ano hasi no hasi de hasi o hirotta (She picked up chopsticks at the side of that bridge.)
2.6. watasi no shoogai no sigoto wa shoogai tuduke tai (Public relations is my lifelong career.)

Table 5.3.
Minimal pairs used in the evaluation experiment. Native speakers and semi-native speakers read 79 minimal pairs.

aoi	aka	ame	arawareru	ari	awa	anzan	igai
ikoo	izi	ijoo	izen	ituka	imi	irai	iru
umi	umu	uri	uru	engi	enzin	endoo	oi
oku	ono	omoi	on	kaki	kata	kiru	kendoo
gengo	kookai	korera	goosee	goen	goyoo	saku	shoogai
zisin	zidai	zidoo	juunin	juubyoo	juuman	joodan	joobu
suihee	sentoo	tikaku	terasu	doojoo	nagai	niwa	neru
nobi	nomi	hasi	hayaku	ban	hurui	budoo	henzi
mago	momo	moru	yamu	yameru	yoi	yooi	yoogo
yoozi	yoojoo	yoozin	yokusuru	wakai	wan		

Table 5.4.
Recognition accuracy percentages according to speaker type and segmental limitations.

speaker type	number of speakers	all words (n=157)	words excluding long vowels (n=112)	2-mora words (n=56)
natives	19	64.4	70.1	86.8
semi-natives	3	67.7	69.7	91.0
both groups	22	64.9	70.1	87.4

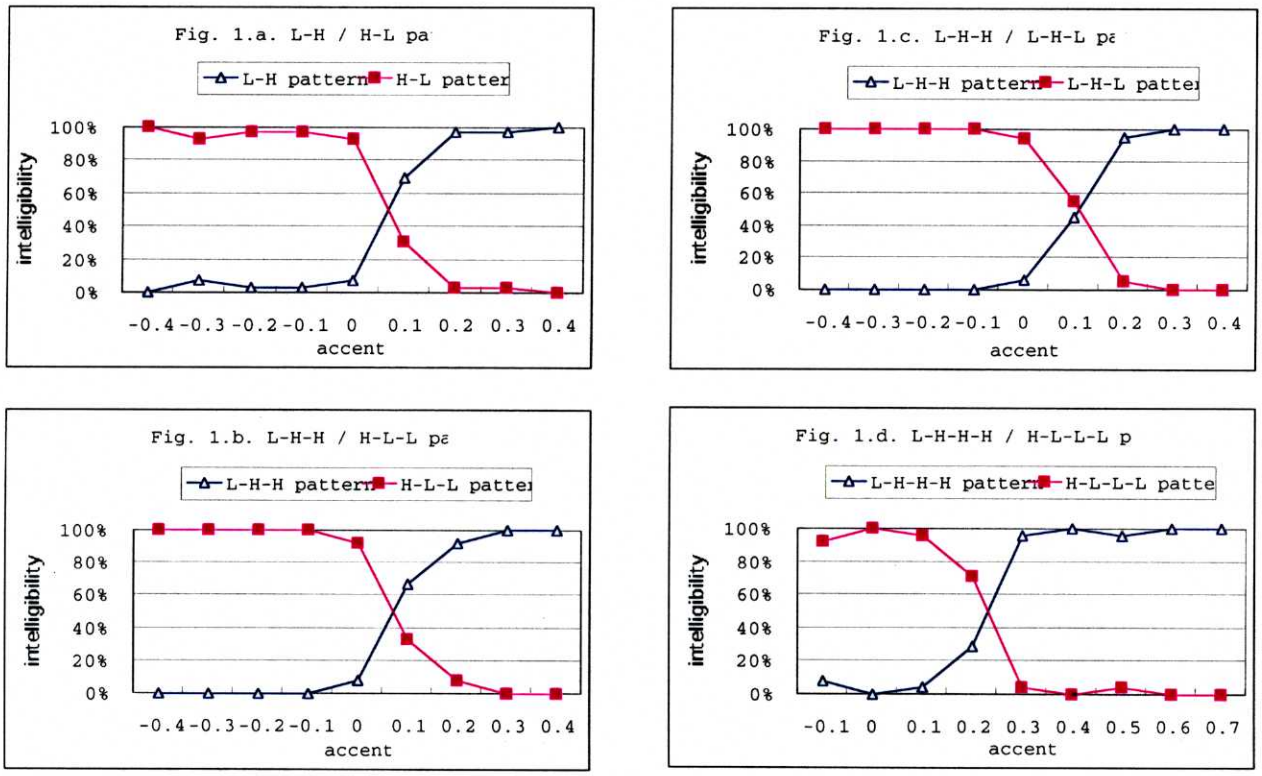


Figure 5.1.

Subjects' responses in the pitch accent perception experiment, according to ACCENT values and word length. Word length is measured by the number of moras in the word. ACCENT values are plotted on the horizontal axis; pitch heights between adjacent moras is H-L towards the left and L-H towards the right of the charts. The subjects' responses are plotted on the vertical axis. Subjects' reponses correspond to intelligibility scores for that ACCENT value.

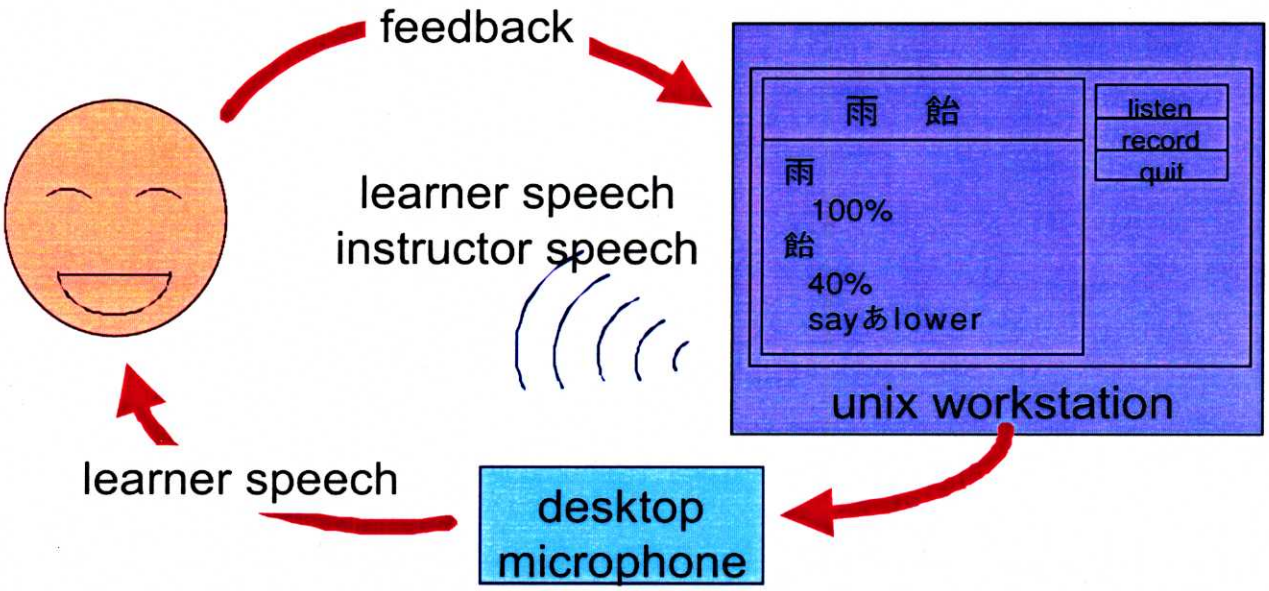


Figure 5.2.
System-user interaction of pitch accent system. System judges pitch accent patterns and calculates intelligibility scores.

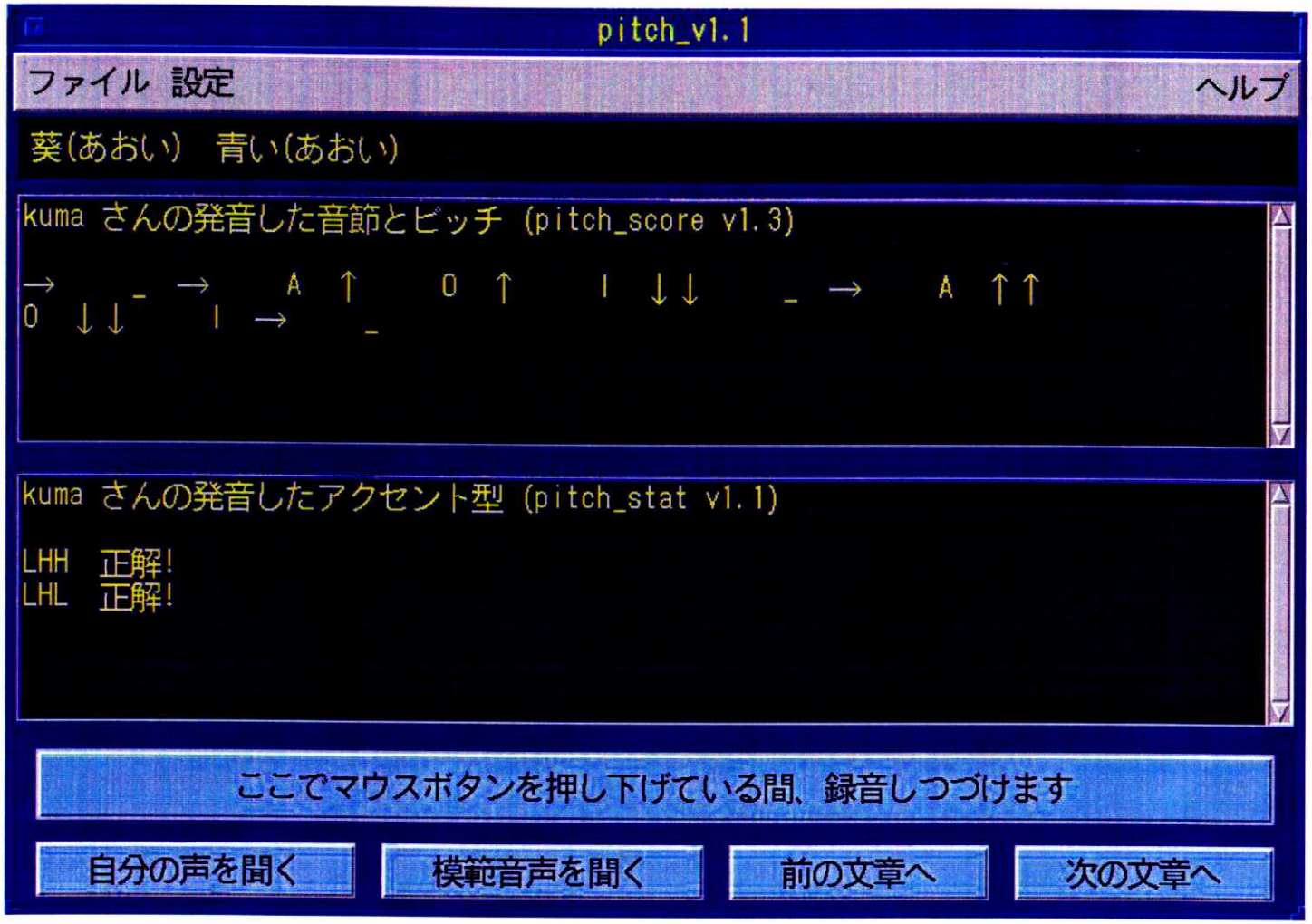


Figure 5.3.

Pitch accent trainer screen example. The system asks the learner to say "aoi aoi" (LHH vs LHL). The number of up and down arrows denote ACCENT values. The system detected a flat pitch contour at the beginning of the utterance, then found that from the first [a] to [o] and then to [i] was a rise (thus the accent pattern for the first word is LHH). From there the pitch fell to zero (interword silence). From the second [a] to [o] it rose rapidly, then from [o] to [i] fell sharply (thus the accent type for the second word is LHL). The system declares the utterance correct and congratulates the learner in the bottom window.

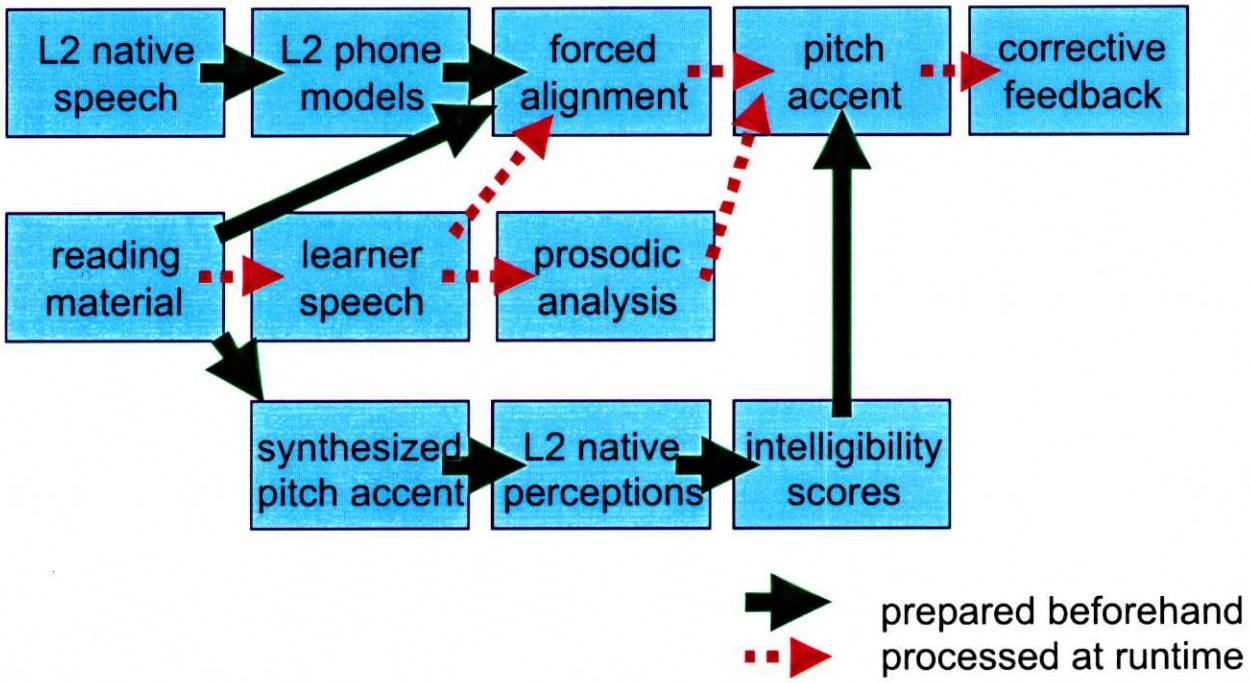


Figure 5.4.

