A SYLLABLE-BASED, PSEUDO-ARTICULATORY APPROACH TO SPEECH RECOGNITION

by

LI ZHANG

A thesis submitted to the Faculty of Science
of the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
The University of Birmingham
Sept 2004

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

The prevailing approach to speech recognition is Hidden Markov Modeling, which yields good performance. However, it ignores phonetics, which has the potential for going beyond the acoustic variance to provide a more abstract underlying representation.

The novel approach pursued in this thesis is motivated by phonetic and phonological considerations. It is based on the notion of pseudo-articulatory representations, which are abstract and idealized accounts of articulatory activity. The original work presented here demonstrates the recovery of syllable structure information from pseudo-articulatory representations directly without resorting to statistical models of phone sequences. The work is also original in its use of syllable structures to recover phonemes. This thesis presents the three-stage syllable based, pseudo-articulatory approach in detail. Though it still has problems, this research leads to a more plausible style of automatic speech recognition and will contribute to modeling and understanding speech behaviour. Additionally, it also permits a 'multithreaded' approach combining information from different processes.

ACKNOWLEDGEMENTS

First of all, my thanks go to my supervisor, Dr. William Edmondson. Without his support, guidance, discussion and enthusiasm, I would have never completed this thesis.

I also thank my thesis group members, Dr. Peter Coxhead and Prof. Xin Yao for their helpful comments and discussion. Additionally, thanks must go to the support team. None of this would have been possible without them.

I would like to thank Prof. Martin Russell and his research staff: Michael Wong and Shona D'arcy for their valuable help on technical problems. I have learnt a lot from them and no words can exactly express my gratitude.

The research was funded by both an ORS award from the Universities UK (formerly CVCP) and a research student grant from the School of Computer Science, the University of Birmingham. I am very grateful for this financial support. I owe my thanks to Dr. Peter Hancox too not only for his giving me an opportunity to be a research student in the school, but also for his help throughout these years.

Finally, I would like to thank Ming and my parents for their patience and support throughout. Especially, my thanks go to my grandparents for their invaluable encouragement and love throughout my life. This thesis is for them.

CONTENTS

LIST OF FIGURES VIII

LIST OF TABLES X

1 INTRODUCTION 1

1.1 LAYOUT OF THE THESIS 4

2 SPEECH AND SPEECH RECOGNITION TECHNIQUES 5

2.1 SOME PHONETIC AND PHONOLOGICAL CONCEPTS 5

2.2 SYLLABLES 7

2.2.1 Syllable structure 7

2.2.2 Sonority 8

2.3 REVIEW OF SPEECH RECOGNITION TECHNIQUES 10

2.3.1 Hidden Markov Models 10

2.3.2 Articulatory speech recognition 12

2.3.3 Syllable-based approach to speech recognition 20

2.3.4 Alternative approaches 25

2.4 TIMIT - SPEECH DATABASE DESCRIPTION 26

2.4.1 Time-aligned phonetic transcription 28

2.5 EARLIER DOCTORAL WORK AS BACKGROUND 29

2.5.1 Pseudo-articulatory features as a natural approach to speech recognition 30

3 ARTICULATORY-ACOUSTIC MAPPING 33

3.1 ARTICULATORY-ACOUSTIC MAPPING PROBLEM 33

3.2 BACKGROUND MATERIAL 37

3.2.1 Speech data 37

3.2.2 Formant representation 37

3.2.3 Feature description 43

3.2.4 Multiple regression analysis 45

3.3 MATERIAL AND METHODS 47

3.3.1 Analysis step 47

3.3.2 Plosives 48

3.4 VOWEL MODEL 49

3.5 CONSONANT MODEL 51

4 DERIVATION OF PSEUDO-ARTICULATORY TRAJECTORIES – RECOGNITION I 54

4.1 A SCHEMATIC VIEW OF THE RECOGNITION PROCESS 54

4.2 SPEECH DATA – TEST SET 55

4.3 BRUTE SEARCH 56

4.4 RECOGNITION RESULTS AS PSEUDO-ARTICULATORY TRAJECTORIES 58

4.5 EVALUATION BY RESYNTHESIS 60

4.6 DISCUSSION AND CONCLUSIONS 62

5 DERIVATION OF RECOVERED SYLLABLE PATTERNS – RECOGNITION II 66

5.1 USING SYLLABLES IN SPEECH RECOGNITION 66

5.1.1 Articulatory pattern in the syllable 68

5.2 SYLLABLE RECOVERY USING IDEALIZED PSEUDO-ARTICULATORY TRAJECTORIES 71

5.2.1 Syllable recovery 72

5.3 SYLLABLE RECOVERY USING DERIVED PSEUDO-ARTICULATORY TRAJECTORIES. 76
5.3.1 Smoothing derived pseudo-articulatory trajectories *76*
5.3.2 Evaluation by resynthesis *78*
5.3.3 Syllable recovery *79*
5.4 RECOGNITION RESULTS OF RECOVERED SYLLABLE PATTERNS 79
5.4.1 Result format *80*
5.4.2 Recognition results *80*
5.5 DISCUSSION AND CONCLUSIONS 81

6 FINDING A PHONEME SEQUENCE – RECOGNITION III 85

6.1 DYNAMIC PROGRAMMING 85
6.2 PHONEME RECOGNITION RESULTS 87
6.3 EVALUATION 94
6.4 DISCUSSION AND CONCLUSIONS 96

7 CONCLUSIONS 98

7.1 SUMMARY AND CONCLUSIONS 98
7.2 FUTURE DIRECTIONS 100
7.2.1 One possible choice of new features *104*
7.2.2 Refinement of syllable model *105*

APPENDICES 106

APPENDIX A TIMIT DATA 106
Appendix A.1 Phone representation *106*
Appendix A.2 Speech data transcription *109*
APPENDIX B REGRESSION COEFFICIENTS 112
APPENDIX C CONSONANT FEATURE VALUES 113
APPENDIX D RECOGNITION RESULTS – RECOVERED SYLLABLE PATTERNS 114
APPENDIX E RECOGNITION RESULTS – PHONEME CANDIDATES AND TRANSITIONS .. 131

Appendix E.1 Phoneme candidates and transitions for individual sentences..... 131
Appendix E.2 Confusion matrix 158
APPENDIX F EVALUATION RESULTS OBTAINED BY HRESULTS FOR NEW SPEECH DATA
..... 162

BIBLIOGRAPHY 165

INDEX 176

LIST OF FIGURES

FIGURE 2.1 THE HIERARCHICAL SYLLABLE STRUCTURE 8
FIGURE 2.2 SONORITY WAVE OF PLUM 9
FIGURE 3.1 EXAMPLE SPECTROGRAMS WITH OVERLAID FORMANT TRAJECTORIES FOR WORD "OFF" AND WORD "SUIT" (SPEAKER ID: DR1/MMRP0) 42
FIGURE 4.1 A SCHEMATIC VIEW OF THE RECOGNITION PROCESS 55
FIGURE 4.2 TRAJECTORIES FOR FEATURE "HIGH" FOR ONE SENTENCE: IDEALIZED AND RECOGNIZED 58
FIGURE 4.3 TRAJECTORIES FOR FEATURE "BACK" FOR ONE SENTENCE: IDEALIZED AND RECOGNIZED 59
FIGURE 4.4 TRAJECTORIES FOR FEATURE "ROUND" FOR ONE SENTENCE: IDEALIZED AND RECOGNIZED 59
FIGURE 4.5 TRAJECTORIES FOR FEATURE "TENSE" FOR ONE SENTENCE: IDEALIZED AND RECOGNIZED 59
FIGURE 4.6 SIGNAL WAVEFORMS FOR ONE SENTENCE: TOP – ORIGINAL, BOTTOM – SYNTHESIZED SPEECH 62
FIGURE 5.1 THE LAYERED SYLLABLE MODEL 69
FIGURE 5.2 AN EXAMPLE OF THE SYLLABLE MODEL 69
FIGURE 5.3 THE TOP SECTION SHOWS THE SPECTROGRAM OF THE UTTERANCE " THERE IS USUALLY A VALVE. ". THE MIDDLE 4 TRACES SHOW THE IDEALIZED FEATURE TRAJECTORIES OF HIGH, BACK, ROUND, TENSE. THE BOTTOM SECTION SHOWS IN SCHEMATIC FORM THE RECOVERED SYLLABLE POSITIONS. 75
FIGURE 5.4 TRAJECTORIES FOR FEATURE "HIGH" FOR ONE SENTENCE: DERIVED AND SMOOTHED 76
FIGURE 5.5 TRAJECTORIES FOR FEATURE "BACK" FOR ONE SENTENCE: DERIVED AND SMOOTHED 77

FIGURE 5.6 ONE EXAMPLE RESULT OF RECOVERED SYLLABLE PATTERNS 82
FIGURE 5.7 THE DERIVATION OF MULTIPLE INDEPENDENT SOURCES OF INFORMATION 84
FIGURE 6.1 PHONEME RECOGNITION RESULTS 89
FIGURE 6.2 ISKRA'S RECOGNITION RESULTS 95
FIGURE 7.1 SUMMARY OF WORK DONE 101

LIST OF TABLES

TABLE 3.1 ACOUSTIC CORRELATES OF CONSONANTAL FEATURES [LADEFOGED 1993]..... 41
TABLE 3.2 NEW FEATURE VALUES FOR VOWELS [ISKRA 2000] 45
TABLE 3.3 COEFFICIENTS OF DETERMINATION 51
TABLE 6.1 A FRAGMENT OF THE CONFUSION MATRIX 91
TABLE 6.2 ONE PSEUDO EXAMPLE OF RECOVERED SYLLABLE RESULTS 92
TABLE 6.3 EVALUATION RESULTS FOR THE NEW SPEAKER (ID: DR1/MDAC0) 93
TABLE 6.4 EVALUATION RESULTS FOR THE PREVIOUS SPEAKER (ID: DR1/MMRP0) 94
TABLE A.1 PLOSIVES (STOPS) 106
TABLE A.2 FRICATIVES AND NASALS 107
TABLE A.3 AFFRICATES 107
TABLE A.4 SEMIVOWELS AND GLIDES 107
TABLE A.5 VOWELS AND DIPHTHONGS 108
TABLE A.6 OTHERS 108
TABLE B.1 REGRESSION COEFFICIENTS 112
TABLE C.1 CONSONANT FEATURE VALUES 113
TABLE F.1 EVALUATION RESULTS FOR 8 UTTERANCES 162

1 INTRODUCTION

We never think carefully about our ability to produce and understand speech and usually do not pay much attention to its nature and function. It is not surprising, therefore, that many of us ignore the great influence of speech on the development of human society. Wherever people live together, they use speech to communicate with one other. This is true for people who even live in the most isolated island. In fact, speech is our basic ability which sets human beings apart from other mammal animals and gives us the ability to think abstractly [Denes and Pinson 1993].

Why is speech so important? It has made contributions to the development of human civilization by providing people with the ability to share experience, exchange ideas and transmit knowledge. Actually there are many other communication methods available in human society, such as sign language used by deaf people, smoke signals used in jungles, pistol signals used in a mission and various writing systems. Among all of these, speech is, under most circumstances, the most efficient and convenient communication means for human society [Denes and Pinson 1993].

Yet some people may believe that writing is the most important means for communication since books and newspapers play very important roles in our daily life and they provide more durable ways of knowledge transmission. Additionally, written material and publications are extensively required. However, the use of mobiles, televisions and radios has expanded greatly as well. Since people rely heavily on the use of the most easy and convenient way for communication, the amount of the information exchanged by speech is undoubtedly still the largest [Denes and Pinson 1993].

Because of its importance to human civilization, speech is worth our careful study. Also a careful study may provide engineers with a better understanding of the speech mechanism so that this knowledge can be used to develop efficient communication systems for different applications. New technologies have advanced the research in

speech production and speech perception. With the development of computer technology, scientists have implemented these two processes in an automatic manner: speech synthesis and speech recognition. Particular efforts have been made in speech processing for recognition. Automatic speech recognition has been a goal of research for more than four decades and has inspired many science fiction wonders such as the highly intelligent self-reconstructed robots in the Terminator series of movies. However, in spite of the various attempts at designing an intelligent machine that can recognize the spoken word and comprehend its meaning, and in spite of the enormous research efforts spent in trying to create such a machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments.