Process flow of pitch accent system. Pitch information is extracted and matched against native listener judgements on synthesized pitch patterns.

6. PHONE QUALITY

6.1. INTRODUCTION

In this chapter we describe a method to teach pronunciation skills related to phone quality. By phone quality we refer to segmental differences between phones. Previous chapters dealt with mistakes in phone duration and pitch; both of these teaching systems assumed target-language segmental sequences appearing in correct order. This chapter deals with detecting incorrect segments that arise from the learner's unfamiliarity with phones in the target language.

When a non-native person starts learning a foreign language, his native language's phonetics and phonology affect the production of his target language phones. For instance, the syllabic structure of his native language may carry over to his target language (this issue is treated in section 6.2 phone insertion). Target language phones that are absent from his native language but must be pronounced somehow may be substituted by native language phones (this issue is treated in section 6.3 phone substitution). When phones are deleted, substituted, or inserted, foreign accent speech is often the result (this issue is treated in section 6.4 phone insertion, substitution, and deletion).

Pronunciation errors that involve phone deletion, substitution, or insertion can be corrected if the errorneous phone's phonetic properties and phonological context are known. Different articulatory gestures yield different spectral characteristics that distinguish one phone from another. By recognizing phones from their spectral characteristics, we can understand the articulatory difference between right and wrong phones. Such differences become the basis of corrective feedback to the learner. Techniques on how corrective feedback is given to the learner are explained over the following sections. In particular, we show how errors can be ranked in order of severity, so that the learner can correct his most obvious errors first (this issue is treated in section 6.5 ranking pronunciation errors).

6.2. PHONE INSERTION

6.2.1. ANAPTYXIS

Some languages have syllabic structures more restricted than others. For example, on the one hand, Japanese has a maximum of two syllable-initial consonants and one syllable-final consonant; the consonant inventory is not large; and the number of allowed consonant combinations is small. On the other hand, English allows more consonants at syllable-initial and syllable-final positions, and possible consonant combinations are abundant. Native speakers of Japanese speaking English have difficulty pronouncing consonant clusters prohibited by Japanese phonological rules. When Japanese speakers say English words, Japanese syllabic structure often carries over to English. A clear

example is loan words from English into Japanese. Loans invariably undergo anaptyxis or proparalepsis [5][15][20][32].

The same phenomenon happens when Japanese speakers learn English. Inserting vowels within consonants clusters or after syllable-final consonants is prevalent. Anaptyxis mutilates the syllable and stress structure of English, and anaptyctic speech is incomprehensible to native speakers of English even after considerable exposure to Japanese-accented speech. However most Japanese teachers of English overlook anaptyxis because they understand anaptyctic speech perfectly and are unaware of the severe impact anaptyxis has on intelligibility. Vowel insertions have remained perhaps the leading cause of miscommunication by Japanese speakers, probably exceeding the oft-cited [l] vs [r] problem because anaptyctic phonological alterations affect the entire utterance rather than a few segment types.

With this problem in mind, we implemented a system for automatically detecting inserted vowels in Japanese-accented English. Students receive corrective feedback from the system that identifies where vowels were inserted and how to pronounce the target utterance correctly. The remainder of section 6.2 describes the component technologies used in the language learning system, and evaluation experiment results.

6.2.2. ANAPTYCTIC VOWEL DETECTOR

The system's core is a speech recognizer running in forced alignment mode (i.e., phone labels are obtained given a correct transcription of the utterance or otherwise tightly constrained language model). The speech recognizer is the same forced-aligner as those used in the previous chapters, except that both English and Japanese HMMs are used.

HMMs for Japanese and English are trained separately on language-dependent native-speaker speech data as in regular monolingual speech recognition. The two HMM sets are used together during the recognition phase. In order to combine the two models, the set of features used in the HMMs must be identical (we use monophones with 12th-order melcepstra, their deltas and delta-deltas, and delta and delta-delta power). In addition, the training data's acoustic characteristics (sampling frequency, number of sampling bits, level of background noise, frequency response of microphone, and so forth) should match as closely as possible.

Implementing our method is straightforward because it uses only native speech of English and Japanese to train acoustic models. Training HMMs on non-native speech is not necessarily practical. The first reason is because building non-native corpora in scales comparable to existing native corpora is a major undertaking to start with. The second reason is we probably need even more data than existing native corpora because non-native speech probably has wider variance than native speech (this assumption is reasonable because by definition non-natives span the range between total nativeness and non-nativeness).

Our system can expect strong support from language teachers because it involves their pedagogical expertise from the beginning of system development (as opposed to delivering the final product to teachers who are seeing it for the first time and are not predisposed to using it). Collaboration between speech engineers and language teachers is crucial to successfully deploying speech-enabled CALL systems in the field.

English HMMs are used for phones that are correct phones. They are referred to as "obligatory phones" because they must appear in the practice material. Japanese HMMs are used for vowels that are inserted in erroneous speech. These phones are referred to as "anaptyctic vowels". We use Japanese HMMs for anaptyctic vowels because we found from subjective listening tests that learners tend to say Japanese phones when applying Japanese syllable-structure rules; that is to say phonetic and phonological influences of L1 into L2 coincide.

The pronunciation lattice for the anaptyctic vowel detector consists of obligatory phones (which the forced aligner will always detect) and anaptyctic vowels (which the forced aligner will detect if found in the speech signal). The possible locations of anaptyctic vowels are predetermined by the system; we use knowledge of Japanese phonology to automatically find locations where [o], [u], or [i] might be inserted (see Appendix B for phonological rules of anaptyxis). The task of the forced aligner is to determine whether an anaptyctic vowel exists in a particular location.

The system displays English words, phrases, or sentences on the computer screen and instructs the learner to read them aloud. The reading material consists of English words containing many consonant clusters and syllable-final consonants that trigger vowel insertion. In addition, many of the words have been loaned into Japanese, making it likely that learners will mispronounce them. Table 6.1 shows the list of words and phrases trained by the system. Figure 6.1 shows the overall system-user interaction. Figure 6.2 shows the process flow of the system. Figure 6.3 shows an example of the feedback display. The system alerts the learner whenever anaptyctic vowels are recognized. Figure 6.4 shows an example of a phone network for detecting vowel insertions. The learner reads the material again while taking care not to insert vowels.

6.2.3. EVALUATION EXPERIMENT

19 native speakers of Japanese (16 male, 3 female), all University of Tokyo undergraduate students with no prior experience with the system, used the system and read all words and phrases once each. Of the 19 subjects, 16 subjects (14 male, 2 female) used a close-talking noise-cancelling microphone (Sennheiser HMD-25) in a fairly noisy computer terminal room; there were multiple conversations happening in the vicinity of the subjects. The remaining 3 subjects (2 male, 1 female) used a desktop microphone (Sony ECM-K8 electret condenser in high-gain, cardiod-directivity mode) in the same computer room; the microphone was placed under the computer monitor where noises from the computer's fan and harddisk were audible.

A native speaker of English determined where anaptyxis occurred via visual and audio inspection of all recordings. Error rates for phone-level speech recognition and for detecting anaptyctic vowels (including false alarms) were calculated by comparing the system's recognition results with hand-labeled phone sequences. Phone recognition error rate is the sum of substitutions, insertion and deletion divided by the total number of underlying correct phones. Anaptyctic vowel detection error rate is the sum of anaptyctic vowel insertions and deletions divided by the number of possible anaptyctic vowel locations in the reading material.

Results from an evaluation experiment are given below in table 6.2. Figure 6.5 includes two scatter plots comparing human judgements with machine-generated scores. The system was slightly more sensitive than the human scorer. This might help detect errors conservatively. Speech recorded in noisy environments seems to be graded less reliably. Close-talking microphones raised the correlation between human and machine to over 0.9.

6.2.4. CONCLUSION

The performance of the prototype system's component technology was verified. The next step is evaluating how effectively learners learn pronunciation skills using the system.

The system can be improved by measuring the duration of elognated fricatives (e.g. [sh] [s]) that become moraic when anaptyxis occurs. For instance, "push" [p u sh: u], where the word-final obstruent [sh] is moraic and a vowel is added, resulting in a 3-mora pronunciation of a 1-mora word.

6.3. PHONE SUBSTITUTION

6.3.1. TOKUSHUHAKU PHONE QUALITY

In chapter 5, we discussed teaching Japanese tokushuhaku by measuring their intelligibility based on their durations. It is a serious pronunciation error to delete tokushuhaku, because deletion means reducing the duration to zero, which in turn reduces intelligibility to zero. All tokushuhaku must be pronounced, but obligatory phones spoken by non-natives are not always spectrally correct [7]. Nonnatives may substitute tokushuhaku phones with phones from their native languages. This might happen for long tokushuhaku vowels said by native speakers of English. Elongated English vowels sometimes become diphthongized, as in "keeri" (accounting) becoming "keiri" because many native speakers of English cannot say long [e]. Substitutions are conspicuous examples of foreign accent.

In this section, we discuss how to target language phones that are absent from the learner's native language but must be pronounced somehow may be substituted by his native language phones. We implemented a system for automatically detecting substituted vowels in English-accented Japanese

tokushuhaku. Students receive corrective feedback from the system that identifies where vowels were substituted and how to pronounce the target utterance correctly. The remainder of section 6.3 describes the core technology, which is detecting obligatory phones replaced by similar phones from the learner's native language.

6.3.2. SUBSTITUTED-VOWEL DETECTOR

The system's speech recognition component is similar to the one described in section 6.2. The main differences are (a) the substituted-vowel detector at least two vowels (one Japanese and one English) as tokushuhaku vowel candidates, while the anaptyctic vowel detector uses only one (a Japanese vowel), (b) the substituted-vowel detector always recognizes a single vowel from the candidates, while the anaptyctic vowel detector may or may not recognize a vowel, and (c) the substituted-vowel detector uses Japanese HMMs for obligatory phones (same definition as given in section 6.2.2) plus both Japanese and English HMMs for tokushuhaku candidates, while the anaptyctic vowel detector uses English HMMs for obligatory phones and Japanese HMMs for anaptyctic vowels. Of these differences, difference (c) merely reflects the difference between English speakers learning Japanese and vice versa. Difference (a) and (b) are the fundamental differences between the two systems.

When designing the speech recognizer, phone insertions can be modeled as substitutions if the system accommodates null phones. Null phones are phones with zero duration (thus with no spectral value). Phone insertions can be modeled as a null phone being substituted by a non-null phone. In addition, phone deletions can be modeled as non-null phones being substituted by a null phone. Using null phones can simplify the speech recognizer's phone lattices by generalizing insertions and deletions as subsets of substitutions. However systems that allow null phones are not necessarily practically useful, first because a special HMM must be made for them, and second because the language model must specify where null phones may occur (otherwise null phones can appear infinitely anywhere in the speech signal). The first problem is not difficult but there is no practical gain because solving the second problem is identical to specifying insertions and deletions independently. The sole benefit of considering null phones is when specifying insertions, substitutions and deletions abstractly within the speech recognizer's language model.

The pronunciation lattice for the substituted-vowel detector consists of obligatory phones and substitution candidates. Exactly one canidate will be chosen from each of the candidate sets. Candidate sets are placed at tokushuhaku vowel locations. The task of the forced aligner is to determine which candidate vowel was said.

We use knowledge of Japanese language teaching to determine which English vowels to add to the candidate sets. For instance, to capture a common mistake of diphthongizing or reducing vowels, we created candidate sets such as:

a = { j_a, j_a:, e_aa, e_ae, e_ah, e_aw, e_ay, e_ey, e_ax, e_axr }
o = { j_o, j_o:, e_ao, e_ow, e_oy, e_ax, e_axr }

where phone labels prefixed with "j_" denote Japanese phones, and "e_" denote American English phones.

Human instructors give learners general articulatory advice before and after the learning session. The system prompts learners to say Japanese tokushuhaku words in the same way as the duration-measuring system described in chapter 5. In addition to tokushuhaku intelligibility measurements, the system alerts the learner whenever non-Japanese vowels are recognized. The learner reads the material again while applying the teacher's articulatory instructions. If the learner's pronunciation does not improve after a practice session, he can ask the instructor for help. Separating the mechanical repetitions from higher-level instruction frees the teacher for more important tasks. The system-user interaction is identical to the anaptyctic vowel detector (see figure 6.1). Figure 6.6 shows the process flow of the system. Figure 6.7 shows an example of corrective feedback.

6.3.3. EVALUATION EXPERIMENT

The system detects mispronounced tokushuhaku duration and quality, such as shortened or lengthened tokushuhaku, diphthongized long vowels, palatalized plosives and inserted liquids. Figure 6.8 shows an example of forced-alignment results of the word "kado" pronounced in three ways. Recognition results closely match human judgements.

6.3.4. CONCLUSION

The paper discussed a CALL system for the teaching of tokushuhaku phone duration and quality to native speakers of American English. By using phone models for other languages, teaching phone duration and quality of other languages may be possible. Future work is planned on teaching fundamental frequency contours, such as lexical pitch accent and pitch contours of set phrases.

6.4. PHONE INSERTION, SUBSTITUTION, AND DELETION

6.4.1. CATEGORICAL RECOGNITION OF NATIVE AND NON-NATIVE PHONES

In sections 6.2 and 6.3 we developed the idea of automatically detecting, measuring and correcting non-native pronunciation characteristics in terms of phone insertions and substitutions. The proposed ideas were effective for their limited applications, but more comprehensive treatment of phone quality is required to account for systemic differences in L1 and L2 (i.e., how L1 phones are substituted for novel L2 phones), realizational differences (how similar phones in L1 and L2 can be uttered with different phonetic realizations), and structural differences (L1 phonotactics carrying over to L2 production). Specifically, we need a measure of segmental non-nativeness to correct so-called "foreign accents" in general.

Previous work in this area includes measuring non-nativeness using statistics obtained from speech recognizers using HMMs. Some systems use HMMs for L2 (the learner's target language) trained on native speakers of L2 [8]. These systems often use speaker-adaptation to accommodate the learner's speech. However these systems do not distinguish non-native speakers from nonstandard native speakers, because the phonetic and phonological effects of L1 (the learner's native language) are ignored. Other systems use HMMs for L2 trained on non-native speakers of L2 [22]. Such systems have the potential advantage of modeling the learners' pronunciation patterns more accurately. Collecting training speech data from a sizeable number of learners may be a practical problem, especially when the training speech data is to be stratified according to the speaker's pronunciation ability. Regardless of whether L2 HMMs are trained on L2-natives or L2-non-natives, the primary challenge of using HMM-derived statistics is in translating the statistics into articulatory phonetic terms — namely quantifying how serious the pronunciation error is, and indicating how to correct pronunciation errors. Most pronunciation scoring systems measure the reliability of their scores by correlating them with human judgements. Problems of this method are that human judgements do not necessarily correlate highly among themselves, and that interhuman correlations set an upper bound on performance beyond which higher reliability cannot be proven.

By contrast, we use a method that automatically measures the pronunciation quality of phones produced by non-native talkers by using a bilingual phone recognizer (the acoustic models are based on [28] and [35]). The learner receives categorical, phone-based information that is readily understood. The system detects errors in the choice of phones, reports the degree of non-nativeness of the learner's pronunciation, and suggests ways to improve speaking ability. Prior knowledge of L1 and L2 phonetics, phonology and language pedagogy are combined to identify non-native ar-

ticulatory gestures that result in pronunciation errors, thus allowing informative corrective feedback to the learner.

Our technique can be applied to any language pair by using appropriate knowledge of phonology and acoustic phone models. The remainder of this section explains details of this method's implementation for native speakers of Japanese learning American English, along with results of feasibility experiments.

6.4.2. SEGMENT CLASSIFIER

The system's speech recognition component is similar to the one described in sections 6.2 and 6.3. Whereas the section 6.2 described a system for detecting vowel insertions, and 6.3 a system for detecting vowel substitutions, here we discuss a system for detecting insertions, substitutions and deletions of any segment at any location. Pedagogical considerations dictate which phones to study or ignore. The result is a a more generalized approach than the two systems previously explained.

We implemented a system for native speakers of Japanese learning American English. Our technique can be used for other language pairs. In section 6.5, we describe a system for native speakers of English learning Japanese. The system-user interaction is shown in figure 6.9. Figure 6.10 shows the process flow. Figure 6.11 shows an example of the feedback display.

The phones in the HMM sets we used are listed in table 6.3. We use slightly broad monophone HMMs to absorb pronunciation variability among non-natives. Using triphones is possible, but cross-language phone sequences cannot be modeled by the HMMs we used because both English and Japanese HMMs were trained on monolingual native speech data. If triphones were used, then only monolingual phone transitions would be modeled in detail. Cross-language transitions will essentially be the same as monophones. The study of cross-linguistic phone sequences is an theoretically intriguing. Given time we may understand how individuals employ different strategies, most probably depending on their language abilities, aptitudes and other factors. But training cross-linguistic biphones or triphones that are good estimators of the population of non-native learners seems impractical at this point, because such language models require speech corpora containing substantial amounts of cross-language phone transitions.

The advantages of bilingual HMMs include (a) detecting L1 accents at the phone level, (b) instructing learners how to correct their pronunciation using techniques employed by language teachers, and (c) measuring pronunciation ability. Identifying L2 phones being substituted by L1 phones shows how the L2 phone was articulated, and knowing the articulatory differences between the L1 and L2 phones allows us to provide the learner with feedback similar to that given by language teachers.

As the learner's pronunciation improves over the course of practice, his choice of interlanguage allophones will become closer and closer to his target language's. We hope ultimately they become identical. The process of the learner losing his "foreign accent" might be described as the shrinking process of his interlanguage allophone sets. An American English speaker learning Japanese might cease to substitute [ax] for /a/, for instance, reducing by one the number of elements in his interlanguage /a/ phoneme. We can measure the learner's accent loss by measuring the extent of allophone set reduction:

$$\text{allophone reduction ratio} = \frac{n_{start} - n_{current}}{n_{start} - n_{correct}} \times 100$$

n_{start} : Number of allophones at start of training, including both correct L2 phones and incorrect ("accented") L1 phones.

$n_{current}$: Number of allophones currently in learner's language model. Decreases as pronunciation ability becomes increasingly consistent.

$n_{correct}$: Number of correct target phones. Typically 1, but occasionally several.

As the learner progresses in his training, his allophone reduction ratio will increase from 0 percent (totally non-native) to 100 percent (totally native), indicating that he has met his goal for that L2 phone. In the CALL system's graphical user interface, allophone reduction ratios are displayed in progress gauges labeled "accent loss gauge" (figure 6.12).

The system does not add allophones; once an interlanguage allophone is removed from the system's language model for a particular learner, the allophone is never restored. Capability to do so may be necessary for true dynamic modeling of the learner's pronunciation behavior, such as when the learner pronounces phones inconsistently compared to previous practice sessions.

6.4.3. EVALUATION EXPERIMENTS

First, an experiment designed to measure the accuracy of the system was run under simulated conditions. 340 non-native phones were studied using American English as L1 and Japanese for L2. Results show that (a) 84 percent of the learner's phones flagged by the system as having an L1 accent influenced by L1 phone x were judged by a human native L2 listener as indeed being influenced by phone x, and (b) 91 percent of phones judged by the system as sounding perfectly L2 were

judged by the human native listener as indeed being free of an L1 accent. Overall, the native listener agreed with the system 88 percent of the time.

This experiment was repeated using Japanese for L1 and American English for L2. Out of a total of 391 non-native phones, a native L2 listener agreed that (a) for 81 percent of the time, phones flagged by the system as being influenced by L1 phone X were in fact influenced by phone X, and (b) for 95 percent of the time, phones declared by the system as accent-free did in fact sound perfectly native. Overall, the native listener agreed with the system 89 percent of the time.

These results suggest that our system is a useful component technology for foreign language pronunciation teaching. Future work includes determining the learner's mispronunciation habits by identifying tendencies in his incorrect interlanguage allophones (for instance, if a learner tends to palatalize, his habit will appear as his preference towards palatal sounds).

A second experiment studied the distribution of e_phones and j_phones. 9 native speakers of English (5 male, 4 female) and 23 non-native speakers (20 male, 3 female, all native speakers of Japanese) each read 10 English sentences. The sentences are listed in table 6.4. Table 6.5 shows the ratio of English and Japanese phones detected in the utterances. Results show that the detection rate of English phones is significantly higher among native speakers of English, as expected.

There were no significant differences between e_phones and j_phones according to interlanguage phoneme type and speaker group. The sole exception was [l], where natives and non-natives were clearly distinguishable. Table 6.6 shows results of [l] and [r].

However, [r], the companion to [l], did not show significant differences across speaker groups. The reason for this for English speakers may be because syllable-final [r] can be shortened to become part of vowels while [l] cannot; this is reflected in the choice of monophones related to [r] ([r], [axr] and [er]) and [l] ([l] only). There may be a similar reason for Japanese speakers. In Japanese-accented English, [l] is pronounced but [r] is often deleted: consider that "call" becomes epenthesized in Japanese ([kooru]) while "car" does not ([kaa]).

Thus the fact that [l] is a strong indicator of non-nativeness is probably because [l] is often replaced by the Japanese flap, while [r] is often reduced and subsequently subsumed by rhotacized vowels.

6.5. RANKING PRONUNCIATION ERRORS

6.5.1. SEGMENTAL NON-NATIVENESS

As we saw in section 6.4, it is possible to automatically measure the pronunciation quality of phones produced by non-native talkers by using a speech recognizer incorporating native phone models of both Japanese and American English. We used HMMs for Japanese and English that were trained separately on language-dependent native-speaker speech data but were bundled to-

gether during recognition. Doing so removed the speaker's physiological aspects from his speech while retaining the systematic differences in phonetics and phonology between the two languages. The system gave categorical results — for instance, "your English phone X sounded like Japanese phone Y" — which are pedagogically useful because we knew how to instruct learners who say phone Y instead of X. Categorical results do not necessarily rank pronunciation errors in order of severity, however. This is important because it guides learners towards rectifying the worst errors first.

This section proposes a method to quantify the degree of a phone's non-nativeness based on how it is falsely detected in native speech. We used this measure to implement a pronunciation-training system with adjustable sensitivity. When the system looks only for strong indicators of non-nativeness, learners can focus on gross errors. When the system looks for both strong and weak indicators of non-nativeness, learners can improve upon minor mispronunciations as well.

To show that the technique we proposed in section 6.4 is applicable to teaching languages other than English, this section describes a system for teaching the phone quality for Japanese to native speakers of American English.

6.5.2. PRONUNCIATION ERROR SORTER

Before measuring the non-nativeness of Japanese phones, we first improved the baseline performance of our previous system by taking the following steps: (a) use gender-specific acoustic models in the speech recognizer, (b) use only clean speech input via a high-quality soundcard, and (c) in cases where the correct Japanese phone and its incorrect English replacement are so acoustically similar that no pedagogical advantage is gained from discriminating them, remove the English phone.

After augmenting the system's baseline performance, we studied the distribution of English phones confused with Japanese phones. On the one hand, some English phones are occasionally falsely detected in Japanese native speech. Such English phones are weak indicators of non-nativeness because such phones naturally occur in Japanese native speech, albeit relatively infrequently. On the other hand, some English phones occur rarely in Japanese native speech. These English phones are strong indicators of non-nativeness because they are conspicuous.

We studied the relative frequency of English phones in Japanese native speech in the speech of 30 Japanese natives each reading 59 Japanese sentences containing 711 phones. Grading the speech by the system showed that the type of false detections differed according to speaker. For instance, one speaker's English phone [a] was misrecognized as the Japanese phone [aa] phone 5.7 percent of the time, while the [a] for a different speaker was misrecognized as [ao] at 5.7 percent.

We experimented with two kinds of thresholds to allow for English phones being detected. The first type of threshold was based on the total number of Japanese phones in the speech material:

THRESHOLD_ONE = (number of possible Japanese phones) x CONSTANT_ONE

For instance, the test material contained 105 occurrences of the phone [a]. Setting the value of CONSTANT_ONE to 0.2 allows up to 21 phones to be recognized as English phones.

The second type of threshold was based on the maximum number of each English phone falsely recognized in the speech of multiple Japanese natives:

THRESHOLD_TWO = (maximum number of English phones falsely recognized in Japanese-native speech) x CONSTANT_TWO

For instance, among 30 native speakers in the training set, the Japanese phone [ay] occurred most frequently in one speaker whose speech yielded [ay] 7 times. Setting the value of CONSTANT_TWO to 2 allows up to 14 occurrences of [ay].

6.5.3. EVALUATION EXPERIMENT

The two thresholds were evaluated on the speech of 7 Japanese natives (the "native test set"), 2 Japanese semi-natives with over 10 years of Japanese experience (the "semi-native test set"), and 7 non-natives with less than 1 year of Japanese language training (the "non-native test set"). Results are shown in table 6.7.

Setting CONSTANT_ONE to 0.2 removed most false detections in the native test set. False detections were significantly reduced in the semi-native test set (from 11.9 to 2.0 percent), and somewhat reduced in the non-native test set (from 22.3 to 8.5 percent). Setting CONSTANT_TWO to 2 removed all false detections in the native test set. False detections were moderately reduced in the semi-native test set (from 11.9 to 5.1 percent), and somewhat reduced in the non-native test set (from 22.3 to 8.2 percent).

These results show that the two thresholds correctly (a) treat native speakers as natives, (b) identify semi-natives as close-to-natives, and (c) point out errors among non-natives.

The two thresholds were added to our pronunciation-training system with learner-adjustable controls. When a CONSTANT is set high, the learners can focus on gross errors. When a CONSTANT is set low, learners can improve upon minor mispronunciations as well.

6.6. DISCUSSION

Automated learning systems were developed for teaching the pronunciation of phone quality (the spectral characteristics of each phone) to entry-level non-native adult learners of Japanese or En-

glish. Most non-natives speak languages whose phonetic inventories do not include all of the phones found in the target language. The system asks the learner to read words, phrases or sentences. Speech recognition technology is used to determine whether the learner produced phones correctly in the target language, substituted it with a phone from his native language, deleted the phone, or inserted extraneous phones. The system informs the learner what mistakes were made and advises how to pronounce more correctly. Learners can adjust the sensitivity of non-nativeness so that learners can focus on gross errors, or improve upon minor mispronunciations as well.

The proposed methods remove individual physiological aspects from the learner's speech by using a speaker-independent bilingual phone recognizer. Two key technologies were involved: (1) using speech recognition to identify phones that are deleted, inserted, or substituted for L2 phones, and (2) ranking mispronunciations in order of non-nativeness. The know-how incorporated into the system might be applied to various areas, including (1) improving acoustic distance measurements for speaker adaptation, (2) improving speech recognition performance by subdividing the user population according to their dialect, and (3) improving spoken language system user-interfaces by identifying the user's native language.

Table 6.1.
Words and phrases trained by the anaptyctic vowel detection system.

Words

1.1. touch
1.2. but
1.3. class
1.4. drama
1.5. extra
1.6. train
1.7. lunch
1.8. please
1.9. jinx

Short phrases and word sets

2.1. Please pay promptly.
2.2. lake, lakes, flakes
2.3. fight, jinx, fast
2.4. The trains were filmed in the Alps.
2.5. Extra drama class, please.

Loan words in carrier phrases (note: only words in [brackets] are graded)

3.1. I have an [atlas] and [album] at home.
3.2. I have a [grapefruit] and [salad dressing] at home.
3.3. I have a [hard boiled egg] and [yoghurt] at home.
3.4. I have [cold cream] and [shaving cream] at home.
3.5. I have an [evening dress] and [turtle neck sweater] at home.
3.6. I have a [crossword puzzle] and [emerald] at home.
3.7. I have a [clarinet] and [flute] at home.
3.8. I have a [trenchcoat] and [knapsack] at home.
3.9. I have a [part-time maid] and [butterfly] at home.

Table 6.2.
Anaptyctic vowel detection accuracy for each word or phrase. Average percentages shown for each word or phrase. n=19.