Currently speech recognition systems based on hidden Markov modeling (HMM) have achieved the best performance. The underlying assumption of HMM (or any other type of statistical model) is that speech signal can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined (estimated) in a precise, well-defined manner [Rabiner and Juang 1993]. The HMM method provides a highly reliable way of recognizing speech for a wide range of applications and integrates well into systems incorporating both task syntax and semantics. However, it ignores details of the vocal tract (co-articulatory effects) and linguistic processes (e.g. morpho-phonemic constraints). Further, its good recognition performance seems to block the way for the development of alternative approaches, which incorporate more phonetic knowledge but cannot possibly provide equally high performance in the short term.

HMM relies heavily on the use of statistics to model the variability of speech. One of its major drawbacks is that it ignores most of what can be learnt from phonetics – the latter has the potential for going beyond the acoustic variance to provide a more abstract underlying representation. Thus using phonetic knowledge is one of the motivations of this research. Pseudo-articulatory representations (PARs) established by earlier work [Edmondson, Iles and Iskra 1996; Edmondson, Iskra and Kienzle 1999], which represent linguistic generalizations and idealizations of articulation and the articulator positions,

have been used. Their abstraction provides the potential to deal with problems such as many-to-one mapping or coarticulation [Iles 1995; Iskra 2000]. They can serve as a perfect bridge to link the acoustic representation of the speech signal and a sequence of syllables and phones, the results of the recognition process. Because of these advantages, PARs will be used independently of HMM as the driving force of the recognition process in this research.

Conventional HMM is mainly based on phones for modeling spoken words. However, time and research have shown that phones are too small an acoustical unit to model temporal patterns and variations in continuous speech [Hamaker, Ganapathiraju and Picone 1997; Ganapathiraju, Hamaker, Ordowski, Doddington and Picone 2001]. It also ignores information in the signal which can be used to derive syllable structural information or syllable timing independently of phone sequences. More recently, attention has shifted to a larger acoustic context. Research has shown that speech recognition systems based on syllables can overcome some of the disadvantages caused by phone modeling systems [Hamaker et al. 1997; Ganapathiraju et al. 2001]. These are another two primary motivations of this work. Our original approach shows that syllable structural information can be directly recovered from the speech stream in the form of PARs without reliance on phonetic segment identification [Zhang and Edmondson 2002a; 2003; see also Edmondson and Zhang 2002 for background theoretical work on syllable structure]. The work is also original in its use of syllable structures to recover phonemes. This research permits speech recognition to proceed from a ‘multithreaded’ base, which is important regardless of the detail of our approach. Additionally, independent measures of timing information and sonority may provide more independent contributions and future work will consider these possibilities. In general, this research leads to a more plausible style of automatic speech recognition and will contribute to knowledge of phonetics and speech behaviour.

1.1 Layout of the thesis

The reminder of the thesis is structured as follows. Chapter 2 gives a technical review of related work, especially identifying what recent research has been done in this area. Chapter 3 introduces the relevant background material and the basis for recognition experiments. This includes the details of mapping procedure, vowel model, and consonant model. Chapters 4, 5, and 6 contain the main stages for speech recognition processing. In Chapter 4, we present the work on the implementation of pseudo-articulatory trajectories and their evaluation procedure as the first stage of the recognition processing. In Chapter 5, we introduce how to recover syllable patterns from PAR trajectories as the second stage of processing. We will present the idea, implementation, and the recognition results of recovered syllable patterns. In Chapter 6, the recognition work has continued using the recovered syllable patterns to find candidate phoneme sequences. The statistical analysis of the final recognition results is also presented and shows that the recognition results are very promising. The overall experiments have been conducted on another speaker (ID: dr1/mdac0) from the TIMIT corpus in order to show this approach's potential for speaker independent recognition. We conclude in Chapter 7 by summarizing the work and highlighting possible areas for future research.

2 SPEECH AND SPEECH RECOGNITION TECHNIQUES

The approach to speech recognition pursued in this thesis is motivated by phonetic and phonological considerations. Therefore, before proceeding to discuss different speech recognition techniques, an explanation of a number of phonetic and phonological terms and phenomena will be provided, as well as a description of syllables. Following the discussion of various speech recognition techniques, we provide a description of the TIMIT speech database, since all the speech material used in this research comes from it. Finally, we review earlier doctoral work, which serves as the basis for the novel approach presented in this thesis.

2.1 Some phonetic and phonological concepts

From west to east, south to north, people speak to each other. All the sounds they produce constitute the universal set of human speech sounds. *Phonology* is concerned with the ways in which these speech sounds form systems and patterns in human languages. *Phonetics* is a part of phonology and provides the ways for describing speech sounds. A relatively small set of phonetic features characterizes all these sounds.

Speech sounds are represented in two levels of representation – a concrete (phonetic) one and an abstract (underlying) one. The concrete phonetic level of representation of a speech sound is called a *phone* (represented by []), while the abstract level of representation is called a *phoneme* (represented by //). One realization of a certain phoneme is called an *allophone* of this phoneme. The minimal phonological units are phonemes since they cannot be divided into smaller successive units. Each phone corresponds to exactly one phoneme on the abstract level. But this does not mean that the phones on the concrete phonetic level are discrete units with clearly defined boundaries in real speech because of the fact that articulatory feature transitions from one phone to the next do not necessarily happen at the same time. Some phones, such as nasalized vowels, occur because nasalization is held over (or anticipated) from an adjacent consonant. Thus phone segmentation becomes the first phonological analysis. There are

larger phonological units than individual phonemes: *syllables*, which play an important role in phonological analysis. Additionally, segments and syllables are not only phonological units, but also phonetic units. A detailed discussion of syllables is provided in the following section.

The importance of phonology is that it enables one to change one word into another by simply changing one sound. Such a pair of words is called a *minimal pair* if the two words have identical phonological structure except for one sound segment that occurs in the same place in the string. For example, “pan” and “ban”: except for the initial sound, these two words are identical in phonological structure in English. And the two distinctive sounds represent two different phonemes. Different realizations of the same phoneme (allophones) are not phonetically exactly identical for one speaker, though all allophones of a certain phoneme show phonetic similarities and share most of the same phonetic features between them. These allophones never occur in the same phonological context and are in complementary distribution. In “pine”, the initial sound [p] is aspirated before a stressed vowel, while in “spine”, it is not after a syllable initial sound [s]. These two allophones of /p/ complement each other, which means in English the aspirated [p] does not occur after [s]. Moreover, two realizations of the same allophone can be phonetically different. These phonetically different realizations are different phones.

Distinctive features are a set of features playing a distinctive role in the phonological discrimination of sounds [Laver 1994]. They are usually binary ones with the value of presence or absence. Each phoneme can be described by a group of distinctive features. The English phoneme /s/ can be described with the set of features: [+consonantal], [-sonorant], [-voiced], [+coronal], etc. Each phoneme differs from all the other phonemes by at least one distinctive feature. *Phonetic features* account for the differences between the phonetic realizations of a segment. They may have multiple values. The set of distinctive phonological features used to account for the phonological discrimination of phonemes is smaller than the set of phonetic features used to account for the differences between the phonetic realizations of these units [Laver 1994].

2.2 Syllables

In his introductory phonetic textbook [1993], Ladefoged concedes that there is no agreed phonetic definition of a syllable. He notes that although nearly everybody can identify syllables, it is difficult to define a syllable with precision. There have been various attempts to define the syllable either in terms of properties of sounds, such as sonority or prominence, or from the perspective of the speaker, principally the notion that a syllable is a unit of organization for the sounds in an utterance. However, none of these attempts has yielded a completely satisfactory definition. This thesis does not attempt to provide a solution to this intriguing linguistic problem, but for the convenience of discussion, we simply assume that a syllable is the smallest articulatory coherent span of speech in the sense that every speech utterance must contain at least one syllable (see section 5.1.1 for further details).

Syllables play an important role in language structure and in language use. In speech production, the syllable is the most efficient abstract unit on which to base phonological encoding [Levelt, Roelofs and Meyer 1999]. In speech perception, syllables play a multitude of roles, some universal, some language-specific.

In speech recognition processing, syllables can contribute both as segmental and supra-segmental units. Syllables vary in weight or complexity, and in pattern in many languages in ways that support or enable prosodic structuring in speech – within words and across groups of words. Obviously, such useful information provided by syllables can be exploited in speech recognition systems. Before considering specific syllable-based systems, we will discuss syllable structure and the sonority contour.

2.2.1 Syllable structure

The internal syllable structure can be described in various ways. Traditionally, first of all, a syllable can be regarded as a sequence of phonemes with a vowel and up to three consonants before and after. Any sequence of phonemes for one syllable is limited by phonotactic constraints.

Another conventional approach to the analysis of syllable structure shows a syllable has a hierarchical structure. Every syllable has a nucleus. It can be a vowel, a syllabic liquid or a nasal. The sequence of consonants before the nucleus is called onset and the one after the nucleus is called coda. Nucleus and coda constitute the subsyllabic unit – rhyme. The hierarchical structure is shown in figure 2.1. In addition, the onset and coda are optional. They both may not be evidenced when the syllable consists of only a vowel. Not only does the sequence of segments, which constitutes a syllable, need phonotactic constraints, but also the syllable structure needs these constraints to determine the maximal onset principle [Ewen and van der Hulst 2001] or the allocation of a consonant to one syllable or the next.

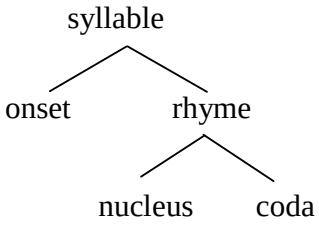


Figure 2.1 The hierarchical syllable structure

2.2.2 Sonority

A slightly less conventional approach, but with a good tradition, notes that speech segments vary along a continuum of sonority values. The ‘sonority theory’ presented in this section is far from being universally accepted. It is adopted here because it lends itself particularly well to the descriptions of syllables in English. “The sonority of a sound is its relative loudness compared to other sounds, everything else (pitch etc.) being equal” [Giegerich 1992]. Speech sounds can be ranked by their relative sonority: low vowels are the most sonorant of all the sounds while voiceless oral stops are of minimal sonority. All other sounds are ranked in between these two extreme points of sonority scale. Moreover, Selkirk [1984] even provided a table of speech sounds with sonority values. The significance of this approach for our purpose is in two aspects.

Firstly, with the knowledge of the sonority scale/values, it is known that syllable nuclei are associated with peaks of sonority, which can be used to predict the right number of syllables for a great majority of English words. Moreover, it is also well known that syllables are sonority waves with the nucleus as the sonority peak and the onset and coda as the sonority troughs. The smoothed sonority contour does provide another form of explanation for the constraints on the consonant sequences. For example, a word like *plum* conforms to the expectation derived from the sonority theory. The phonemic representation is /plahm/. According to the sonority scale, /p/ is less sonorant than /l/, which is less sonorant than /ah/. And /m/ is less sonorant than /ah/. The graphic representation of the sonority wave of *plum* is in figure 2.2, showing a single sonority peak and smoothed troughs for consonant sequences. In general, the sonority theory gives us a general understanding of what syllables are: "syllables are associated with peaks in sonority in such a way that in a given string of phonemes, every syllable corresponds to a single sonority peak" [Giegerich 1992]. This definition is probably universal. It can be used for the descriptions of syllables of all languages. However, it still leaves a number of questions unsolved and fact unexplained. For example, in English words *asks* or *six* do not conform to the expectations derived from the sonority theory, which makes the use of sonority for modelling syllables still an open research topic.