1.1.	93.8
1.2.	100.0
1.3.	87.5
1.4.	87.5
1.5.	85.4
1.6.	87.5
1.7.	100.0
1.8.	93.8
1.9.	84.4
2.1.	80.0
2.2.	87.5
2.3.	95.0
2.4.	91.4
2.5.	89.8
3.1.	70.3
3.2.	86.6
3.3.	84.4
3.4.	82.1
3.5.	84.8
3.6.	83.0
3.7.	89.0
3.8.	93.8
3.9.	95.3
ALL	88.4

Table 6.3.
Phones used in the bilingual HMM set.

English phones (45 phones).

aa ae ah ao aw ax axr ay b ch d dh eh el em en er ey f g hh ih ix iy jh k l m n ng ow oy p r s sh t th uh uw v w y z zh

Japanese phones (40 phones).

sp N a a: b by ch d e e: f g gy h hy i i: j k ky m my n ny o o: p py q r ry s sh t ts u u: w y z

Table 6.4.

List of sentences read by the subjects.

1. Good morning.
2. Do you have a car?
3. Can you speak Japanese?
4. My girlfriend likes red flowers.
5. School starts the first week of April.
6. Look at the pretty fish in the warm water.
7. There are many birds and animals in Australia.
8. Australia is a big country without large rivers.
9. Her mother lives near School Drive and Garden Street.
10. The library is next to the nice Japanese garden in the park.

Table 6.5.
Distribution of all English and Japanese phones detected according to the subjects' nativeness.

nativeness	gender	e_phone%	j_phone%	total phones
native	male	39.3	60.7	1331
native	female	43.8	56.3	1072
non-native	male	30.4	69.6	5360
non-native	female	35.7	64.3	804

Table 6.6.
Distribution of English and Japanese phones for the target English phones [l] and [r] according to the subjects' nativeness and gender.

nativeness	gender	phone	e_phone%	j_phone%	total phones
native	male	l	64.6	35.4	65
native	female	l	48.1	51.9	52
non-native	male	l	18.9	81.2	260
non-native	female	l	20.5	79.5	39
native	male	r	58.8	42.2	154
native	female	r	41.9	58.1	124
non-native	male	r	53.6	46.5	620
non-native	female	r	49.5	50.5	93

Table 6.7.
False detections depending on use of thresholds.

speaker	no threshold	CONSTANT _ONE=0.2	CONSTANT _TWO=2.0
native 1	10.69	0.46	0.00
native 2	9.06	0.12	0.00
native 3	12.08	0.32	0.00
native 4	11.15	0.00	0.00
native 5	7.43	0.00	0.00
native 6	14.52	0.00	0.00
native 7	13.59	0.25	0.00
native mean	11.22	0.16	0.00
semi-native 1	10.60	1.76	5.34
semi-native 2	13.12	2.21	4.76
semi-native mean	11.86	1.99	5.05
non-native 1	14.05	1.16	4.65
non-native 2	20.21	8.13	8.94
non-native 3	20.56	8.36	8.13
non-native 4	14.29	4.30	1.65
non-native 5	30.66	16.14	19.40
non-native 6	39.37	13.47	7.55
non-native 7	17.07	7.67	6.74
non-native mean	22.32	8.46	8.15

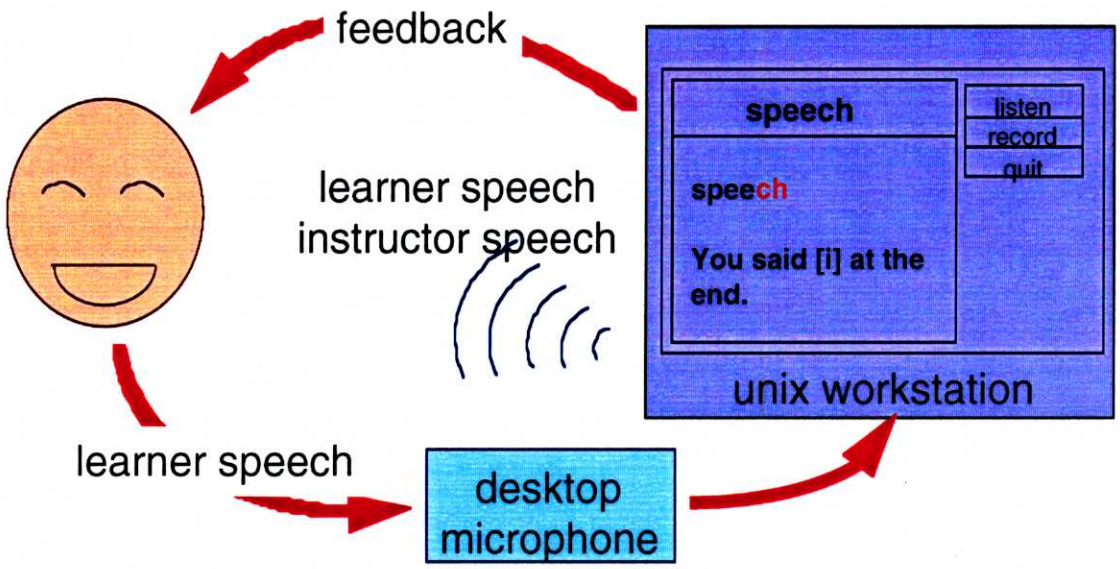


Figure 6.1.

System-user interaction of anaptyctic vowel detection system. System points out where vowels were inserted. The learner's task is to remove such vowels. This figure shows the screen in English; the actual system uses both Japanese and English, and with more detailed feedback (see figure 6.3 for a screen image).

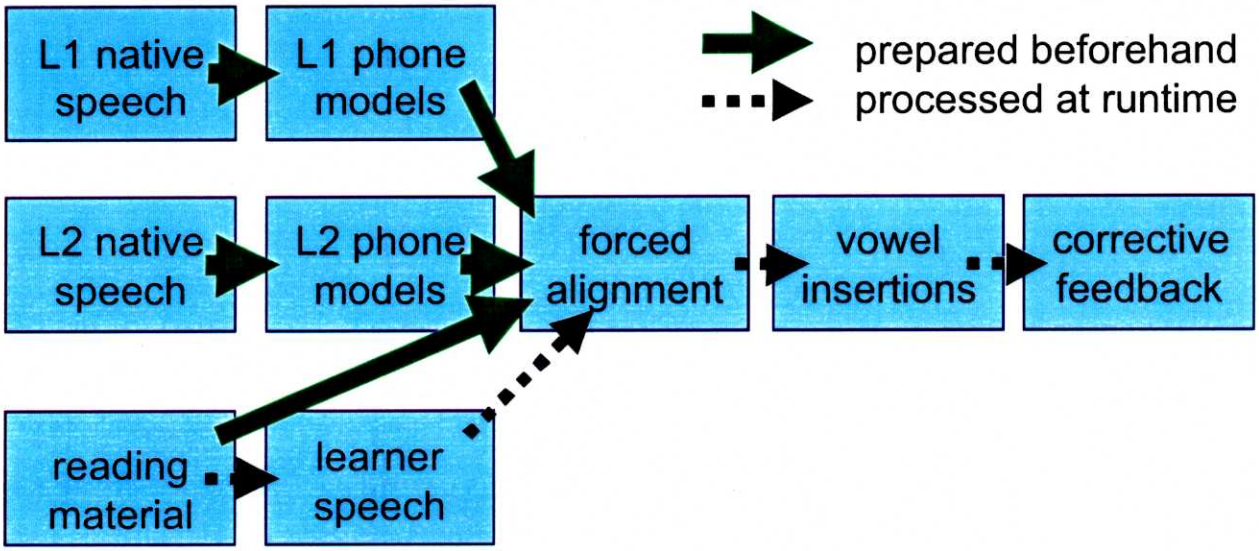
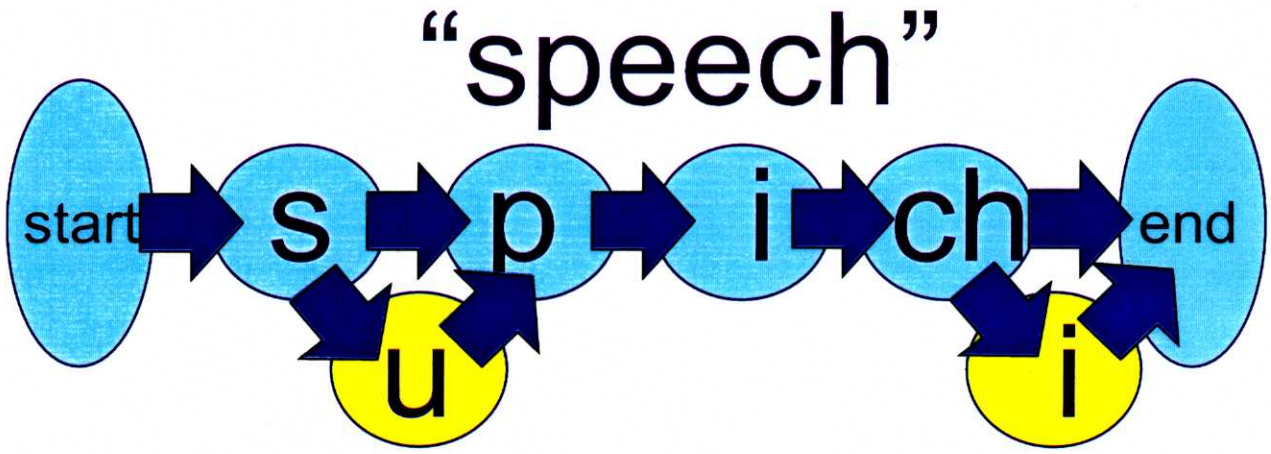


Figure 6.2.

Anaptyctic vowel detector process flow. L2 (Japanese) phone models are used only for detecting vowel insertions. All obligatory phones use English phone models only.

Figure 6.3.

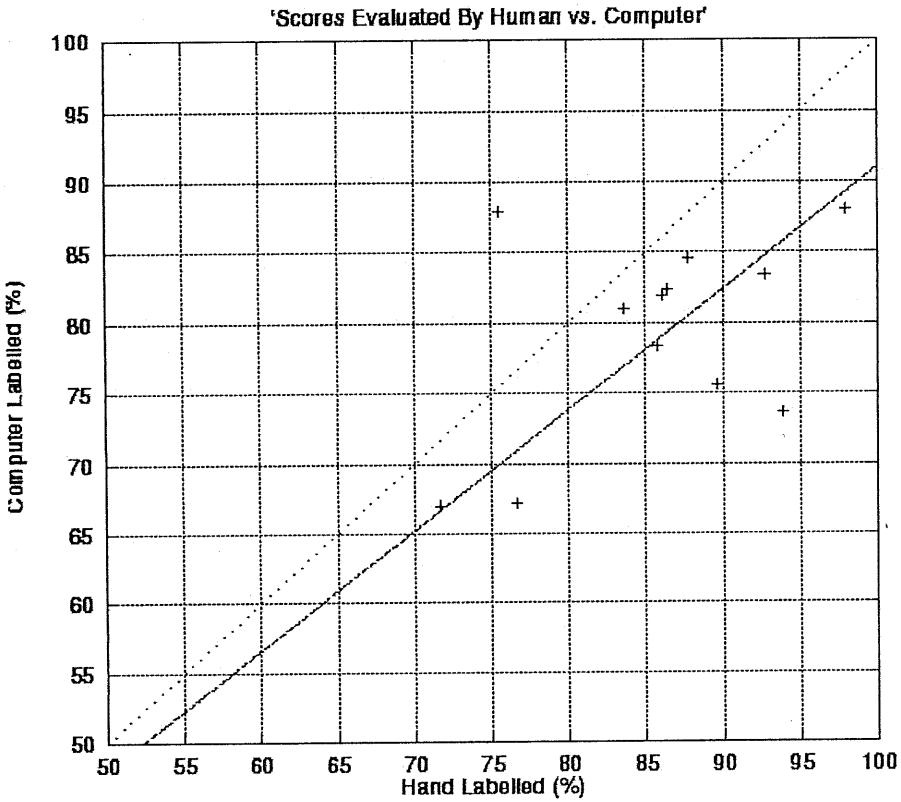
Anaptyctic vowel detector screen example. The system prompted the learner to say "I have a trenchcoat and knapsack at home." The words "trenchcoat" and "knapsack" are the words being graded; both exist as loan words in Japanese, and learners often make mistakes reading them because the loans' pronunciations carry back over to English. The system has detected Japanese vowels [o] after the first and last "t" in "trenchcoat", plus the vowel [u] after "p" and "ck" in "knapsack". The learner's task is to repeat the sentence trying not to insert vowels.



(a) Phone network for the word "speech."

(b) Phone lattice corresponding to figure (a).

Figure 6.4.
Phone network for anaptyctic vowel detection. Figure (a) shows the network for the word "speech", where the correct (and obligatory) phones are shown in the main path, and the incorrect (and optional) phones are shown as alternate paths. Figure (b) shows the equivalent phone lattice for figure (a). Anaptyctic vowels are paired with null phones, the latter corresponding to the correct path.



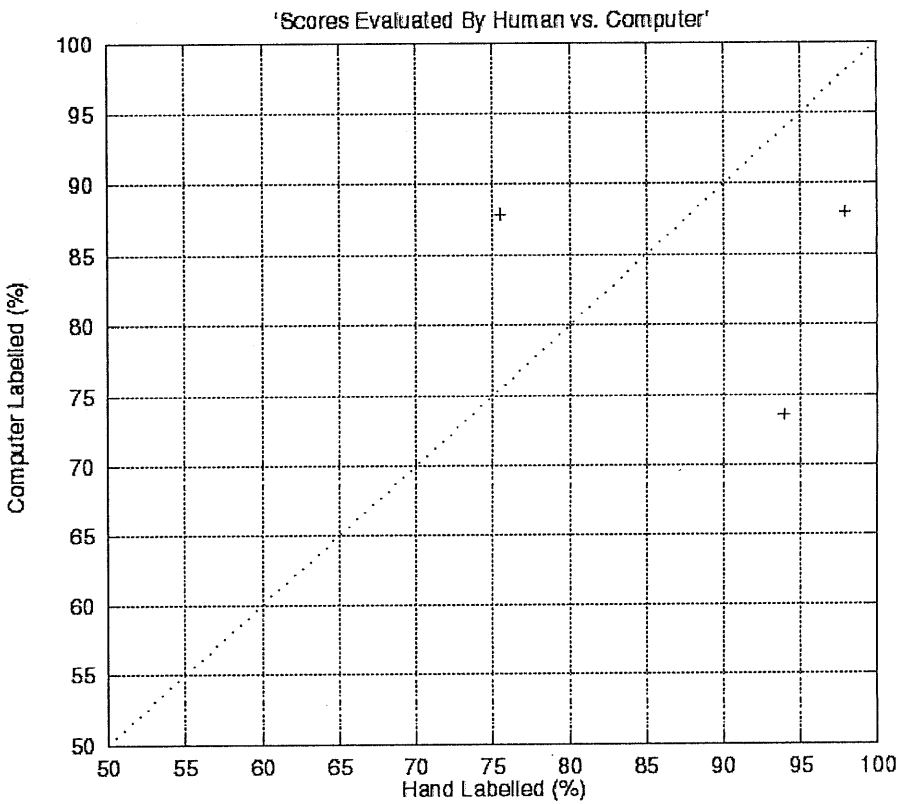


Figure 6.5.

Results of comparisons between system-detected anaptyctic vowels and hand-labeled data. The left chart (chart A) shows scatter plots for speech data collected using a close-talking microphone. The right chart (chart B) is data recorded on a desktop microphone. Correlation for chart A is 0.99. The regression line for chart A is slightly below y=x; this means the system gave lower scores because it detected more vowel insertions. Chart B suggests that recognition is inaccurate in noisy environments, which may have implications for when learners wish to study together in groups.

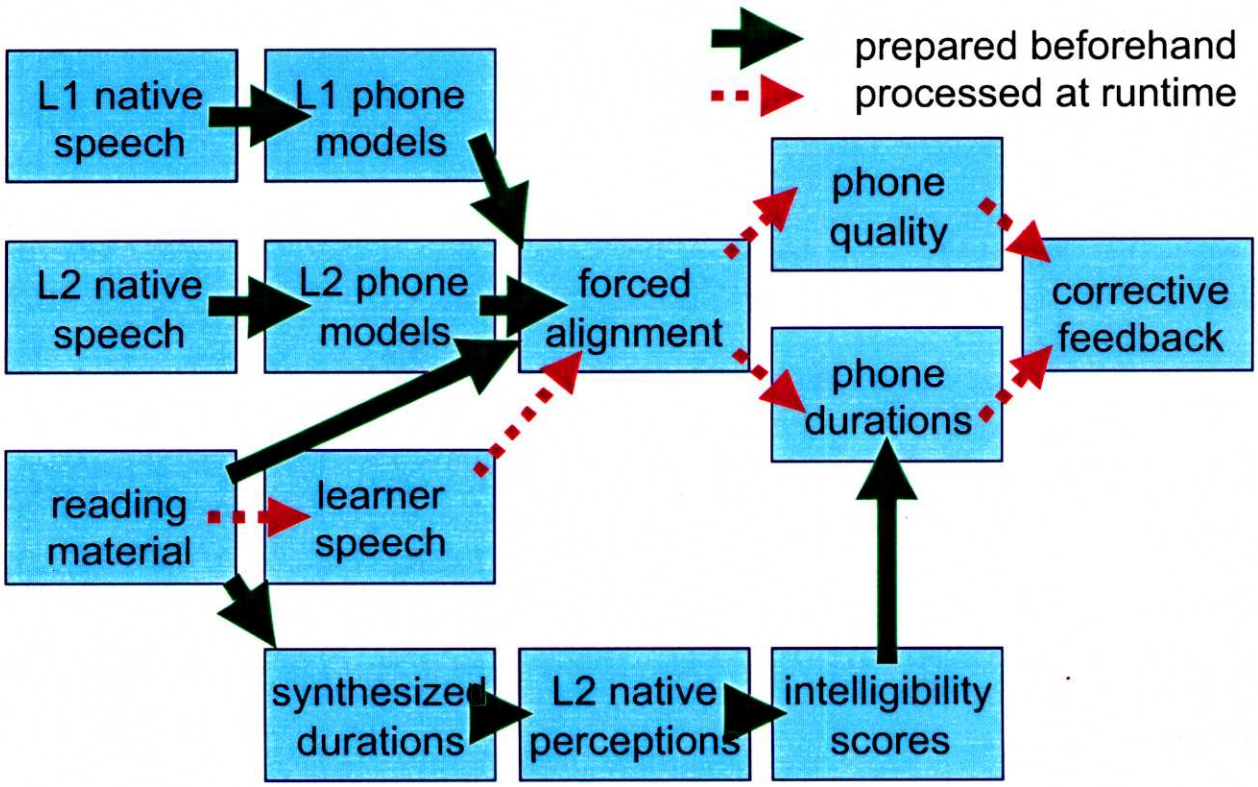


Figure 6.6.

Process flow of hybrid phone-duration and phone-quality system. Phone quality and duration are measured in parallel.

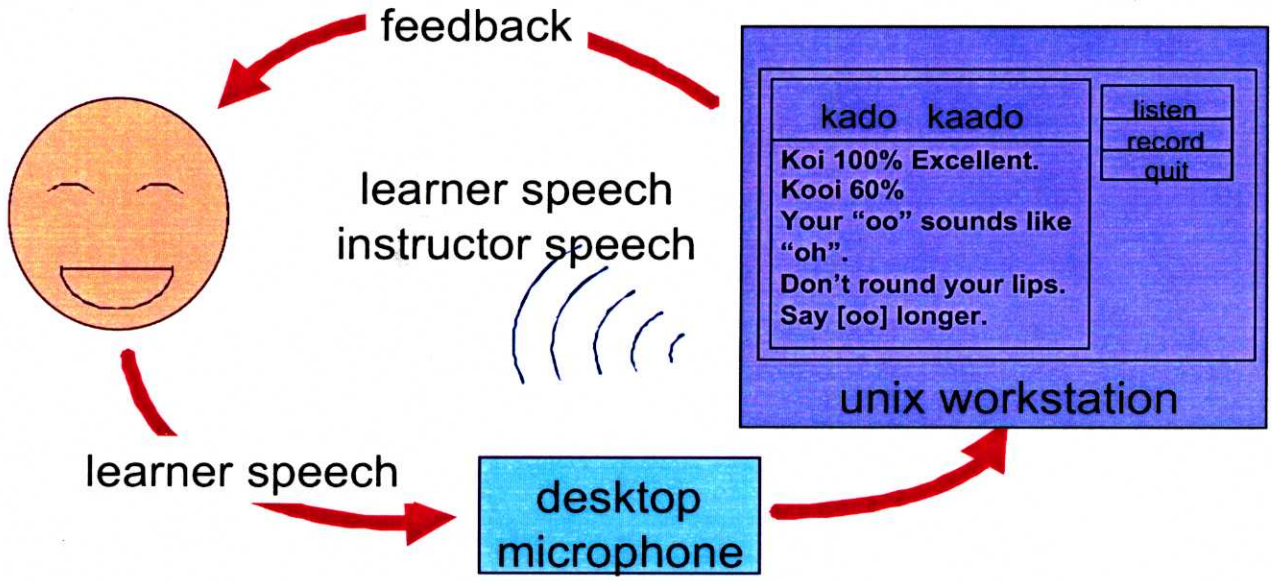


Figure 6.7.

System-user interaction of vowel-substitution detector. In addition to phone duration information, the learner receives categorical articulatory advice on vowel quality. This figure shows the screen in English only. The actual system uses both kana and English, and with more detailed feedback.

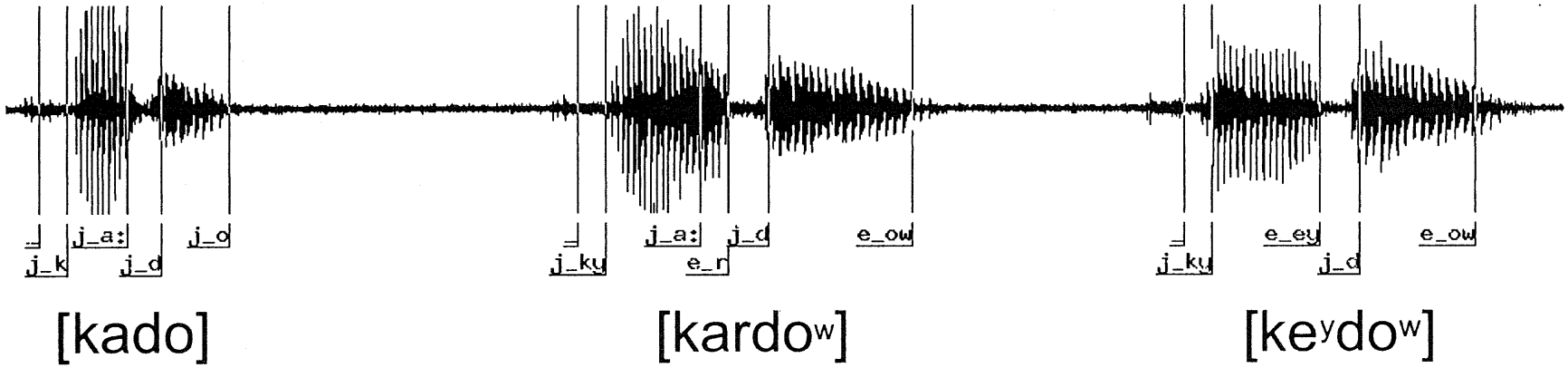


Figure 6.8.

Recognition results of three renditions of "kado." The vowels shown range from correct Japanese [a] and [o] (shown in waveform segment (a)) through American-English-accented Japanese [ar] and [ow] (waveform segment (b)) to literal reading of the word string "kado" by and American English speaker with no knowledge of Japanese pronunciation or romanization (waveform segment (c)). Phone labels prefixed with "j_" denote Japanese phones, and phones prefixed with "e_" denote American English phones.

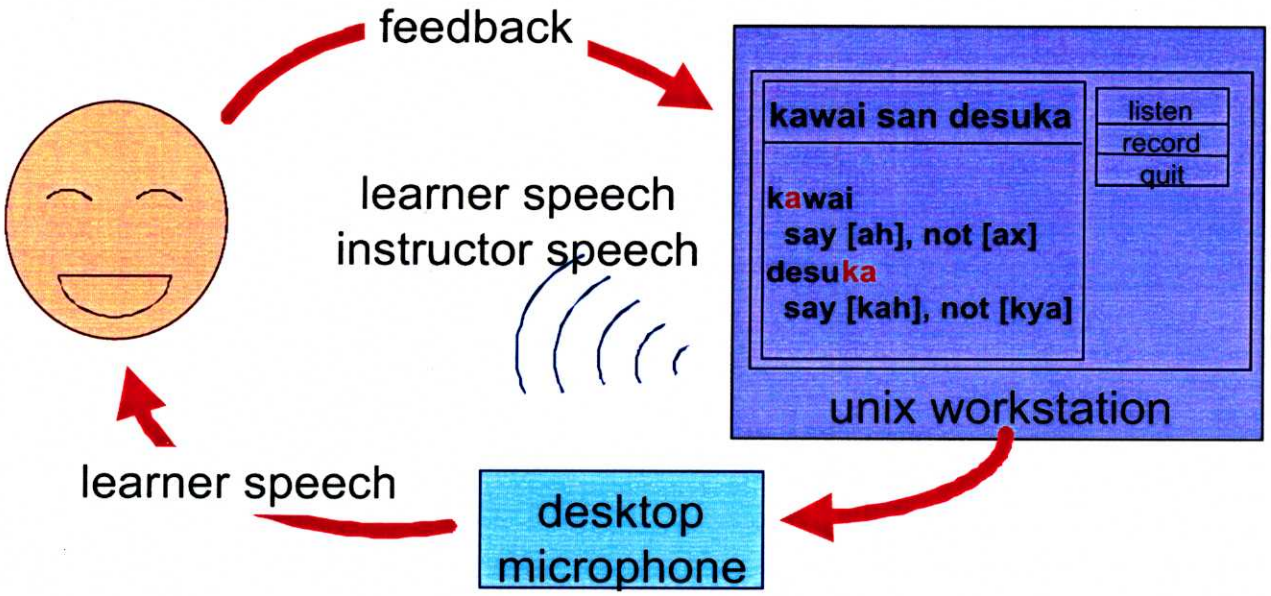


Figure 6.9.

System-user interaction of phone-quality system. Learner receives categorical articulatory advice on phone quality. This figure shows the screen in English only. The actual system uses both kana and English, and with more detailed feedback.

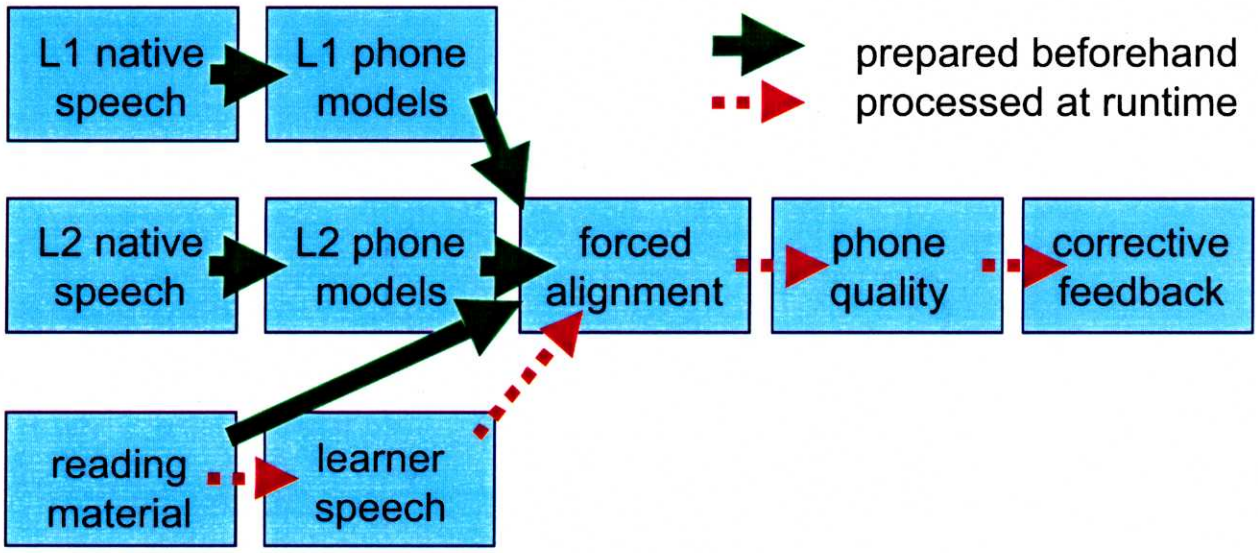


Figure 6.10.