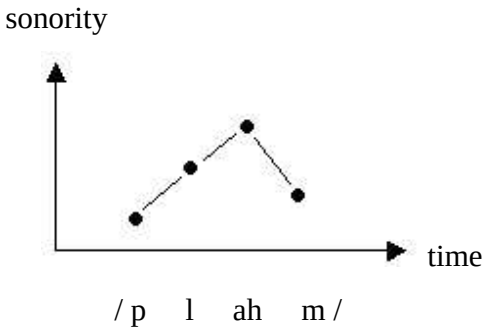


Figure 2.2 Sonority wave of *plum*

Secondly, the sonority theory, outlined here, has shifted the attention away from the conventional sequential arrangement of segments to an articulatory process model. For example, Beckman, Edwards and Fletcher [1992] have defined sonority in terms of impedance forward of the vocal folds. Obviously, this has the potential for use in speech recognition processing.

2.3 Review of speech recognition techniques

In this section, we review various speech recognition techniques from the literature which are closely related to the topic of this thesis. We begin by introducing widely used statistical method of characterizing the spectral properties of the frames of a pattern, namely the hidden Markov model (HMM) approach. Then we focus on the introduction of articulatory speech recognition. Following this, we consider the implementation and discussion of syllable-based approach to speech recognition. Finally, we also review some other approaches to speech recognition.

2.3.1 Hidden Markov Models

The basic methodology of hidden Markov models was invented by A. A. Markov, a Russian mathematician in the early 1900s during his studies of word statistics in literary texts. Since the mid 1970s, statistical concepts based on HMMs have been applied to speech recognition processing. Since the 1980s, HMMs have become the most popular speech recognition method. The mathematical theory has led to practical computational solutions, which consistently achieved the best recognition results to date.

There are many properties which make HMMs attractive for use in speech recognition. The most important of all is that they can represent/model events whose statistical properties change with time. This is particularly the case since it is possible to use a HMM to model the spectrum changes that occur during a word. These changes are related to the sequence of sub-word speech sounds, for example the sequence of phonemes, which make up the word [Denes and Pinson 1993].

The HMM method provides a highly reliable way of recognizing speech for a wide range of applications and integrates well into systems incorporating both task syntax and semantics. The current syllable-based HMMs provide 53.4% accuracy rate for syllable identification using TIMIT database. This equates to a phoneme accuracy rate of 72.8% [Stewart, Ming, Hanna and Smith 2002]. They used the syllabified hand-labelled transcripts of the TIMIT sentences which resulted from their previous work for all of the experiments. The total vocabulary of syllables used in their experiments was composed of 9414 distinct syllables. This vocabulary was composed of all the unique syllables found in the training and testing data of the TIMIT database. Syllable HMMs were trained on the training data and were used to identify syllables in the test sentences. The syllable-based experiments were identification experiments instead of recognition experiments, which meant that the syllable HMMs were used to identify discrete syllables with known boundaries.

Hidden Markov modelling is currently the most popular approach to speech processing for recognition. It has achieved the best recognition results so far and with the help of powerful language models and careful dialogue design[1] it is reliable enough to be implemented in commercial products. However, HMM relies heavily on the use of statistics to model the variability of speech, such as coarticulation effects and inter-speaker differences, and this technique has nothing in common with the actual mechanisms of speech production or perception. Lee [1989a], for example, states that "HMM is a very inaccurate model of the speech production process". Therefore, it seems desirable to incorporate linguistic and articulatory knowledge into the current speech recognition systems. Some people try to combine phonology and phonetic knowledge with hidden Markov modelling [King, Taylor, Frankel and Richmond 2000], while others seek alternative recognition techniques. In the next section some of these systems will be discussed in detail.

[1] By controlling the dialogue structure through asking well-bound rather than open questions, the recognition task can be considerably simplified. With such a limited task it is possible to implement speech recognition in commercial systems.

2.3.2 Articulatory speech recognition

Recently, speech recognition systems based on articulatory features such as “voicing”, “labial” or the positions of the lips and tongue have gained increasing attention and interest, because such systems are more robust to noise, reverberation and speaker variability. From a general point of view, these systems are also interesting since they adopt phonological and phonetic concepts, which provide a richer description of a speech activity than the sequence of HMM-states, which is the most prevailing technique for automatic speech recognition (ASR) to date. In this section, we will discuss some speech recognition systems which incorporate articulatory features into some stage of the recognition process.

Markov, Dang, Iizuka and Nakamura [2003] report an automatic speech recognition system (for Japanese) where articulatory features extracted from the human speech production system, in the form of articulatory movement data, are effectively integrated in the acoustic model for improved recognition performance. The articulatory data are used only for the acoustic model training. In the recognition process only acoustic data in form of standard mel-frequency cepstral coefficients (MFCC) features are used, since the authors believe that articulatory features are very difficult or impossible to obtain during recognition. They adopt the hybrid HMM and Bayesian Network (BN) model, which makes possible the integration of different speech features by modeling probabilistic inter-dependencies. This work also demonstrates a way to apply details of the speech production mechanism to an ASR system.

First of all, the articulatory data have been collected using the electromagnetic midsagittal articulographic (EMMA) system [Okadome and Honda 2001]. They are collected by placing four receive coils on the tongue surface in the midsagittal plane, and one coil for each of the upper lip, lower lip, maxilla incisor, mandible incisor and the velum, respectively. Then in their coordinate system, the maxilla incisor is chosen as the origin. The articulatory data obtained from the other eight observation points are time-varying vectors with 16 components accounting for both the x- and y- coordinates. With

the first and second order coefficients included, the final articulatory data are 48 dimensional. In order to confirm the validity of the articulatory data for speech recognition processing, a preliminary experiment using HMM has been conducted using the acoustic data alone as well as both the acoustic data and articulatory data. The results show that recognition process outperforms the baseline system when using the combined features.

In the next step of the processing, they use a combination model of HMM and BN to integrate the extra articulatory data into the recognition process. In this model, the BN parameters are obtained using standard statistical algorithms using both articulatory and acoustic data. During recognition, they only use acoustic feature vectors. The first experiment was conducted using speaker dependent models. The same speech data have been used in the test set as those used in the preliminary experiment. The results show that all the accuracy rates of HMM/BN are higher than those of the baseline (using acoustic data only) HMM. And the same tendency has been obtained using multi-speaker models in the following experiments. Additionally the HMM with combined acoustic and articulatory feature vectors even outperforms the HMM/BN system. This means there is more potential for utilizing the articulatory data in ASR.

Commonly used acoustic features such as MFCC are primarily representations of the short time spectral characteristics of speech. The articulatory features, however, are an attempt to characterize the motions and changes of the sound production system. Since the natures of the two features are different, the articulatory data provide additional useful information for speech recognition, which is not included in the acoustic models. This work is one attempt to combine the acoustic and articulatory features. Though their combined HMM/BN model has already shown its superiority over the conventional HMM in almost all their experiments, where the combination is exploited in training, there should be further value from combining articulatory movement data in the recognition processing (test set). In general, their system can be regarded as the first step towards the utilization of articulatory features in speech recognition processing.

Leung and Siu [2002] investigate multiple ways to further improve the combination of an articulatory feature (AF) model and an acoustic feature (ACF) model. First, a multiple-distribution AF model is proposed, which increases the model's resolution by separately modelling different sub-phone segments. Then the asynchronous combination of this multiple-distribution AF model with an ACF HMM, which provides flexible combination of states between these two models, is discussed. Finally, the AF information has been used for the ACF model training so that the ACF model is optimized to give the best recognition performance when combining with AF model for decoding. Indeed, among the combined and individual feature systems, it proves to be the best one.

First the baseline system of articulatory features is introduced, which is a hybrid HMM/ANN system using 5 different AF estimators and a phone estimator with N outputs where N is the number of phones. A hybrid HMM/ANN system is different from traditional HMM since the observation probabilities of the HMM are estimated by an artificial neural network (ANN). Additionally, the AF estimators and the phone estimator are implemented using Multi-Layer Perceptrons (MLPs). According to Leung and Siu, the 5 feature estimators used in the baseline system are "voicing MLP", "rounding MLP", "front-back MLP", "manner MLP" and "place MLP" with 27 values as the output.

In the baseline system, the AF model uses only one observation distribution per phone. However each phone can be affected by its context, thus the AF characteristics are different from one part of the phone to the others. By using different observation distributions for different phone segments a multiple-distribution AF model is obtained. The two AF models are evaluated using constant and estimated transition probabilities. The performance of the multiple-distribution AF model is less sensitive to the change of the state transition probabilities. This means that some state transition information has been already included in the distributions of different phone segments. Even under the same conditions, the single-distribution AF model still performs lower than the multiple-distribution system.

Secondly, the combination of an ACF HMM with the multiple-distribution AF model is explored. Not only is the alignment of the ACF HMM states and the AF model states needed, but also the combined state transition probabilities need to be calculated. Because it is unlikely that the same state index of the two models capture the same portion of phone, asynchronous state combination has been chosen to provide flexible combination of states between these two models. In order to improve recognition performance, they re-estimate the combined state transition probabilities for the combined system. Moreover, the ACF HMM model has also been re-trained with the knowledge that the AF model will be used. Actually, the phone MLP of AF models should also be re-trained by incorporating the ACF HMM information. But this is not done in their current systems.

Finally, experiments were conducted using different combinations of the two AF models and the ACF HMM. The synchronous state combination is applied to the single-distribution AF model, since it has no state synchrony issue. Both synchronous and asynchronous state combinations are applied to the multiple-distribution AF model. The systems with only individual features are also tested. The asynchronous combined system of the ACF HMM and the AF model with multiple distributions performs the best, providing an absolute improvement of 1.8% compared with the ACF HMM alone.

This work indicates that it is really crucial to know how to combine the acoustic information with the articulatory features in one combined system. They succeed in implementing the system using an asynchronous combined model. The parameters of individual models can be re-estimated to optimize this combined model's performance. However, they only re-train the parameters of ACF HMM using AF information. Further work is needed to re-train the AF model so that it incorporates the ACF HMM information, which may lead to the increase in the recognition percentages.

Metze and Waibel [2002] present a speech recognition system, which incorporates dedicated detectors for phonological or articulatory features into conventional context-

dependent HMM sub-phone models, using (what they call) a stream architecture. They use articulatory features as phonological distinction for speech sounds only and are not concerned with the relationship with actual articulatory movements. By using the combination of scores at the log-likelihood level, they provide a promising approach for the fusion of features and standard models. They set the system up with 76 binary phonological features such as “palatal” or “coronal” etc. But their main goal is to show how supporting a conventional ASR with only a few streams of articulatory features can improve the recognition results significantly.

Their first experiment is conducted on the read Broadcast News task (readBN). They use the 76 linguistically motivated questions to set the system up, which are used during construction of the decision tree for context-dependent modelling. This setup permits a very flexible combination of existing models with detectors for articulatory states. Then the feature detectors are built in exactly the same way as acoustic models for existing speech recognizers. In their experiments, the Janus [Finke, Geutner, Hild, Kemp, Ries and Westphal 1997] speech recognition toolkit is used for training. A relevant detail of the acoustic training is that they only use the middle frames because the phonological properties may not be evidenced at the beginning and ending parts of a phone. The classification output of the feature detectors on readBN data indeed approximates the canonical feature values very well. Since there are a number of feature detectors available, the next step is to choose which ones are to be used in the recognizer. They decided to use three different approaches to incrementally add feature streams to the baseline HMM system using equal weights for all states and streams. But the three approaches perform approximately equally. In all cases, the word error rate decreases to a minimum when using 6 to 9 features, then slowly starts to increase again. In the future work, they plan to use other methods for feature selection and the determination of stream weights. Finally, the best feature system, using 8 features: “affricate”, “palatal”, “glottal”, etc, has been obtained with word error rate 11.6%, while word error rate for the baseline system without feature streams is 13.4%.

In order to test their approach on a large number of speakers and on spontaneous speech under clean condition, ESST data has been used. Adding the same features as those used in the above best feature system and with the same weight values, word error rate reduces to a minimum 23%, while the one of the baseline system is 23.5%. However, only a slight improvement in performance has been achieved for Switchboard (SWB) task.

Though their approach improves ASR performance on many different tasks, several points need to be noted. First, they only use binary phonological features in their approach instead of continuous feature values, which indeed improved the performance on a small test-set of hyperarticulated data. However it is obviously inadequate for spontaneous or sloppy speech, which needs a more complex articulatory model. Using continuous feature values it is possible to provide a more appropriate modelling of articulatory feature trajectories for spontaneous speech. Moreover, using continuous values can reduce the number of features needed to describe a set of sounds (in comparison to binary valued features).

The work presented by Stuker, Metze, Schultz and Waibel [2003] is an extension of the work presented by Metze and Waibel [2002] by incorporating cross- and multilingual feature detectors into an HMM based speech recognition system. In their earlier work, they have shown that articulatory features can compensate for inter-language variability and can be recognized across languages. Following Metze and Waibel, Stuker et al. use a similar stream architecture to incorporate feature detectors into HMM. By carefully selecting feature detectors and stream weights in a discriminative way, the obtained word error rate reduction is very close to the reduction when using monolingual feature detectors. Additionally, the combination system of a standard English model and feature detectors from many languages achieves a better performance than the system using language specific feature detectors.

First of all, 5 languages have been chosen for their research: Mandarin Chinese, English, German, Japanese, and Spanish. The GlobalPhone global unit set is used, which is based

on the International Phonetic Alphabet created by IPA, to produce a global set of features on the five selected languages. The features are divided into two categories: one for consonants including manner and place of articulation (aspirated, plosive, bilabial, dental, etc), and the other for vowels including vertical and horizontal features (round, open, front, central, etc). In general, there are 20 features for consonants and 10 features for vowels. In their experiments, they are all used as binary features. For every language and every feature, they train two models – one for the presence of the feature and one for the absence of the feature. The training is done in exactly the same way as for acoustic models for phonemes. In the test set, every detector is tested on all the five test sets in order to examine if it is possible to detect features across languages. The classification accuracy for all features is between 93% and 95% when tested on the language that they are trained on. Otherwise, for the cross-lingual evaluation, the accuracy ranges from 83% to 88%. In order to produce the multilingual feature detectors, all the possible combinations of 2 to 5 languages are used for training using the 'Multilingual Mixed (MM)' method, which means when training MM models, data from different languages are used to train acoustic models. They use the generic term MM*n* to refer to a set of feature detectors that have been trained on *n* languages.

The same stream mechanism has been used to incorporate AF into a conventional HMM as in the work of Metze and Waibel [2002]. This requires the selection of appropriate stream weights for standard model and the feature streams. One solution is to use fixed stream weights. They have also produced stream weights using 'Discriminative Model Combination (DMC)', which is an approach that can be used to integrate multiple acoustic and/or language models into one log-linear posterior probability distribution. In this approach, different models are combined in a weighted sum at the log likelihood level. Experiments are conducted by using both fixed stream weights and the weights produced by DMC.

In their first experiment, they adopt one monolingual, one cross-lingual, and two multilingual scenarios using fixed stream weights. The English test set consists of

standard English models with English, German, MM4, and MM5 feature detectors. The MM4 detectors are trained on the four languages other than English, the MM5 on all five selected languages. The Chinese test set consists of standard Chinese models with Chinese, Japanese, MM4 and MM5 feature detectors. The MM4 detectors are trained on the four languages other than Chinese. The German and Japanese feature detectors have been chosen as the cross-lingual scenarios because the German feature detectors achieve the best average cross-lingual classification accuracy on English and the Japanese detectors show the best cross-lingual accuracy on Chinese. 0.05 is chosen as the weight for the feature streams. The feature detectors are added in the order of their classification accuracy on English, and Chinese respectively. The results show that the best performance is obtained for English by adding 9 English feature detectors. The system with the German detectors outperforms the ones with MM4 and MM5. For Chinese, the system with 4 Chinese feature detectors gives the best performance, while the cross-lingual and multilingual systems all perform lower. Then the same experiments are conducted again using DMC adapted stream weights. Though there are no improvements for the monolingual systems, the cross-lingual and multilingual systems achieve better performance using the new weights obtained from DMC. For English, the performance of the cross-lingual and multilingual systems is very close to the monolingual system, while for Chinese, the four systems achieve the very same recognition performance. In further experiments, an English recognizer with DMC selected feature detectors from Chinese, English and Spanish outperforms the system with monolingual feature detectors.

Their work suggests that incorporating cross-lingual and multilingual articulatory feature detectors into an HMM based recognition system provides significant performance improvements. The selection of optimal stream weights is very important in their systems for successful incorporation of articulatory features into the conventional HMM. They also suggest that there is still a need for a better method to select stream weights so that future systems would depend on context-dependent stream weights for sub-phonetic units. They envisage that using articulatory feature detectors will lead them to abandon

the conventional concept of the “beads-on-a-string” (cf conventional HMM) for speech modeling.

2.3.3 Syllable-based approach to speech recognition

The most prominent approach to speech recognition is the use of phones for modeling of spoken words. However, time and research have shown that phones are too small an acoustical unit to model temporal patterns and variations in continuous speech. Thus, a need exists for a new technique capable of exploiting both the spectral and temporal characteristics of continuous speech. The focus has shifted to a larger acoustic context. The syllable is one such acoustic unit whose appeal lies in its close connection to human speech perception and articulation, its integration of some co-articulation phenomena, and the potential for a relatively compact representation of conversational speech. Consequently, syllable-based modelling of speech has been widely used in not only ASR systems for languages that are considered more explicitly syllabic (e.g. Mandarin Chinese [Lee 1997] and Japanese [Nakagawa, Hanai, Yamamoto and Minematsu 1999]), but also for English-language ASR systems.

Syllable-based HMM systems

In the latter half of the twentieth century, the focal point of speech research is the problem of large vocabulary continuous speech recognition (LVCSR). Wide-ranging research and application have kept LVCSR technology at the fore-front level. Phone-based HMM technology has been used in the most successful systems to date. However, these systems still fall far short in performance. Not only is the triphone a relatively inefficient decompositional unit due to the large number of frequently occurring patterns, but also it is not suitable for integration of spectral and temporal dependencies because it spans an extremely short time duration Hamaker et al. [1997]. Thus, a need exists for a new technique capable of exploiting both the spectral and temporal characteristics of continuous speech.

In response to this need, the Institute for Signal and Information Processing (ISIP) at Mississippi State University has provided a novel approach to LVCSR, using the syllable as the fundamental acoustic unit, which is capable of overcoming these significant issues. Hamaker et al. [1997] noted that their research achieved only limited success using phone-based approaches in tasks such as Switchboard. Then their research effort has shifted and focused on exploration of the syllable as a unit of recognition in the context of the SWB task initiated by ARPA and DOD. Their success on this task involved the following three phases: development of strong training and test data sets, design of phone and syllable baseline system, exploration of new techniques to accentuate the strengths of syllable-based modeling. The most important parts were integration of infinite-duration modeling and monosyllabic word modeling.

The first goal of their system is to develop a syllable-based SWB system. As is well-known, the most critical issue in a syllable-based approach is the number of syllables required to suit the application, which is a challenge for building a context dependent syllable system. Hamaker et al. used a syllabified lexicon [Ostendorf et al. 1997] for this stage, which consisted of over 70,000 word entries for SWB and required 9,023 syllables for complete coverage of the 60+ hours training data [Greenberg 1997]. They kept the model topology for the syllable models similar to their previous context-dependent phone system, except that each syllable model was allowed to have a unique number of states. According to them, the number of states was selected to be equal to one half the average duration of the syllable, measured in 10 msec frames. In addition, the duration information of each syllable was obtained from a force alignment based on a state-of-the-art triphone system. The training procedure for syllable models was similar to their previous one for the context dependent phone system, minus the clustering stage.

Since only a limited syllable coverage was achieved in the baseline system, their second goal is to develop a mixed phone and syllable system – a hybrid system to handle words not covered by the syllabary. Firstly, they chose the syllables from the syllabary that occurred at least 20 times in the training database for training. In total a set of 2,419

syllables was selected. Then each unique word in the training database can be classified into three categories – syllable-only, phones-only, and mixed groups. Unfortunately, many models in this system were poorly trained. In order to solve this problem, they built of a system consisting of 800 most frequently used syllables and word-internal context dependent phones. Interestingly, these 800 syllables covered almost 90% of the training data, leaving the remaining 10% replaced by their underlying phone representation.

Their system, which used 632 context independent syllables with 200 monosyllabic words, context dependent phones and a finite-duration topology for testing, has achieved 49.1% of word error rate, which has outperformed their previous context independent monophone system with word error rate 62.3% and context dependent phone system with word error rate 49.8%.

As an addition to the SWB task, Ganapathiraju et al. [2001] also developed a speaker-independent alphadigit recognition system applicable to many telephony applications. The corpus they chose was the OGI alphadigit corpus [Noel 1997]. According to them, it consists of approximately 3,000 subjects, each of whom spoke some subset of 1,102 phonetically-balanced prompting strings of letters and digits. All experiments were performed on a training set of 51,545 utterances from 2,237 speakers and evaluated on 3,329 utterances from 739 speakers. They have developed three systems: cross-word triphones, word-internal triphones and context-independent systems. Evaluation on the OGI alphadigit data confirmed the gain they achieved by using syllables as a replacement for triphone units. The word error rate for syllable-based system is 10.4%, an absolute 2.9% decrease in the word error rate for a cross-word triphone system.

The syllable seems to be an intuitive unit for representation of speech sounds. Without much difficulty listeners can identify the number of syllables in a word [Greenberg 1999] and, with a high degree of agreement, even the syllable boundaries. Perhaps such behaviour makes the syllable a more stable acoustic unit for speech recognition. The stability and robustness of syllables in the English language are further supported by

comparing the phone and syllable deletion rates in conversational speech in HMM systems. Ganapathiraju, Goel, Picone, Corrada, Doddington, Kirchoff, Ordowski and Wheatley [1997] showed in their experiment that the deletion rate for phones was 12%, compared to a deletion rate for syllables of less than 1%. This suggests that syllable modeling becomes a necessary component in phone-based systems for dealing with the high phone deletion rate problem in conversational speech. In addition, the previously mentioned SWB and alphadigit systems also confirm the use of an acoustic unit with a longer duration, e.g. a syllable-sized unit, may facility the exploitation of temporal and spectral variations simultaneously. Syllables are obviously useful processing units since they yield significant improvements in speech recognition performance while maintaining a manageable level of complexity.

Though Hamaker et al. and Ganapathiraju et al.'s work in syllables are promising, they have just scraped the surface of syllables' potential. Their systems are clearly deficient in a number of areas, including the representation of ambisyllabics in the lexicon and the integration of syllables and phone models in a mixed word entry. Their current systems do not exploit the temporal modeling advantages inherent to the syllables.

Another significant difficulty with syllable-based HMM systems is the severe sparse data problem due to the uneven distribution of syllables in natural speech. Ganapathiraju et al. and Hamaker et al. have shown that syllables are unevenly distributed in speech data, with the result that some syllables appear frequently and some appear very infrequently. Indeed, it is also shown that in the TIMIT database some syllables that appear in the testing data are not seen at all in the training data. Usually, the problem has been avoided by only building models of a subset of all syllables, i.e. those which have high enough frequencies in the training data. However, it is recognized that this reduced the potential benefits that could be gained from any syllable-based speech recognizer [Stewart et al. 2002]. So far, it seems not possible to solve this problem totally for HMM-based systems, since HMM technology is not based on abstract linguistic models but based on acoustic details.

Other syllable-based systems

Another syllable-based speech recognition system is presented by Chang [2002]. He stated that current-generation speech recognition systems assume that words are readily decomposable into constituent phonetic components ("phonemes"), while a detailed linguistic dissection of state-of-the-art speech recognition systems indicates that the conventional phonemic "beads-on-a-string" approach is of limited utility, particularly with respect to informal, conversational material. The study indicates that there is significant gap between the observed data and the pronunciation models of current speech recognition systems. In addition, it also shows that many important factors affecting recognition performance are not modeled explicitly in these systems. This is also the motivation of the work presented in this thesis.