Process flow of phone-quality system. The phone-quality system uses a bilingual recognizer to give the learner categorical articulatory advice on phone quality.

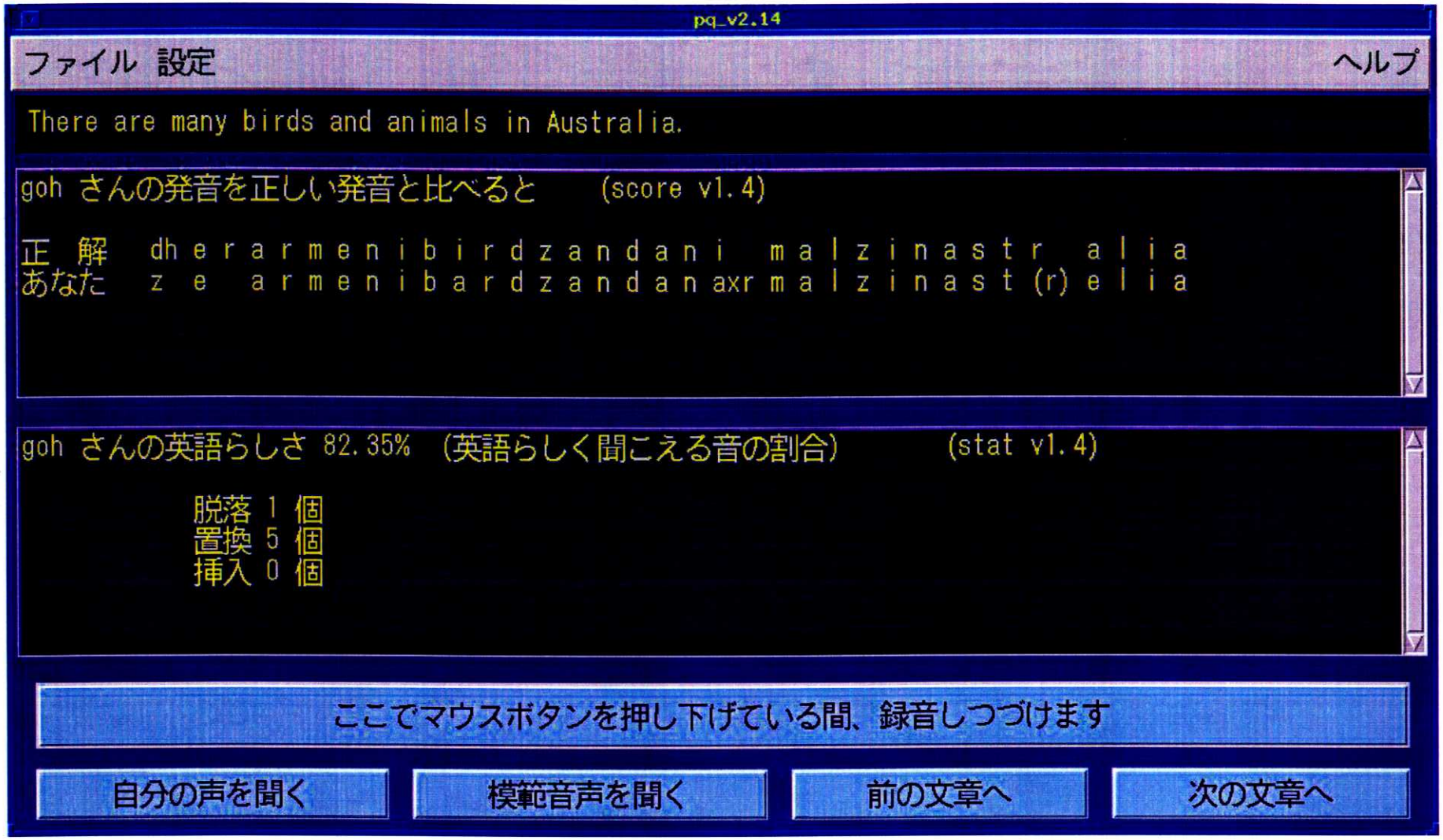


Figure 6.11.

Segment classifier screen example. The system asks the learner to say "There are many birds and animals in Australia." Starting at the left of the utterance, we find the system detected a phone substitution ([dh] was pronounced as [z]), a phone deletion ([r] was dropped), and more substitutions ([i] in "birds" became [a], [i] in "animals" became [axr], and [r] in "Australia" was pronounced with a Japanese [r]). No insertions were found in this utterance.

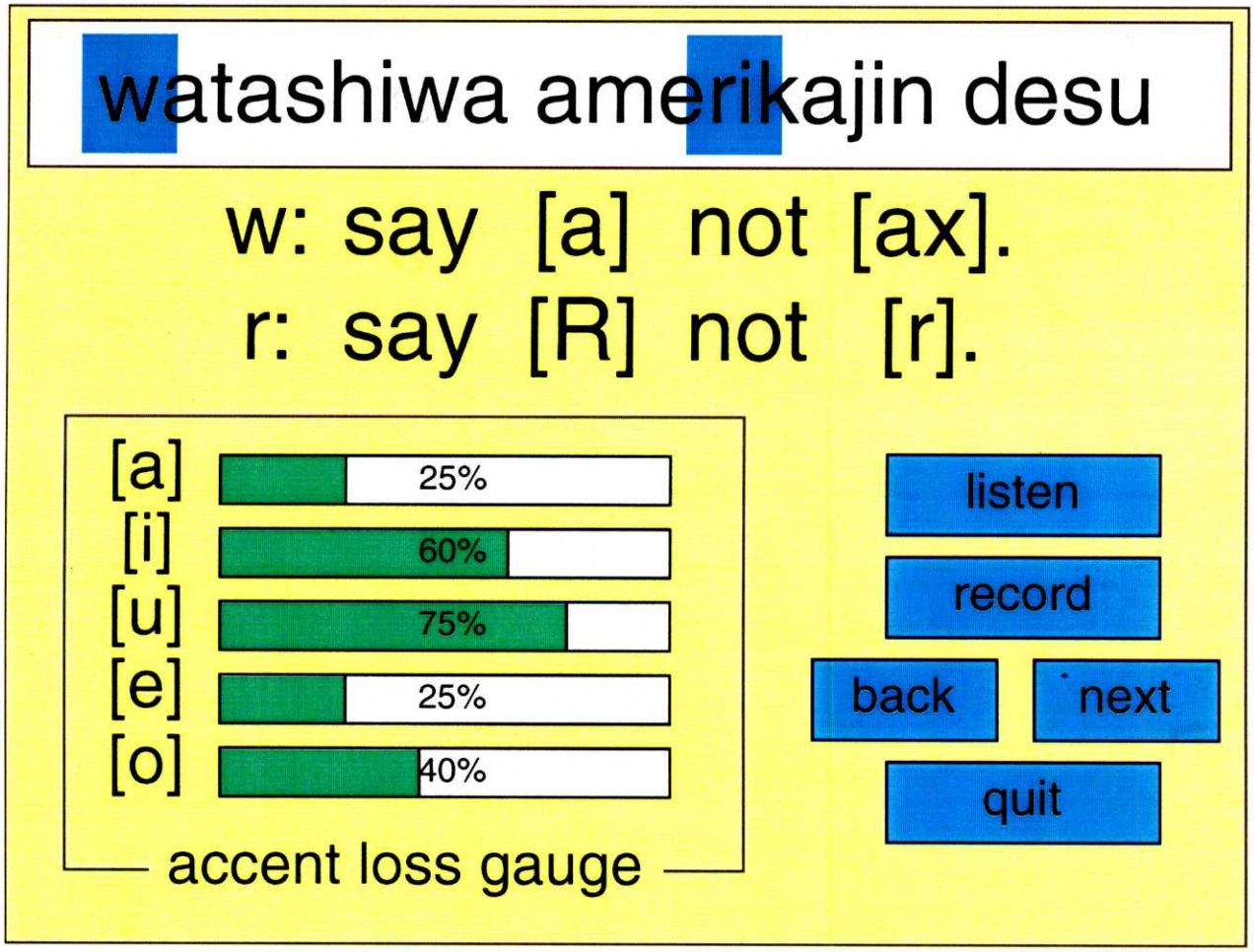


Figure 6.12.

Accent loss gauge screen example. The learner receives categorical articulatory advice on phone quality for each mispronounced phone, and an "accent loss gauge" indicating allophone reduction ratios. For explanation purposes, the figure is in English and shows only vowels.

7. CONCLUSION

This dissertation advances the state of the art by combining existing technologies with newly created knowledge. The research tasks provide multiple viewpoints towards the larger issue of automated pronunciation learning. CALL systems were built to teach three pronunciation skills: phone duration, pitch, and phone quality. This section lists our major findings and directions for future work.

7.1. FINDINGS

The system for teaching phone duration provides corrective feedback similar to human teachers by measuring phone duration using speech recognition technology. We learned that (1) learners adjust their speaking rate to the system's prompts, (2) native speakers distinguish long and short tokushuhaku clearly, (3) instructing learners to say a phone longer or shorter is easy to understand, and (4) learners rapidly acquire skills.

The system for teaching phone quality provides corrective feedback while disregarding individual physiological aspects of the learner's speech by measuring L1 accents by comparing L1 and L2 phones using a speaker-independent bilingual phone recognizer. We learned that (1) the number of phone types in the learner's interlanguage allophone inventory decreases with pronunciation practice, (2) the size of the learner's interlanguage allophone inventory is one measure of the learner's level of foreign accent, (3) the frequency distribution of interlanguage allophones being selected as the L2 phone is another measure of the learner's level of foreign accent, and (4) the system helps learners rapidly acquire phone quality skills.

The system for teaching pitch contours automatically grades prosody by measuring pitch patterns and intonation contours using speech recognition technology and prosodic analysis. We learned that (1) native speakers clearly distinguish pitch changes at syllable boundaries, (2) instructing learners to say a syllable higher or lower is easy to understand, and (4) learners rapidly acquire skills.

7.2. FUTURE WORK

Teaching phone duration can be improved by gradually lowering intelligibility scores as the durations of long phonemes become overly long. This study ignored this aspect because we found that learners do not elongate long phones to the extent of sounding unnatural or incorrect (undoubtedly because learners wish to comply with target language norms). However at least in theory such erratic behavior is possible. Perception experiments using synthetized long phones can determine the goodness of phones.

A related issue involves multimorphome superlong tokushuhaku, such as the 6-mora-long [o] in "denwa bangoo o oo ojini osieru" (give the telephone number to my great uncle). Superlong tokushuhaku occur across multiple morphemes. Morpheme boundaries are perceived by pitch accent patterns. Superimposing pitch accent on phone duration instruction may help learners acquire skills in this area.

The pitch contour system could be improved by teaching pitch contours of phrases and sentences with more than one acceptable intonation contour. Here again the concept of quantitative measurements of intelligibility rears its head. Manageable tasks might include categorizing various intonation contours into various meanings — for instance "kawai san desuka" with a falling tone means "So you are Mr. Kawai," (indicates confirmation of information) while a rising tone means "Are you sure you are Mr. Kawai?" (indicates incredibility). Simple tasks such as these can be based on perception experiments. More advanced tasks might include intonation as discourse structure markers — for example utterance-final intonation patterns functioning as cues for turn-taking behavior.

Teaching phone quality might be significantly improved by quantitatively measuring the intelligibility of phones as a function of phone quality. On the one hand, we could perhaps extrapolate intelligibility from rank-order or qualitative judgements of native speakers. On the other hand the correlation among native speakers regarding pronunciation has never been high; in fact this has been the limiting factor for many pronunciation grading systems. Thus unfortunately the goodness of phone quality may never be quantifiable.

A fundamental issue that this dissertation has not touched upon is the training of hearing skills. Spoken language competence necessarily requires both production and reception skills. This dissertation has not dealt with teaching speech perception to non-native learners primarily because hearing skills have been the focus of CALL research up to this point; this paper attemps to balance the scale of research efforts by proposing learning systems for speech production. Certainly this paper does not intend to minimize the importance of speech perception training. In fact the systems proposed in this paper would be meaningless without matching teaching in listening comprehension. An example of this would be tokushuhaku training, where learners must learn how to produce long and short vowel lengths correctly, but must also learn how to tell them apart in connected native speech because where localized fluctuations in speech rate can obliterate clear durational distinctions.