Motivated by these findings, he analyzed spontaneous speech with three important, but often neglected, components of speech (at least with respect to English automatic speech recognition). These components are: stress accent, articulatory-acoustic features (AFs) and syllables. They are also the key components motivating the work presented in this thesis. Chang provided an alternative approach to speech modeling, one in which the syllable assumes preeminent status and is joined to the lower as well as the higher tiers of linguistic representation through incorporating prosodic information such as stress accent. According to his experiments using concrete examples and statistical results, a systematic relationship between the realization of AFs and stress accent in conjunction with syllable position is found, which can be used to provide an accurate and parsimonious characterization of pronunciation variation in spontaneous speech. Moreover, he also developed an approach to automatically extract AFs from the acoustic signal.

Based on the results of these studies, he proposed a syllable-centric, multi-tier model of speech recognition, which explicitly relates AFs, phonetic segments and syllable constituents to a framework for lexical representation, and incorporates stress-accent information into recognition. Finally he developed a test-bed implementation of the model using a fuzzy-based approach for combining evidence from various AF sources

and a pronunciation-variation modeling technique using AF-variation statistics extracted from data. Experiments using a limited-vocabulary speech data showed the great advantage of incorporating AF and stress-accent modeling within the syllable-centric, multi-tier framework, particularly with respect to pronunciation variation in spontaneous speech. Furthermore, he has extended his system to provide speech-based interface to specific applications such as telephone directory assistance and internet information retrieval.

In recent years, there have been more and more studies incorporating syllable-level information in English-language speech recognition systems. Wu [1998] also builds speech recognition systems incorporating information from syllable-length time scales and Jones, Downey and Mason [1997] report a syllable-based speech recognition system for a British English corpus.

In fact, there has been an increasing interest in the notion that the syllable may be the binding unit of speech around which information at various linguistic tiers is organized [Greenberg 1999], in contrast to the traditional phonetic-segment perspective of spoken language. This view echoes work in the linguistic study of the phonological organization of speech [Goldsmith 1990]. From linguistic analysis and the experiments from the above systems, it is known that syllables are not only more stable in the speech signal compared to phonetic-segments and but also play an important role in speech perception. Moreover, much prosodic information that is important for word recognition, such as stress-accent level and speaking rate, is directly tied to the syllabic representation of speech. All this suggests that syllable-level information should have a significant impact on speech recognition performance and it should be beneficial to model such syllable-related factors explicitly in automatic speech recognition systems.

2.3.4 Alternative approaches

Research in automatic speech recognition by machine has been carried out for almost four decades. During this period, a huge amount of progress has been achieved. It is

worthwhile to briefly review some other research highlights, for example, the knowledge-based approach, the artificial neural networks approach and the pattern matching approach.

Knowledge-based systems use explicit rules formulated for the entire speech recognition problem based on existing knowledge about speech production and perception. The development of this approach accompanied developments in artificial intelligence, particularly the introduction of the Expert System technique. Artificial neural networks (ANN), modeled on the analogy with human brain [McCullough and Pitts 1943], are constructed from interconnected nodes or processing elements (neurons) and learn patterns by experiencing samples of these patterns. In ANN, usually there are three types of layers: the input layer (receives information from the external environment), the output layer (communicates the network's decisions to the external environment), and the hidden layer (one or more, which communicate with each other and the output layer). Groups of nodes arranged in these layers. And these processing elements can be arranged differently. There are three standard ANN topologies or architectures [Rabiner and Juang 1993; Lippmann 1989]. ANNs have failed to perform better than other methods when confronted with problems such as time processing, context information or large vocabularies. However, they can achieve state-of-the-art results for speaker-independent large-vocabulary speech recognition task [Robinson, Hochberg and Renals 1996]. In the pattern matching approach, each pattern is represented by a sequence of vectors evenly distributed along the time axis [Mariani 1989]. These vectors can be the result of various speech signal processing techniques, for instance, linear prediction coding, (fast) Fourier transformation or cepstral analysis. In the recognition process, the distance is computed between the incoming "to be recognized" pattern and each of the reference patterns.

2.4 TIMIT - speech database description

The speech material used in this research work is from the TIMIT speech database. Since it was first released in 1988 by Texas Instruments and the Massachusetts Institute of Technology, TIMIT has been widely used for the development and evaluation of speech

recognition systems [Garolfo, Lamel, Fisher, Fiscus, Pallett and Dahlgren 1993]. There are 6,300 sentences in the database with 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States.

There are three different types of sentences in the database, which are designed for different recognition purposes: dialect sentences, phonetically-compact sentences and the phonetically-diverse sentences. The intention of designing the dialect sentences is to emphasis the dialectal differences of the speakers. There are two dialect "shibboleth" sentences uttered by all 630 speakers. The phonetically-compact sentences are created in such a way as to provide a good coverage of pairs of phones from each speaker. 450 phonetically-compact sentences are included in TIMIT, and each sentence is uttered by seven speakers, with five such sentences spoken by each speaker. The 1,890 phonetically-diverse sentences are chosen from existing text sources and the Playwrights Dialog with the criterion of maximizing the variety of allophonic contexts. Each sentence is uttered by only one speaker with three of these sentences read by each speaker.

The speech material in the database has been divided into two parts: training and testing. 20-30% of the corpus is used in the test set, while the rest is used in the training set. There are different speakers for training and testing and no speaker and sentence can appear in both sets. The test set must cover all the phonemes, preferably in multiple contexts. Finally, all the dialect regions are represented in both subsets with at least one male and one female speaker from each dialect.

There is a core portion in the test data which contains 24 speakers: two male and one female from each dialect region. 192 phonetically-compact and phonetically-diverse sentences have been used as the core test material. A more extensive complete test set is available as well. It includes 168 speakers and 1,344 utterances. The remaining part of the database is used for training. The superimposed division of the database into training and testing makes evaluations possible between different systems, since the same speech material for training and testing has been used for different research projects.

TIMIT provides four files associated with each utterance:

- a speech waveform file (.wav)
- an orthographic transcription of the words the speaker said (.txt)
- a time-aligned word transcription (.wrd)
- a time-aligned phonetic transcription (.phn)

Other supporting documents are also available for the TIMIT database, such as the documents describing the phonetic conventions used in the transcription, speaker attributes, etc.

Since apart from the speech waveform files, time-aligned phonetic transcriptions are extensively used in this research, it seems appropriate to discuss them in more detail.

2.4.1 Time-aligned phonetic transcription

Time-aligned phonetic transcriptions are extremely useful for the development and evaluation of various speech recognition systems. That is the reason why the designers of the TIMIT database make great efforts to provide these valuable files. There is a set of labels used in TIMIT which includes not only phoneme labels, but also some allophones [Zue and Seneff 1988]. The complete label set is in Appendix A. In it, each plosive is described as the combination of a closure and a release. There are also some syllabic nasal and liquid allophones. A single label is used to represent each diphthong and each retroflexed vowel. Additionally, four allophones are used for reduced vowels. There are also some minor allophone distinctions in it as well.

In most cases, the criteria for boundary assignment are very well defined. But in the problematic cases, a systematic approach is used. For example, in the case of a prevocalic stop followed by a semivowel (e.g. “trend”), the unvoiced portion of the semivowel is always absorbed into the release portion of the stop, and the transition between the semivowel and the adjacent vowel is slow and continuous. Therefore, a consistent rule

for the latter is first adopted which assigns one-third of the vocalic region in the signal to the semivowel and the rest to the vowel, leaving the problem of gemination at word boundaries unsolved.

In order to solve the above problem, the phonetic sequence is produced manually based on careful listening and visual examination of the speech signal. Since manual labelling proves too time-consuming for such a large speech database, an alignment program is used to automatically align the speech waveform with the phonetic sequence. The program assigns 5 ms chunks of speech to one of the five broad-class labels (sonorant, obstruent, voiced consonant, nasal/voicebar, silence) using a non-parametric pattern classifier. A search strategy guided by some phonetic rules is used to align the output and the phonetic transcription. If needed, algorithms based on phonetic context are used to provide further segmentation. Finally, transcription errors are corrected manually again by critical listening and visual examination of the relevant portions of the utterance.

In spite of their enormous efforts, the authors admit that errors undoubtedly exist. Using a program to automatically produce alignment may have increased the objectivity of the judgement, but may introduce errors. In a formal evaluation, 75% of the automatically generated boundaries are within 10 ms of those entered manually, which are very good results for an alignment program. But still 25% of incorrect boundaries exist.

2.5 Earlier doctoral work as background

Earlier work [Edmondson et al. 1996; 1999] has established the notion of pseudo-articulatory representation (PAR). Briefly, the PAR is based on the linguist's conception of binary distinctive features – a set of parameters which are both language and speaker independent, and which also categorize speech sounds in terms of a very abstract and atemporal model of the vocal tract. The PAR takes this abstract model and injects realism in two ways – by making the feature values continuous where appropriate, instead of binary, and by providing values as a continuous function of time, instead of segment by segment.

Both Iles [1995] and Iskra [2000] have used PARs to explore their feasibility in speech synthesis and recognition. While Iles worked with formants, Iskra worked with cepstral coefficients. Their work has demonstrated the value of the general approach, with Iles managing to show recognition feasibility using formants in addition to his main work on synthesis, and Iskra demonstrating some synthesis in addition to recognition. Iskra has looked specifically at the use of PARs to handle the speech recognition task.

In her work, the recognition system consists of two stages. The first stage is responsible for the transition from the acoustic representation of the incoming signal to the pseudo-articulatory representation with feature trajectories available as a function of time. The second stage is the movement directly from the pseudo-articulatory representations to the phonetic level of description and produces a sequence of phone labels. The recognition results can be regarded as promising. However, she was unable to develop her approach any further because of the difficulty in recovering articulatory information irrespective of any supposed segment labels. Thus, her system performed still far short of its potential and it can be improved in a number of ways. The work presented in this thesis can be regarded as one of the possible ways for improvement.

2.5.1 Pseudo-articulatory features as a natural approach to speech recognition

Statistical approaches commonly use phone-based models (sometimes context-dependent phone models such as diphones or triphones, or syllable models) for the convenience of processing, while a feature-based approach offers a more linguistically justified solution. Using diphones, triphones or syllables, as the basic units for training hidden Markov models require a large speech database and substantial computing power. Though these are all available nowadays, context-dependent models and syllable models are increasing in popularity. Moreover, HMMs have little relevance to any articulatory process taking place during speech production or the vocal tract. They convert the problem of speech recognition to the computation of probabilities. Features, on the other hand, have deep roots in linguistic theory. They can be defined as the building blocks of particular sounds

and can be mapped onto their respective articulatory configurations. They can be justified by finding reference to articulatory phenomena taking place during speech production.

Since coarticulation can be described as certain effects of a particular sound extending onto the neighboring sounds, these effects can be very well accounted for in terms of features. In another word, there is no need for rigid segmentation for phone boundaries, since features are allowed to overlap. Consequently, the problem of coarticulation is also more readily tackled with features. In HMM approach, the basic training units can be carefully chosen to incorporate these effects, which may bring results, but in the end leaves us with certain speech units which are unjustified in articulatory terms. The feature-based approach, on the other hand, avoids imposing strict phoneme boundaries and suffices to allow features to overlap onto the neighboring sounds. Since this is exactly what happens at the articulatory level, this approach is just a reflection of what take place during speech production.

Features are not just descriptive of articulation (phonetic features), but are also used to represent linguistic processes (phonological features). Phonological features make it possible to generalise some language phenomena and find the underlying grounds for them. Also phonological features play an important role in linguistic models. In phonology features commonly adopt binary values whereas there is more space for flexibility within phonetics. Since pseudo-articulatory features possess the level of abstraction of phonological features and the descriptive power of phonetic features, they have bigger scope and become more general. Because of their generality, modelling speech in terms of features is a viable method of dealing with the articulatory-acoustic mapping problem. Slight changes in the acoustic signal are mapped onto the same features. Features become, therefore, a successful way of abstracting from the variance present in the acoustic signal and of deriving a more general underlying representation of speech.

It seems not appropriate to pursue an approach just because the current computer technology and facilities can provide all the power it needs, especially if this approach has nothing to do with the actual processes taking place in the real world. An approach which takes advantage of these processes is bound to succeed in the long run since it leads to deeper understanding of human behaviour and all it tries to do is to build a model of these processes which offers the best possible fit. This is the key motivation for this work presented in this thesis. Any progress in this area, therefore, contributes to the general science of phonetics and linguistics and constitutes a part of the bigger enterprise to understand human behaviour.

3 ARTICULATORY-ACOUSTIC MAPPING

This chapter describes the mapping procedure which will subsequently serve as the basis for recognition. First of all, the articulatory-acoustic mapping problem is illustrated. Different articulatory strategies can lead to what is perceived as the same acoustic signal. Then the proposed solution – a pseudo-articulatory level of description – will be described in more detail. Next speech material and formant parameter representations are provided. Then the derivation of pseudo-articulatory features is introduced and a pseudo-articulatory description for vowels is provided which is suitable for the current speech database and the application. Following this, an explanation of the statistical technique of multiple regression analysis and the choice of the regression equations are stated. Next the analysis step is adjusted for the eventual recognition processing. When everything is ready, the mapping is produced for vowels using multiple regression analysis. A satisfactory vowel model is obtained. For some of the consonants, the formant parameters provided by the TIMIT database are unrepresentative because they are influenced by neighbouring sounds. Accordingly, we choose the consonant model obtained by Iskra [2000] using cepstral coefficients as the phoneme reference.

3.1 Articulatory-acoustic mapping problem

The articulatory-acoustic mapping problem is a many-to-one mapping problem. Different articulatory configurations can result in the same phoneme. It is proposed that there exists an abstract level of representation where the articulatory variance is overlooked. From the above description it may sound as if this abstract level is equivalent to the phonological level (i.e. phonemes). It will be argued here that the phonological level is too abstract; the phonetic level, on the other hand, can be too detailed and too close to the acoustic signal. Therefore, a new type of description has been proposed in terms of *pseudo*-articulatory features. First of all, the articulatory-acoustic mapping problem will be presented. Then the proposed solution – a pseudo-articulatory level of description – will be described in more detail.

The articulatory-acoustic mapping as a many-to-one mapping can be illustrated by the ventriloquist effect. Although a ventriloquist uses alternative articulatory strategies, because of the need to suppress the articulatory gestures which are visible to the outside world, the produced speech is perfectly intelligible [MacKay 1987]. Another example is of speakers trying to speak with an object in their mouth such as a cigarette or a sweet. These objects make it impossible to use the usual gestures in the production of speech by blocking certain parts of the vocal tract and, for instance, preventing the tongue from reaching its target configuration. Even so, the speech produced with such obstacles is intelligible. That means that articulatory gestures which in textbooks are presented as leading to the production of a particular sound, are not (all) necessary to achieve that effect.

It is clear from the above that the same phoneme can be produced using different articulatory configurations. Turning this problem round, for a given sound it is effectively impossible to be 100% sure what its detailed articulatory configuration is. There is no one-to-one relationship between the two. In order to make speech intelligible all these differences have to be overlooked at some point and all these levels unified, leading to a single abstract representation. The unifying representation proposed here is that of pseudo-articulatory features. On the one hand, they are like phonological features in being distinctive and abstract, and serving primarily the purpose of classification. On the other hand, they resemble phonetic features in taking a larger range of values than binary and thus accommodating more phonetic detail. They represent linguists' generalisations and idealisations of articulation and the articulator positions. As such they serve the purpose of an intermediate abstract representation providing the link between the articulatory description of sounds and their acoustic correlates.

It is a well-established fact that the phonologist's view of articulatory activity is idealised and abstract. This view does permit linguistic generality to be distinguished from both the articulatory reality and the acoustic reality. Thus, for example, my tokens of "research student" will differ from each other, and from other persons'. But the

abstraction/idealization permits them all to be categorized as tokens of the type RESEARCH STUDENT. This cannot be overstated – the linguist's insight permits speaker and language independent specification of the vocal tract configuration during speech. This must be so because we can understand one another despite the fact that we do differ in the production of tokens (both from time to time by a single individual, and across a population). However, the linguist's vocal tract specifications look very much like articulatory specifications, and this can be exploited [Edmondson et al.1999].

Pseudo-articulatory representations (PARs) are derived from the phonologist's specifications by identifying a sub-set of distinctive features presumed to be especially significant and assigning them values on a continuous scale from 0 to 100. This has the effect of mapping many binary features onto somewhat fewer continuously variable features; in both cases the assumption is that the vocal tract configuration targets for linguistic use can be specified in the multidimensional space defined by abstract features. The specification looks articulatory so we refer to the representation of vocal tract targets in these terms as *Pseudo-articulatory representations*.

The utilisation of PARs requires that distinctive features based on the descriptions of vocal tract configurations are matched with evidence of the actual acoustic activities. In reality, articulations are variable for non-linguistic reasons, especially between speakers. The ventriloquist problem in speech processing actually provides the solution: we do not need to know exactly where each bit of the articulator is because all we need are the details captured pseudo-articulatorily or linguistically controlled. The promise of PARs is that they represent the linguistic detail only – they convert a vocal tract specification into an articulatory target (and vice versa). Other useful non-linguistically controlled details of the real articulation are not captured or expressed by PARs. However, some aspects of speaker characterization may be specified in linguistically relevant terms, and thus included in PAR space. On the whole, we believe the utilisation of PARs provides both a linguistic handle on the acoustic evidence of articulation and a way of isolating other

aspects of speech behaviour which are relevant (e.g. speaker identification), but not linguistic.

Moreover, the utilisation of PARs recovers non-segment information from speech, which gives more opportunities for further processing. It is possible to make the derivation directly from PAR space to the descriptions of phonemes, or make the processing from PARs to an intermediate level, which may provide important additional information for the derivation of phonemes. In reality, articulator movements are relatively slow and the produced speech consists of phones and transitions between adjacent phones. This is also one motivation of this research work – to recover what happens in real speech linguistically. The previous work [Iskra 2000], which recovers phonemes directly from PARs, provides only a sequence of phones and their durations without any other relevant information. Additionally, since it is a coarse processing without knowing the locations of the transitions and the wanted targets, this approach performs comparatively poorly and is a step to be avoided. In contrast, the intermediate level, which can provide transition and target information, seems very valuable for further processing. A novel articulatory description of syllable model has been explored and used [Edmondson and Zhang 2001; 2002] to provide the necessary and important intermediate information, which is very crucial for the derivation of phonemes in the further processing. Since the new approach provides transition and the wanted target information, it is a finer processing than previous one and is expected to perform better.

Finally, it is also one of the challenges of this work to demonstrate that the articulatory-acoustic mapping can work both ways. The two-way mapping seems plausible if it is performed at an abstract enough level which can discard all the variance present in the acoustic signal. Current statistical approaches to speech recognition concentrate on the acoustic parameters only. The models are trained on as much speech data as possible in order to accommodate all the variance in the acoustic signal. Most of the time no reference is made to any phonetic or linguistic knowledge, which has the potential of overlooking this acoustic variance and retrieving the more abstract underlying

representation. That is also the motivation of this work, which has adopted pseudo-articulatory features.

3.2 Background material

3.2.1 Speech data

The speech data used for the mapping processing come from a single male speaker (ID: dr1/mmrp0) from the TIMIT database. He represents the dialect region of New England. This amounts to 10 sentences: 2 dialect sentences, 5 phonetically compact ones and 3 phonetically diverse sentences.

3.2.2 Formant representation

The resonances of the vocal tract are called *formants*. Frequency is a technical term for an acoustic property of a sound – namely, the number of complete repetitions (cycles) of variations in air pressure occurring in a second. Formant frequencies are the resonance frequencies of the vocal tract. Every configuration of the vocal tract has its own set of characteristic formant frequencies. The sounds we produce can be described in terms of how fast the variations of the air pressure occur, which determines the fundamental frequency of the sounds.

The shape of the vocal tract determines formant frequency values. The vocal tract starts as a single tube in the pharynx, but separates into two branches at the soft palate, one through the mouth and the other through the nose. When the soft palate is raised, shutting off the nasal cavities, the vocal tract is a tube about seven inches long from the glottis to the lips and the principal resonances of this tube are at 500Hz, 1,500Hz, 2,500Hz, 3,500Hz, and 4,500Hz [Denes and Pinson 1993]. Depending on different physical dimensions of the vocal tract and the configuration of the articulators (such as tongue and lips), formant frequencies will change accordingly. The lowest formant frequency is called the *first formant*; the next highest frequency, the *second formant*, and so forth. Additionally, the first three formants are most prominent. Any two speakers will not

produce the same formant frequencies when they produce the same resonant sound, because of the different physical dimensions of their vocal tracts. The ratio between the frequencies, however, will be similar, leading to the perception of the same sound.

The traditional articulatory descriptions of vowels are related to the formant frequencies. The first formant frequency decreases as a speaker goes from low vowels to high vowels. Vowel height is inversely related to the first formant. The second formant frequency decreases as a speaker moves from the front vowel in “heed” to the back vowel in “who’d”. But the correlation between the degree of backness of a vowel and the second formant frequency is not as consistent as that between the vowel height and the first formant frequency. However, there is a better correlation between the distance between the first two formants and the degree of backness: the two formants are far apart in front vowels and close together in back vowels.

The degree of lip rounding also affects the formant frequencies. Briefly, lip rounding is inversely related to the frequencies of the higher formants: as sounds become more rounded, the frequencies of the higher formants decrease and vice versa. But the situation is complicated because the effect is greater in the second formant for back vowels, and in the third formant for front vowels.

“The acoustic structure of consonants is usually more complicated than that of vowels. In many cases, a consonant can only be said to be a particular way of beginning or ending a vowel” [Ladefoged 1993]. This is particular true of voiced stop consonants. Each of these consonants has a rapid movement of the lips or tongue before or after another sound such as a vowel. The resonances of the vocal tract, the formants, are being produced while the stop closure is being formed. The voiced stop consonants are distinguished by the movements of the second and third formants. In general, if a syllable or a word starts with [b], then the second and third formants are increasing in frequency. If the second formant has only a small movement and third formant falls, then the sound is [d]. If the second and third formants are close together, then it is a [g]. The voiceless counterparts ([p], [t],

[k]) are made with same gesture of the tongue or lips. Consequently, the movements of the formants are similar for these two sets of consonants [Ladefoged 2001], though for the voiceless stops, there is no vibration of the vocal folds. At the beginning of a word, they are mainly distinguished by the frequencies of the bursts of noise produced as the stop closure is released.

The liquids and glides are also, like vowels, produced with a source at the glottis, but there is a constriction in the oral part of the vocal tract. The constriction is narrow enough for the airflow to create resistance which increases acoustic losses and that, in turn, increases the bandwidth of certain formants. As a consequence some peaks, especially the second formant and the third formant, are less salient than others. And the liquids ([l], [r]) and the glides ([w], [y]) have their own formant patterns. For [w], at the beginning of a word, there is a noticeably sharp rising second formant, a comparatively steady rising first formant and a steady third formant. [y] has a visible rising first formant, a falling second formant and a drop in the third formant. [l] and [r] differ from the first two sounds because of the abrupt changes in the articulation and normally show irregularities at high frequencies (the third formant). [l] has a low-intensity formant at a very low frequency, another low-intensity formant at about 1,500 Hz, and a distinct break in the pattern before a vowel. On the other hand, [r] is characterized by the very low third formant frequency. If there is an [r] in a word, the third formant will be very low, usually below 2,000 Hz [Ladefoged 2001].

Nasals ([m], [n], [ng]) are like vowels and approximants since they can be characterized largely in terms of their formant frequencies. But since nasals are made by allowing the sound to come out of the nose instead of the mouth, this affects the relative formant amplitudes (loudness). The three consonants usually have very low first formant frequencies. But the third formant for [m] is slightly higher than that of [n] and the second and the third formants are close together in [ng]. They have similar formant movements to the corresponding stops.