8. REFERENCES

[1] Auralog "Auralang version 2.50.3." Auralog, 1995

[2] Bongaerts, T. et al "Age and ultimate attainment in the pronunciation of a foreign language." Studies in Second Language Acquisition, vol 19, no 4, pp 447-465, 1997

[3] Ehsani, F. et al "Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New Call Paradigm." Language Learning & Technology, vol 2, no 1, pp 45-60, 1998

[4] Ehsani, F. et al "Subarashii: Japanese interactive spoken language education." Proc. Eurospeech-1997 (Rhodes, Greece), pp 681-684, 1997

[5] Enomoto, Y. "Phonetic rules for foreign loan words in Japanese." Nagoya Junior College Journal, vol 23, pp 93-100, 1985

[6] Fujisaki, H. et al "Analysis of voice fundamental frequency contours for declarative sentences of Japanese." Journal of the Acoustical Society of Japan (English), vol 5, no 4, pp 233-242, 1984

[7] Hibiya, J. "Pronunciation education outside of Japan." Bulletin of the Phonetic Society of Japan, vol 211, pp 43-48, 1996

[8] Hiller, S. et al "An automated system for computer-aided pronunciation learning." Computer Assisted Language Learning, vol 7, no 1, pp 51-63, 1994

[9] Hiller, S. et al "SPELL: an automated system for computer-aided pronunciation teaching." Proc. Eurospeech-1993 (Berlin, Germany), pp 1343-1346, 1993

[10] Hirose, K. et al "Analysis and synthesis of voice fundamental frequency contours of spoken sentences." Proc. ICASSP-1982 (Paris, France), pp 950-953, 1982

[11] Hirose, K. et al "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features." Proc. ICSLP-1996 (Philadelphia, U.S.A.) pp 378-381, 1996

[12] Imaizumi, H. et al "Realtime extraction of pitch and formants using DSPs and its applications to pronunciation training." Technical Report of the IEICE, SP89-36, 1989

[13] Ishii, E. "Japanese language education in Japan: issues and possible improvement." J. of Japanese Language Teaching, vol 94, pp 2-12, 1997

[14] Kindaichi, H. et al eds "Meekai Japanese accent dictionary, 2nd ed." Tokyo: Sanseedoo, 1981

[15] Kobayashi, Y. "Gemination in loans from English to Japanese." Studies in English Language and Literature, vol 1, pp 97-114, 1992

[16] Leggetter, C. et al "Speaker adaptations of HMMs using linear regression." Report CUED/F-INFENG/TR.181, Cambridge University, 1994

[17] Minematsu, N. "A study on the human process of speech perception." Ph.D. dissertation, University of Tokyo, 1994

[18] Mizutani, O. "Executive summary." in "Japanese speech and Japanese education." Tokyo: Ministry of Education, pp 1-4, 1993

[19] Nihon Hoosoo Kyookai ed "Japanese pronunciation accent dictionary, revised new edition." Tokyo: Nihon Hoosoo Publishing Association, 1985

[20] Otake, T. "Consonant gemination of English loanwords in Japanese." International Budo University Journal, vol 5, pp 101-116, 1989

[21] Ramalakshmi, V. et al "Panel discussion: TJSL pronunciation education abroad." In "The Japanese language in international society," Ministry of Education, pp 146-163, 1997

[22] Ronen, O. et al "Automatic detection of mispronunciation for language instruction." Proc. Eurospeech-1997 (Rhodes, Greece), pp 649-652, 1997

[23] Rooney, E. et al "Training consonants in a computer-aided system for pronunciation teaching." Proc. Eurospeech (Berlin, Germany), pp 561-564, 1993

[24] Saida, I. et al "A spoken language teaching system for nonnative learners of Japanese" in "A study on prosodic features in spoken Japanese and its implications to education." Tokyo: Ministry of Education, pp 137-140, 1991

[25] Saida, I. et al "Speech CAI for non-native speakers of Japanese." In "Annual report of the study group on Japanese prosody and its instruction", Ministry of Education, pp 5-15, 1993

[26] Sakata, M. "Analysis and synthesis of prosodic features in spoken dialogue of Japanese." Unpublished master's thesis: University of Tokyo, 1995

[27] Syracuse Language Systems "Triple Play Plus! Japanese version J1.2." Syracuse Language Systems, 1996

[28] Takeda, K. et al "Common platform of Japanese large vocabulary continuous speech recognition research: construction of acoustic model." IPSJ SIG Notes, paper number 97-SLP-18-3, 1997

[29] Taniguchi, H. "Results of a survey on Japanese pronunciation teaching." In "Prosody and its role in TJSL", Ministry of Education, pp 17-21, 1991

[30] Toki, S. "Pronunciation teaching." In Teramura, H. ed "Japanese and Japanese language teaching" vol 13, Meijishoin: Tokyo, pp 111-138, 1989

[31] Toki, S. et al "Japanese exercises for non-native speakers vol 12: pronunciation and listening (with cassette tape)." Tokyo: Aratake Shuppan, 1988

[32] Tsuchida, O. "Automatic generation of Japanese-accented pronunciations of English words." Unpublished senior project, University of Tokyo, 1997

[33] Witt, S. et al "Language learning based on non-native speech recognition." Proc. Eurospeech-1997 (Rhodes, Greece), pp 633-636, 1997

[34] Woodford, P. "Language testing at ETS: its development and evaluation." J. of Japanese Language Teaching, vol 94, pp 160-170, 1997

[35] Woodland, P. et al "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task." Proc. ARPA CSR Workshop (Arden House), 1996

[36] Young, S. et al "The HTK Book for version 2.1." Cambridge University, 1997

APPENDIX A SPEECH DATA COLLECTION

Material used for collecting speech data from native speakers of American English is included here for reference purposes. Native speakers of Japanese used a similar set of materials that was produced in Japanese kana-kanji.

Experiment for recording Japanese words read aloud by native speakers of English

1. Overview of Experiment

You are asked to participate in an experiment for recording Japanese words read aloud by native speakers of English. You will be given a list of Japanese words printed on a sheet. Prior to recording, you will check the list, and mark on the sheet whether you know the word or not. Then you will wear a microphone on your head, and read the words on the sheet aloud. The entire experiment is expected to last about 20 minutes, not including time waiting for your turn. You will receive 1,000 yen in cash as compensation for participation.

The recordings of your voice will be used solely for the purposes of academic research by the University of Tokyo and Tokai University. Your identity will neither be recorded, nor will it will be disclosed to any third party.

To participate, you must be an adult native speaker of English, and a beginning-level learner of Japanese as a foreign or second language. You must answer a list of questions describing your language background. The risk of discomfort, injury or death by participating in this experiment is negligible. You must sign an agreement that you agree to the purpose of the study.

2. Research Background

A group of researchers at the University of Tokyo are, with the cooperation of Tokai University, developing a computer system for teaching the pronunciation of nonnative languages to adult learners.

Our research addresses the automatic detection, measurement, and correction of nonnative pronunciation characteristics (so-called "foreign accents") in foreign language speech. Acquiring nativelike pronunciation ranks first in desirability among foreign language learners.

Computerized self-learning systems can be used to handle the repetitive tasks of pronunciation teaching. Our systems automatically measure the pronunciation quality of speech produced by nonnative learners. The learner receives detailed feedback on what his or her mistakes are, an estimate of the percentage of native speakers who will understand the learner's utterance the way it was pronounced, and recommendations to improve his or her pronunciation skills.

Your speech will be used to study how beginning learners of Japanese as a foreign language mispronounce familiar or unfamiliar words. The objective of this experiment is to learn what kind of mistakes are made, how often they occur, and whether current speech recognition technology can detect such mistakes. Your speech data will help advance the state of the art of computer-aided pronunciation training.

3. Experiment Instructions

You will first read this set of instructions, sign the agreement form (a copy will be given to you), and answer a list of questions describing your language background. Some questions may appear unusual but are necessary for research purposes (the reasons are explained on the sheet). If you feel uncomfortable responding to any of the questions, leave the answer blank.

You will be given a list of Japanese words printed in hiragana, along with their English translations. These are the words you will read during the recording session. Check the list, and circle *Yes* or *No* depending on whether you know the word or not. Remember, this is not a test. There are many words in the list that even advanced learners are unlikely to know.

Next, we will prepare you for recording. You will wear a head-mounted microphone, similar in shape to headsets worn by aircraft pilots and motorcycle peace officers. This allows high-quality recordings because your voice is clearly captured and ambient noise is minimized. We will adjust our recording equipment for optimal recording quality.

You will then start reading words aloud. You may repeat reading words as many times as you like. **If you cough, sneeze or laugh during talking, you must repeat saying the word.** We will use your last utterance of each word.

There is no time limit as far as we are concerned. You may spend as much time with us as you like. Based on past experience with other participants, we expect that you will finish in about 20 minutes, not including time waiting for your turn.

A researcher or technician will assist you during the entire experiment. If you have questions, please ask. **One question we cannot answer is the correct pronunciation of the reading material in the list.** Your mispronunciations are important data. We cannot contaminate the data by guiding you to correct or incorrect pronunciations. You must guess the best pronunciation possible and say it aloud. **There is no penalty for mispronunciations.**

4. Reading Material

A portion of the reading material is given below to serve as a sample of what you will read during the experiment.

ねる	knead	あおい	blue	おの	axe
いらい	after	あめ	rain	うる	obtain
あり	ant	いる	roast	うむ	give birth
おもい	heavy	にわ	garden	のみ	chisel

5. References

The following academic papers explain the technology and research issues in detail. Online versions are available via the World Wide Web.

Goh Kawai and Keikichi Hirose (1998a) *A CALL system using speech recognition to train the pronunciation of Japanese tokushuhaku.* Proceedings of STiLL (Speech Technology in Language Learning) (Marholmen, Sweden), pp. 73-76, May 1998. Online version at: http://www.gavo.t.u-tokyo.ac.jp/~kawai/goh/980525.pdf

Goh Kawai and Keikichi Hirose (1998b) *A CALL system for teaching the duration and phone quality of Japanese tokushuhaku.* Proceedings of the Joint Conference of the ICA (International Conference on Acoustics) and ASA (Acoustical Society of America) (Seattle, Washington), pp. 2981-2982, June 1998. Online version at: http://www.gavo.t.u-tokyo.ac.jp/~kawai/goh/980626.pdf

Goh Kawai and Keikichi Hirose (1998c) "A bilingual speech recognizer for detecting phone-level pronunciation errors in nonnative speech." Proceedings of the ASJ (Acoustical Society of Japan) 1998 Fall Conference (Yonezawa, Japan), paper number 1-2-24, September 1998. Online version at: http://www.gavo.t.u-tokyo.ac.jp/~kawai/goh/980924a.pdf

Agreement to Participate in Experiment

I, (print name) _____, am an adult native speaker of English, and a beginning-level learner of Japanese as a foreign or second language. I agree to participate in an experiment for recording Japanese words read aloud by native speakers of English. I will answer a list of questions describing my language background. I will be given a list of Japanese printed on a sheet, and mark whether I know the word or not. I will wear a microphone on my head, and read the Japanese words aloud.

The entire experiment is expected to last about 20 minutes, not including time waiting for my turn. I will receive 1,000 yen in cash as compensation for participation.

The purpose of the experiment has been explained to me. I have been told that the risk of discomfort, injury or death by participating in this experiment is negligible. The recordings of my voice will be used solely for the purposes of academic research by the University of Tokyo and Tokai University. My identity will neither be recorded, nor will it will be disclosed to any third party.

I have read and agree to the conditions above.

Signature _____ Date _____.

Contact Information

For questions and comments regarding this study, contact:

Goh Kawai
University of Tokyo
Department of Information and Communication Engineering
Professor Keikichi Hirose Laboratory
Tokyo 113-8656, Japan
tel: +81-3-5804-7459
fax: +81-3-3815-4442
email: goh@kawai.com
http://www.kawai.com/

PARTICIPANT COPY

Agreement to Participate in Experiment

I, (print name) _____, am an adult native speaker of English, and a beginning-level learner of Japanese as a foreign or second language. I agree to participate in an experiment for recording Japanese words read aloud by native speakers of English. I will answer a list of questions describing my language background. I will be given a list of Japanese printed on a sheet, and mark whether I know the word or not. I will wear a microphone on my head, and read the Japanese words aloud.

The entire experiment is expected to last about 20 minutes, not including time waiting for my turn. I will receive 1,000 yen in cash as compensation for participation.

The purpose of the experiment has been explained to me. I have been told that the risk of discomfort, injury or death by participating in this experiment is negligible. The recordings of my voice will be used solely for the purposes of academic research by the University of Tokyo and Tokai University. My identity will neither be recorded, nor will it will be disclosed to any third party.

I have read and agree to the conditions above.

Signature _____ Date _____.

Contact Information

For questions and comments regarding this study, contact:

Goh Kawai
University of Tokyo
Department of Information and Communication Engineering
Professor Keikichi Hirose Laboratory
Tokyo 113-8656, Japan
tel: +81-3-5804-7459
fax: +81-3-3815-4442
email: goh@kawai.com
http://www.kawai.com/

Your Language Background

Your identity will be kept separate from your responses to the following questions. If you don't feel like answering the question, leave the answer blank.

The following questions are about your language experience. Circle the answer that best describes you.

(1) What language(s) do you speak at home?

English Japanese Spanish

other (specify)_____.

(2) Where do you live?

continental U.S. Alaska Hawaii or Pacific

Canada Japan

other (specify)_____.

(3) Where did you grow up?

continental U.S. Alaska Hawaii or Pacific

Canada Japan

other (specify)_____.

(4) Are any of your family members (such as parents, siblings, spouse, significant other) native speakers of Japanese?

No Yes (specify whom)_____.

(5) Have you ever used Japanese to communicate on a daily basis with parents, siblings or other family members who are living with you or have lived with you for at least a year?

No Yes (specify whom)_____.

(6) How long have you been learning Japanese?

_____years _____months

(7) How many times have you visited Japan?

_____months

(8) How long have you been in Japan for this visit?

_____weeks

(9) What is the total duration of time you have spent in Japan?

_____months _____weeks

The following questions are about your gender, age and physical size. People's voice patterns differ depending on physical factors. Computers often process speech differently based on the person's physical characteristics. Circle the answer that best describes you.

(10) Are you male or female?

male female

(11) What is your age group?

below 15 15-20 21-25 26-30

31-35 36-40 41-45 46-50

51-55 over 56

(12) How tall are you?

less than 5'0" 5'0"-5'3" 5'4"-5'7" 5'8"-5'11"

6'0"-6'3" greater than 6'4"

Reading Material

We will now start recording. To make you feel accustomed to the recording task, we start with several English sentences. If you cough, sneeze or laugh during talking, you must repeat saying the sentence. **Pause for two seconds between sentences.**

(1) Good morning.

(2) Do you have a car?

(3) Can you speak Japanese?

(4) My girlfriend likes red flowers.

(5) School starts the first week of April.

(6) Look at the pretty fish in the warm water.

(7) There are many birds and animals in Australia.

(8) Australia is a big country without large rivers.

(9) Her mother lives near School Drive and Garden Street.

(10) The library is next to the nice Japanese garden in the park.