Voiceless fricative is produced by the friction, the resistance to air as it rushes through a narrow gap. Though they do not have formants, they have formant movements. The voiceless fricatives [f] and [th] have energy over a wide range of higher frequencies. They are distinguished from each other mainly by the formant movements. The other voiceless fricatives, [s] and [sh], have more energy with [s] being in a higher frequency range and [sh] in a lower range. The corresponding voiced fricatives ([v], [dh], [z], [zh]) have similar energy distribution, but with additional clear formant resonances. The consonant [h] has noisy forms of the formants in the adjacent vowels [Ladefoged 2001].

Ladefoged summarized the acoustic correlates of some articulatory features in table 3.1. He also noted that these descriptions should be regarded only as rough guides. The actual acoustic correlates depend to a great extend on the particular combination of articulatory features in a sound.

Another widely used parametric representation of speech signal today is cepstral coefficients. These are based on the Fourier spectrum. Iskra [2000] used cepstral coefficients for all the experiments in the first step analysis and testing set using the same speech data uttered by the same speaker. Furthermore, all the available data from that speaker were used by Iskra in the first step analysis (training processing). In general, she used the same speech data for analysis (training) and testing by using the same parameters – cepstral coefficients. Since this research work is based on the same speech data uttered by the same speaker, in order to compare our results with hers for evaluation, we choose to work with formant parameters to set the system up and also for the testing set in order to avoid a too closed testing. Moreover, because of the correlation between formant parameters and the features: “high”, “back”, “round”, formant parameters will capture the necessary and important acoustic information for the derivation of the values of these features. Hence, we will choose formant parameters as the acoustic parameters instead of cepstral coefficients.

Voiced	Vertical striations corresponding to the vibrations of the vocal cords.
Bilabial	Locus[2] of both second and third formants comparatively low.
Alveolar	Locus of second formant about 1,700-1,800 Hz.
Velar	Usually high locus of the second formant. Common origin of second and third formant transitions.
Retroflex	General lowering of the third and fourth formants.
Stop	Gap in pattern, followed by burst of noise for voiceless stops or sharp beginning of formant structure for voiced stops.
Fricative	Random noise pattern, especially in higher frequency region, but depend on the place of articulation.
Nasal	Formant structure similar to that of vowels but with nasal formants at about 250, 2,500 and 3,250 Hz.
Lateral	Formant structure similar to that of vowels but with formants in the neighborhood of 250, 1,200 and 2,400 Hz. The higher formants are considerably reduced in intensity.
Approximant	Formant structure similar to that in vowels, usually changing.

Table 3.1 Acoustic correlates of consonantal features [Ladefoged 1993]

Finally, it is necessary to consider the limitations of the use of formant data to provide acoustic parameters for voiceless consonants. (There is no problem with voiced consonants. For example, the burst produced during the release of a voiced plosive contains recognizable frequencies.) The formant tracker *xwaves* used in the work presented in this thesis produces frequencies for the unvoiced plosive bursts. The turbulence of fricatives usually shows up on speech spectrograms as a fuzzy segment. The formant tracker *xwaves* also produces energy frequencies for fricatives. Figure 3.1 shows some example spectrograms with overlaid formant trajectories provided by

[2] The apparent point of origin of the formant for each place of articulation is called the locus of that place of articulation.

xwaves. Strictly, there are no formants in the central parts of unvoiced plosives and fricatives. According to the manual [ESPS 1996], *xwaves* produces a standard set of frequencies under these circumstances. This may introduce errors into the processing. This is the major limitation of the use of formant data as acoustic parameters[3].

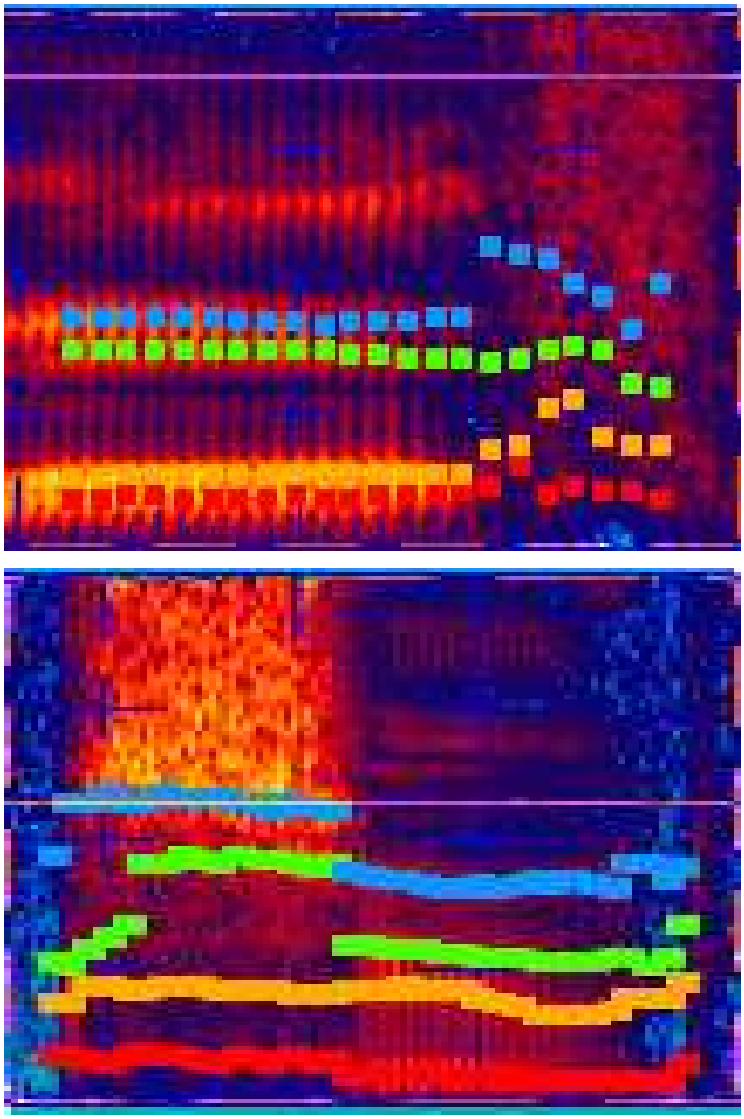


Figure 3.1 Example spectrograms with overlaid formant trajectories for word “off” and word “suit” (speaker ID: dr1/mmrp0)

[3] See for example the high values in the confusion matrix for /t/, /k/ and /s/ in Appendix E.2.

3.2.3 Feature description

Apart from the parametric representation, a description in terms of features is needed too. The same feature descriptions for vowels are used as those defined by Iskra [2000]. Iles [1995] intended to use distinctive features as these provide a minimum set of descriptors to distinguish between phonemes. The binary values were, however, not sufficient to model the transitions between segments in speech synthesis. Therefore, a set of descriptive phonetic features was used and most of them were ascribed continuous values between 0 and 100. Iles defined all the vowels in terms of four features: high, back, round and tense, and derived their values from phonetic textbooks [Atkinson, Kilby and Roca 1988; Fromkin and Rodman 1998; O'Conner 1974].

In linguistics the distinctive features which specify the vocal tract in the binary model are idealized. The model assumes that a specification only has two possibilities (present or absent), but nothing in between. Though such specification seems straightforward (e.g. a sound is either voiced or not), in reality it may not be adequate. For tongue height, or lip-rounding, or tongue body front/back position, as well as for some other features, the notion of binary values is obviously imposed. The idealization is justified since within the overall set of possible configurations the system of binary distinctions can provide the accounts for the production and perception of distinctly different speech sounds [Edmondson et al. 1999]. However, such binary distinctive features cannot create enough descriptive and classificatory space to accommodate all the sounds and account for coarticulatory phenomena in phonetics.

Since the tongue is a very important articulator in the production of vowels, the features describing its position lend themselves very well to the classification and description of this group of sounds. Consequently, the features "high" and "back" are frequently used to characterize vowels [Laver 1994; MacKay 1987]. A range from 0 to 100 was chosen to create enough descriptive and classificatory space to accommodate all the sounds and account for coarticulation. Another feature "round" was added to the first two corresponding to the position and the shape of the lips. This feature can also be inversely

related to higher formants. In order to broaden the descriptive space one more feature was added, “tense”. This feature is mostly presented as distinctive. It corresponds to the amount of muscular tension involved in the production of a sound. Long vowels are described as tense vowels while short ones are lax. This feature has complex acoustic evidence which can be measured in terms of the parameters of the voice source [Ni Chasaide and Gobl 1997]. Features round and tense also were given values between 0 and 100. Since these features are neither phonetic nor distinctive, they will be referred as *Pseudo-articulatory features*. Here follows a description of the features used:

High: continuous feature; takes values between 0 and 100; indicates the tongue position during the production of a sound; 0 corresponds to the tongue being low in the mouth and 100 to the tongue being high.

Back: continuous feature; takes values between 0 and 100; indicates the tongue position during the production of a sound; its value takes 0 if the tongue is in its front position and 100 in the back position.

Round: continuous feature; takes values between 0 and 100; indicates the degree of lip rounding during the production of a sound; 0 corresponds to the lips being unrounded or spread and 100 to the lips maximally rounded.

Tense: continuous feature; takes values between 0 and 100; indicates the degree of muscular tension and articulatory effort present during the production of a sound; 0 corresponds to “lax” and 100 to “tense”.

Iskra [2000] provided new feature definition values for vowels based on the relationship between formant measurements and vowel feature values for American English, as shown in table 3.2. Ladefoged [1993] provides average values of formant measurements for American English vowels. It made the classification more objective to have more reference points (formants), which are physically quantifiable and measured for

American English, which is the variety of English used in the TIMIT database. Using these reference points, the first formant and the second formant were translated into features high and back ranging from 0 to 100. For the feature round, five degrees of roundness were discriminated, and three different values were used for the feature tense. The table below contains new feature values for vowels. The vowel symbols used in the table are taken from TIMIT [Garolfo et al 1993]. The examples are given in their orthographic transcription with the letters corresponding to a particular sound represented in capitals.

vowel	high	back	round	tense	example
ae	12	60	50	20	sAt
ax	50	50	50	10	datA
ah	24	80	50	10	bUt
eh	40	46	50	10	bEt
iy	94	7	0	90	hEEd
ih	70	31	50	10	knIt
uw	88	81	100	90	tOO
uh	60	80	90	10	gOOd
ao	32	95	75	90	cAUght
aa	8	90	50	90	fAther

Table 3.2 New feature values for vowels [Iskra 2000]

3.2.4 Multiple regression analysis

The mapping between the features and the acoustic parameters is done using multiple regression analysis as implemented by Microsoft Excel [Excel 1997] or Matlab [Matlab 2001]. Excel provides a function which uses “the least squares” method to calculate a straight line that fits the data best. Then it returns an array of coefficients describing the line.

The equation used in previous experiments is as follows [Iles 1995]:

$$y_i = a_0 + a_1 h + a_2 b + a_3 r + a_4 t + a_5 hb + a_6 hr + a_7 ht + a_8 br + a_9 bt + a_{10} rt + a_{11} h^2 + a_{12} b^2 + a_{13} r^2 + a_{14} t^2$$

Equation 3.1

where y_i were the successive parameters (cepstral or formant), a_i were the regression coefficients (including the constant a_0) and *h*, *b*, *r*, *t* were the values for "high", "back", "round", and "tense" for each vowel.