The following is a list of Japanese words in hiragana, along with English translations. You may repeat saying the sentences as many times as you like. **If you cough, sneeze or laugh during talking, you must repeat saying the word.** We will use your last utterance of each word. You must guess the word's pronunciation from the hiragana if you don't know the word. We cannot tell you the correct pronunciation of the word. **Pause for two seconds between lines.**

(11)	いらい	request
(12)	やめる	ill
(13)	ねる	knead
(14)	えんじん	engine
(15)	おん	debt
(16)	ようじん	V.I.P.
(17)	いぜん	still
(18)	うみ	sea
(19)	ぶどう	grape
(20)	まご	horseman
(21)	こうかい	open
(22)	じだい	era
(23)	おい	nephew
(24)	あり	ant

(25)	これら	these
(26)	じしん	earthquake
(27)	ばん	evening
(28)	にわ	lawn
(29)	あめ	rain
(30)	じょうぶ	upper region
(31)	のみ	chisel
(32)	どうじょう	training hall
(33)	じゅうびょう	serious illness
(34)	いつか	fifth day
(35)	おの	[a surname]
(36)	ごうせい	synthesize
(37)	ごよう	misuse
(38)	しょうがい	lifetime
(39)	ながい	long
(40)	これら	cholera
(41)	いらい	since
(42)	もる	leak
(43)	うる	gain
(44)	ばん	turn
(45)	あか	red
(46)	いじ	maintain
(47)	はし	bridge
(48)	おの	axe
(49)	てらす	shine
(50)	ようご	nursing
(51)	ながい	[a surname]
(52)	せんとう	head
(53)	きる	wear
(54)	こうかい	regret
(55)	すいへい	seaman
(56)	いがい	unusual
(57)	あおい	hollyhock
(58)	ごよう	official business
(59)	ようじ	infant
(60)	へんじ	reply

(61)	やむ	cease
(62)	いこう	traverse
(63)	ちかく	nearby
(64)	おもい	thought
(65)	いみ	meaning
(66)	あらわれる	wash
(67)	いみ	mourn
(68)	じゅうにん	ten people
(69)	えんじん	circular formation
(70)	どうじょう	sympathy
(71)	じゅうびょう	ten seconds
(72)	いつか	someday
(73)	ようご	term
(74)	かた	model
(75)	いじょう	malfunction
(76)	じゅうにん	inhabitant
(77)	いる	need
(78)	はし	chopsticks

(79)	じゅうまん	fill
(80)	はやく	side role
(81)	あめ	candy
(82)	いじょう	above
(83)	けんどう	regional road
(84)	のび	wildfire
(85)	わかい	young
(86)	うみ	pus
(87)	きる	cut
(88)	あおい	blue
(89)	ようじょう	open sea
(90)	あり	exist
(91)	あんざん	arithmetic
(92)	よくする	improve
(93)	かき	persimmon
(94)	じょうだん	top row
(95)	いぜん	before
(96)	もも	peach

(97)	ようじ	errand
(98)	ねる	sleep
(99)	ようじょう	recuperate
(100)	しょうがい	disability
(101)	うり	sale
(102)	いじ	obstinate
(103)	よくする	bathe
(104)	ぶどう	martial arts
(105)	うり	gourd
(106)	いる	shoot
(107)	じゅうまん	hundred thousand
(108)	へんじ	incident
(109)	おく	put
(110)	よい	evening
(111)	あわ	bubble
(112)	けんどう	swordsmanship
(113)	わかい	reconcile
(114)	あわ	millet

(115)	いがい	except
(116)	おん	sound
(117)	えんどう	roadside
(118)	げんご	original word
(119)	えんぎ	performance
(120)	すいへい	flat
(121)	じどう	automatic
(122)	にわ	two birds
(123)	ごうせい	extravagant
(124)	うる	sell
(125)	おい	hey
(126)	かた	shoulder
(127)	うむ	give birth
(128)	かき	below
(129)	じどう	child
(130)	もる	heap
(131)	さく	bloom
(132)	げんご	language

(133)	じだい	rent
(134)	えんどう	[a surname]
(135)	ふるい	shaker
(136)	よい	good
(137)	ごえん	connection
(138)	せんとう	public bath
(139)	じしん	self
(140)	じょうだん	joke
(141)	ちかく	perception
(142)	のみ	flea
(143)	じょうぶ	robust
(144)	やむ	fall ill
(145)	さく	rip
(146)	うむ	presence or absence
(147)	ごえん	five yen
(148)	わん	bowl
(149)	よういい	easy
(150)	わん	bay

(151)	あらわれる	appear
(152)	ふるい	old
(153)	いこう	after
(154)	おもい	heavy
(155)	えんぎ	origin
(156)	てらす	terrace
(157)	おく	deep
(158)	ようじん	caution
(159)	のび	growth
(160)	はやく	quickly
(161)	あか	dirt
(162)	いこう	go
(163)	まご	grandchild
(164)	あんざん	normal birth
(165)	もも	thigh
(166)	よういい	preparation
(167)	やめる	stop

The following is a list of Japanese sentences in kanji and hiragana, along with English translations. You may repeat saying the sentences as many times as you like. **If you cough, sneeze or laugh during talking, you must repeat saying the word.** We will use your last utterance of each word. You must guess the word's pronunciation from the hiragana if you don't know the word. We cannot tell you the correct pronunciation of the word. **Pause for two seconds between lines.**

(168) つぎの　雨(あめ)の　日(ひ)に　飴(あめ)を　買(か)う。

Buy candy the next rainy day.

(169) それ　以上(いじょう)は　異常(いじょう)だ。

Exceeding that is absurd.

(170) 広(ひろ)い　県道(けんどう)の　そばで、剣道(けんどう)の　練習(れんしゅう)を　した。

We practiced swordsmanship near the wide road.

(171) この　ふるいは、もう　古(ふる)い。

Here's an old shaker.

(172) あの　橋(はし)の　端(はし)で、箸(はし)を　拾(ひろ)った。

She picked up chopsticks at the foot of the bridge.

(173) 私(わたし)の　渉外(しょうがい)の　仕事(しごと)は、生涯(しょうがい)　続(つづ)けたい。

Public relations is my lifelong career.

This is the end. Thank you for your cooperation.

APPENDIX B PHONOLOGICAL RULES FOR ANAPTYXIS

In most cases of anaptyctic vowel insertions in loanwords into Japanese, the following vowel insertion rule applies:

(1) vowel insertion rule

C_ > Co / C = [t][d]

Ci / C = alveorlar affricate

Cu / C = otherwise

Words loaned up till the 19th century often had [i] inserted instead of [u] (i.e., in all non-[t][d] contexts). Consider older and more recent loans:

(2) older loans

excite > ekisaito

text > tekisuto

Texas > tekisasu

(3) newer loans

mix > mikisu > mikkusu

taxi > takushii

sex > sekkusu

Gemination occurs at syllable-final stops, which become moraic obstruents (tokushuhaku) followed by an anaptyctic vowel:

(4) gemination examples

step > sutepp

kit > kitto

black > burakku

bed > beddo

APPENDIX C AUTHOR'S PUBLICATIONS TO DATE

Publications authored by Goh Kawai are listed in reverse chronological order.

[1] Jin–song Zhang, Goh Kawai, and Keikichi Hirose "Subsyllabic tone units for reducing physiological effects in automatic tone recognition for connected Mandarin Chinese." Proceedings of ICPhS (International Congress of Phonetic Sciences) (San Francisco, California), to appear August 1999

[2] Goh Kawai, Ching Siu Lim, and Keikichi Hirose "A CALL system for correcting vowel insertions in English spoken by native speakers of Japanese." Proceedings of ICPhS (International Congress of Phonetic Sciences) (San Francisco, California), to appear August 1999

[3] Goh Kawai and Keikichi Hirose "Teaching the pronunciation of Japanese double–mora phonemes using speech recognition technology." Invited paper for the special issue on language learning using speech recognition technology. Speech Communication, to appear spring 1999

[4] Carlos Toshinori Ishi, Goh Kawai, and Keikichi Hirose "A pronunciation teaching system for Japanese lexical pitch accent." Proceedings of the ASJ (Acoustical Society of Japan) 1999 Spring Conference (Kawasaki, Japan), to appear March 1999

[5] Goh Kawai "Some can, some can't: Japanese learning to speak English." Invited survey for the special issue on learning non–native languages. "Journal of the ASJ (Acoustical Society of Japan), vol 54, no 1, pp 45–50, January 1999

[6] Goh Kawai and Keikichi Hirose "A method for measuring the intelligibility and non–nativeness of phone quality in foreign language pronunciation training." Proceedings of ICSLP (International Conference on Spoken Language Processing) (Sydney, Australia), pp 1823–1826, November 1998

[7] Jinfu Ni, Goh Kawai, and Keikichi Hirose "A synthesis–oriented model of phrasal pitch movements in standard Chinese." Proceedings of ICSLP (International Conference on Spoken Language Processing) (Sydney, Australia), pp 3317–3320, November 1998

[8] Goh Kawai "The mechanism and usefulness of computer–based non–native pronunciation learning systems." 22nd Tokyo Spoken Language Workshop (Tokyo, Japan), November 1998

[9] Goh Kawai and Keikichi Hirose "A bilingual speech recognizer for detecting phone–level pronunciation errors in non–native speech." Proceedings of the ASJ (Acoustical Society of Japan) 1998 Fall Conference (Yonezawa, Japan), paper number 1–2–24, September 1998

[10] Goh Kawai, Chingsiu Lim and Keikichi Hirose "Using speech recognition to detect epenthetic vowels in English spoken by native speakers of Japanese." Proceedings of the ASJ (Acoustical Society of Japan) 1998 Fall Conference (Yonezawa, Japan), paper number 1–2–25, September 1998

[11] Goh Kawai "Highlights of the STiLL (Speech Technology in Language Learning) workshop." IPSJ (Information Processing Society of Japan) Spoken Language Processing, 98–SLP–22–7, July 1998

[12] Goh Kawai and Keikichi Hirose "A CALL system for teaching the duration and phone quality of Japanese tokushuhaku." Proceedings of the Joint Conference of the ICA (International Conference on Acoustics) and ASA (Acoustical Society of America) (Seattle, Washington), pp 2981–2982, June 1998

[13] Goh Kawai and Keikichi Hirose "A CALL system using speech recognition to train the pronunciation of Japanese tokushuhaku." Proceedings of STiLL (Speech Technology in Language Learning) (Marholmen, Sweden), pp 73–76, May 1998

[14] Goh Kawai and Keikichi Hirose "Detecting pronunciation errors of non–native speakers of Japanese using speech recognition algorithms." Proceedings of the ASJ (Acoustical Society of Japan) 1998 Spring Conference (Yokohama, Japan), pp 339–340, March 1998

[15] Goh Kawai and Keikichi Hirose "Applying speech recognition to the teaching of Japanese pronunciation." IPSJ (Information Processing Society of Japan) Spoken Language Processing, 98–SLP–20–17, February 1998

[16] Goh Kawai, Tetsunori Kobayashi, Kazuya Takeda, and Hiroyuki Nishi "The search for a new paradigm for robust speech recognition." IPSJ (Information Processing Society of Japan) Spoken Language Processing, 98–SLP–20–7, February 1998

[17] Goh Kawai and Keikichi Hirose "A CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal, and mora obstruents." Proceedings of Eurospeech 1997 (Rhodes, Greece), pp 657–660, September 1997

[18] Goh Kawai and Keikichi Hirose "Improving the validity of corrective feedback of a CALL system using speech recognition to teach tokushuhaku production." Proceedings of the ASJ (Acoustical Society of Japan) 1997 Fall Conference (Sapporo, Japan), pp 357–358, September 1997

[19] Goh Kawai and Keikichi Hirose "Applying speech recognition to Japanese tokushuhaku pronunciation teaching." Proceedings of the 5th JCET (Joint Conference on Educational Technology) (Tokyo, Japan), pp 217–218, September 1997

[20] Goh Kawai and Keikichi Hirose "Applying speech recognition to the learning of pronunciation of Japanese long vowels, the mora nasal, and mora obstruents." Technical Report of the IEICE (Institute of Electronics, Information, and Communication Engineers), SP97–7, May 1997

[21] Goh Kawai, Takashi Kawasaki, and Keikichi Hirose "Speech rate normalization for CAI systems using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal, and mora obstruents." Proceedings of the ASJ (Acoustical Society of Japan) 1997 Spring Conference (Kyoto, Japan), pp 343–344, March 1997

[22] Goh Kawai, Akira Ishida, and Keikichi Hirose "A study on the reliability of scoring pronunciation of English spoken by Japanese students." Transactions of the IEICE (Institute of Electronics, Information, and Communication Engineers), vol J80–D–II, no 1, pp 367–368, January 1997

[23] Megumi Kameyama, Goh Kawai, and Isao Arima "A real–time system for summarizing human–human spontaneous dialogues." Proceedings of ICSLP (International Conference on Spoken Language Processing) (Philadelphia, Pennsylvania), pp 681–684, October 1996

[24] Goh Kawai, Takashi Kawasaki, and Keikichi Hirose "A CAI system using speech recognition for training pronunciation of Japanese long vowels, the mora nasal, and mora obstruents." Proceedings of the ASJ (Acoustical Society of Japan) 1996 Fall Conference (Okayama, Japan), pp 309–310, September 1996

[25] Goh Kawai "Diversification of spoken language research in the United States: confusion in the post–DARPA era." IPSJ (Information Processing Society of Japan) Spoken Language Processing, 96–SLP–12–7, July 1996

[26] Goh Kawai and Akira Ishida "The effectiveness of a speech recognition system for grading the pronunciation of English read by Japanese students." IEICE (Institute of Electronics, Information, and Communication Engineers) Fall Conference (Tokyo, Japan), September 1995

[27] Goh Kawai and Akira Ishida "An experimental study on the reliability of scoring pronunciation of English spoken by Japanese students." Technical Report of the IEICE (Institute of Electronics, Information, and Communication Engineers), ET95–44, June 1995

[28] Otoya Shirotsuka, Goh Kawai, Mike Cohen, and Jared Bernstein "Performance of speaker–independent Japanese recognizer as a function of training set size and diversity." Proceedings of ICSLP (International Conference on Spoken Language Processing) (Banff, Alberta), pp 297–300, October 1992

[29] Otoya Shirotsuka, Isao Arima, and Goh Kawai "Spoken Japanese sentence recognition based on Hidden Markov Models." Proceedings of the ASJ (Acoustical Society of Japan) 1991 Fall Conference, pp 31–32, October 1991

APPENDIX D AUTHOR'S BIOGRAPHY

1961	Born in Tokyo, Japan.
1980—1984	University of Tokyo, Department of Linguistics (B.Litt.)
1984—1986	International Christian University, Department of Educational Technology (M.A.)
1986—1988	Stanford University, Department of Linguistics (graduate student)
1986—1987	Xerox Palo Alto Research Center, Informations Systems Laboratory (Research Assistant)
1988—1996	SRI International, Speech Technology and Research Laboratory (Research Linguist)
1996—1999	University of Tokyo, Department of Information and Communication Engineering (Ph.D. expected)