In Iles' work [1995], both the cross-products and squares of the features were used in the recognition equation 3.1, requiring 15 parameters to be estimated from the data. Iskra [2000] discarded the square items (equation 3.2). Initially, both equations were tried in this work. As the simpler equation 3.2 produced coefficients of determination r^2 (see page 50) higher than Iles had produced with the longer equation 3.1, equation 3.2 was used in all remaining work.

$$y_i = a_0 + a_1 h + a_2 b + a_3 r + a_4 t + a_5 hb + a_6 hr + a_7 ht + a_8 br + a_9 bt + a_{10} rt$$

Equation 3.2

In order to increase the reliability of the mapping, it seemed desirable to increase the number of data points beyond the very small scope of vowels. A possible candidate was the set of diphthongs because of their similarity to vowels. However, as diphthongs are sometimes described as a movement from one vowel to another, it was clear that the feature values for the starting point might be considerably different from the end point. The end point of a diphthong is more aimed at than actually reached [Ladefoged 1993]. A solution to the problem of feature description for diphthongs came from MacKay who claimed, "Since diphthongs are nothing but a vowel of changing timbre, virtually any combination of starting points and finishing points is theoretically possible. Indeed, non-standard dialects of English and many other languages have a great variety of

diphthongs" [MacKay 1987]. The work of Hatzis and Green [2001] supported this claim using an artificial neural network to obtain intermediate positions for diphthongs on the vowel quadrilateral when testing gliding from /i/ to /ae/ to /a/ to /u/. Atkinson, Kilby and Roca [1988] also agreed that "spectrograms of diphthongs show the corresponding movement from one formant region to another". Therefore, a set of diphthongs was added, described as mid-points between any two vowels. Such mid-point data has been used not only by Iles [1995], but also by Iskra [2000]. Hence the work presented in this thesis continues using vowels and such diphthongs (mid-points between any two vowels). This increases the number of data points to 48. However, strictly the number of degrees of freedom is unaltered. The statistical tests used by Iles and Iskra do not appear to have allowed for. The present work is unaffected because statistical tests have not been used.

3.3 Material and methods

3.3.1 Analysis step

It is very important to adjust the analysis step and the window size to fit better the recognition approach here. Though the recognition procedure will be discussed in great detail in the following three chapters, it is necessary to mention at this point that recognition takes place with a fixed window sliding along the speech signal and for every formant vector (frequency, amplitude, bandwidth) four best matching pseudo-articulatory values for high, back, round, tense are found. Then the syllable structures are recovered using smoothed pseudo-articulatory trajectories. Finally, the recovered syllable patterns are used to find the best matching sequence of phonemes from the reference phone inventory.

Looking at the durations of the recognized phones it was evident that they extended over at least two records (5 ms each) in the preliminary experiments [Iskra 2000]. The results, therefore, should not be affected negatively if the analysis step is increased to 10ms. Using too fine a grid has the disadvantage of producing "noise". Even the smallest changes in the signal are detected in the analysis and, when passed onto the recognition

process, often instead of improving the results provide a large amount of insertion errors to the recognition results. If the analysis is too coarse, on the other hand, there is a chance of overlooking acoustically important information. The trade-off between too fine and too coarse search is an important one and the values have to be chosen with care.

The speech files are analysed with a step value of 10 ms which makes the process significantly more efficient. The window of 10 ms is incremented every 10 ms each step. The derived pseudo-articulatory trajectories in the recognition process are smoother. The recognition results in different stages show less unnecessary variation, or "noise", without losing the important information or missing out phone segments.

As a result of this assessment, the value of 10 ms for the analysis step is found to be more efficient for the work presented in this thesis and is used subsequently in the recognition process reported in this thesis. The window of 10 ms is incremented every 10 ms each step as before.

3.3.2 Plosives

It is possible to distinguish three stages in the production of a plosive: closing phase, closure (occlusion) and release. During the first two, the pressure is slowly building up behind the point of occlusion. Then a sudden burst takes place as this pressure is released [MacKay 1987]. If we look at the signal waveform, the two parts are drastically different. Hence, it seems appropriate to split plosives into two segments: closure and release.

The plosives were also transcribed as consisting of two parts in the TIMIT database: [b], for instance, referring to the release portion and [bcl] to the preceding closure. Therefore, it was relatively easy to incorporate this distinction into the mapping procedure. Instead of treating the whole portion as a single phone, the signal was segmented into a closure and a release based on the TIMIT transcription files. Then an average cepstral vector was calculated for each of those across all the examples [Iskra 2000]. In this way, six extra phones: [bcl], [dcl], [gcl], [pcl], [tcl] and [kcl] were added to the phone inventory. The

closure portions of the affricates [jh] and [ch] were also marked separately as [dcl] and [tcl] respectively. This approach is used throughout this work and in the recognition process the closure and release portions of plosives will be recognised separately.

3.4 Vowel model

The first step in the recognition processing is the establishment of a mapping between PARs and acoustic parameters. This is done in order to set the system up for subsequent processing. 10 vowel formant examples for 10 distinct vowels were chosen from the 10 utterances (shown in the Appendix A.2 (speaker ID: dr1/mmrp0)). The choice of the particular 10 vowels is explained on page 44. The regression analysis used to obtain regression coefficients was based on the feature values of vowels defined by Iskra [2000] (shown in table 3.2) and the obtained 10 vowel formant data from the examples.

Formant frequencies, amplitudes and bandwidths are chosen as acoustic parameters (Iskra [2000] has used cepstral coefficients. We do not use cepstral coefficients at all for the work presented in this thesis). The conditions mentioned in previous sections such as speech data, multiple regression equation and new feature values are used to produce a new vowel model. In the processing, single examples of each vowel are found in the transcription files and their quality is checked by examining the signal waveform. Then portions are cut out in such a way as to exclude the transitions. At that point 18 formant parameters are calculated for each segment using ESPS *xwaves* software.

In order to provide the pseudo-articulatory descriptions, each vowel is described in terms of four features: high, back, round, and tense. The values for those are continuous rather than binary and are allowed to range from 0 to 100. They are derived from the data provided by Ladefoged [1993] with the purpose of representing idealised articulatory positions rather than those of a particular speaker. The complete list of vowels and their respective articulatory descriptions are shown in table 3.2.

The mapping coefficients are calculated using multiple regression analysis. The (shorter) equation 3.2 is used.

The regression analysis is performed using 10 independent variables and 48 data points. Vowels constitute 10 of the data points while remaining ones are diphthongs defined here as mid-points (formant frequencies, bandwidths, amplitudes and feature values) between any two vowels.

Eventually, for each vowel (and diphthong) a set of 18 equations corresponding to 18 formant parameters (6 formant frequencies, 6 formant amplitudes, 6 formant bandwidths) are solved resulting in a set of 11 regression coefficients corresponding to the 10 independent variables plus one constant. These coefficients describe a line which provides the best fit for the data.

The coefficient of determination, which returns the value of comparing estimated and actual y-value, ranges in value from 0 to 1. If it is 1, there is a perfect correlation in the sample – there is no difference between the estimated y-value and the actual y-value. At the other extreme, if the coefficient of determination is 0, the regression equation is not helpful in predicting a y-value. The higher the value, the greater the relation between the variables. The results are multiplied by 100, thus the final results are between 0 and 100, shown in table 3.3. For the sake of comparison, the correlation values obtained by Iles are given too, since this work and Ile's work are all based on formant parameters. In table 3.3, *f*, *a*, and *b* respectively represent formant frequency, amplitude and bandwidth.

Comparison of the new values with Iles' is very encouraging. Sufficient accuracy has been obtained for future recognition processing by employing the shorter equation 3.2 and with the help of *xwaves*. Although the comparison is not entirely justified since in the two cases different speech material was used, we have achieved better results than Iles'.

	r^2	Iles' r^2
f1	93.4	97.8
f2	90.9	94.9
f3	86.9	74.6
f4	95.8	80.0
f5	68.7	34.4
f6	72.5	58.0
a1	96.7	78.2
a2	96.1	89.1
a3	94.1	77.4
a4	68.3	47.8
a5	89.5	82.8
a6	87	45.3
b1	62.3	91.2
b2	89.1	-
b3	77.1	62.8
b4	62.2	50.6
b5	65.5	72.2
b6	76.8	90.9
mean	81.8	72.2

Table 3.3 Coefficients of determination

3.5 Consonant model

In order to find the articulatory descriptions for consonants, features high, back, round, and tense, have also been chosen. It is well known that the phonological binary distinctive features used for the description of consonants are somewhat different in different phonetic textbooks. The difficulties of finding features: high, back, round, and tense for consonants are similar to part of the difficulties of linguistics to find the right set

of features for consonants, which is an open topic. On the articulatory side, the features high, back, round, and tense may not be the features with which consonants are frequently described. In the articulation of consonants, there are other features which play a much more important role in the consonant classification. Therefore, the four features used here may be considered insufficient to describe consonants in terms of articulatory gesture significant to their production. Nor will they suffice to describe the consonants uniquely using only their binary mode. Using continuous values, however, makes more room available for the classification. The four features with continuous values make the description for consonants more efficient than the 11 or 13 binary features in the phonetic textbooks (e.g. in Fromkin and Rodman [1998]). Of these four features, “tense” is the one commonly used for consonants, probably more than for vowels. In the context of consonants, it also signifies the amount of articulatory effort. In practice, it is often equivalent to the presence (lax) or absent (tense) of voicing. The features “high” and “back” can still be used to refer to the position of the tongue during the articulation, e.g., a consonant described as alveolar in terms of the place of articulation will be described as front, another one described as velar will be back, etc. “Round” can also be used to denote lip rounding during the production of a consonant. Additionally, the combination of these four features can efficiently indicate the scope the vocal tract, which relates to formant parameters.

Because of the fact that consonants are hard to describe using these four features and that the same mapping model is supposed to be used, the feature values for consonants are not determined on the basis of phonetic textbooks, but calculated using the vowel model.

However, because of the high diversities or shortage of examples in the context in the TIMIT data, it is quite hard to find representative formant parameters for some of the consonants, especially for plosives and fricatives. It is impossible to produce a satisfactory consonant model with unrepresentative formant parameters. Since cepstral analysis renders itself equally well for the representation of both vowels and consonants, Iskra [2000] had done all the possible experiments using cepstral coefficients to obtain

the consonant model. The same speaker and the same speech data were used. Although it seemed desirable to limit the scope of the experiments to a single speaker, all the available data from that speaker were used in the mapping experiments (to obtain the consonant model) by Iskra. Her statistical analysis of the results shows that the consonant model is good enough to be regarded as the reference. Consequently, we chose Iskra's consonant model as the phoneme reference.

In Iskra's analysis for the consonant model, the same shorter equation 3.2 was used as in the regression analysis for vowels (see section 3.4). The regression coefficients a_i were taken from her vowel model and the y_i values were provided by the cepstral vectors for each consonant. The feature values *h*, *b*, *r*, and *t* were thus unknown. In this way, for each consonant a set of 18 equations were formed where the feature values had to be found. The equations were solved using a brute search mechanism, which will be discussed in the following chapter in more detail.

Though we use Iskra's consonant model as the phoneme reference, we still provide consonant feature values from our analysis. The same procedure as the above has been used. But the regression coefficients a_i are taken from our vowel model and the y_i values are the formant parameters for each consonant (obtained from the 10 utterances from speaker ID: dr1/mmrp0). Brute search mechanism has been used to find the feature values of high, back, round and tense for each consonant. The consonant feature values from our analysis are presented in detail in Appendix C.

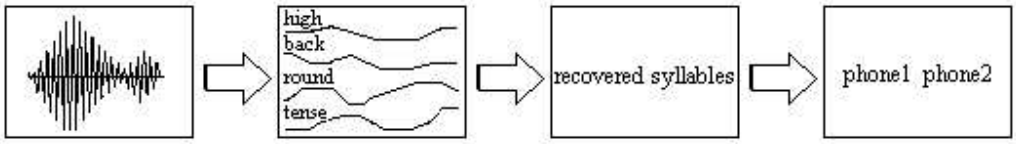
4 DERIVATION OF PSEUDO-ARTICULATORY TRAJECTORIES – RECOGNITION I

This chapter presents the first stage of the recognition process, namely the derivation of pseudo-articulatory trajectories from the incoming speech signal. First of all, a schematic view of the recognition process is introduced with a detailed description of the implementation of the first stage. Then the test set speech data is described. Following this, the brute search mechanism, which is used as the driving force of the recognition process, is outlined. This algorithm produces four closest pseudo-articulatory values for each incoming formant vector. Next, the recognition results are presented in the form of pseudo-articulatory trajectories which are then overlaid with the idealized ones for the purpose of comparison. The idealized trajectories seem to be a reasonable approximation of the new ones, at least on the average, since the newly recovered trajectories clearly contain considerably more peaks and troughs. Then another evaluation means – resynthesis – is resorted to. The quality of the synthesized speech is judged to be very good providing evidence in favour of the articulatory-acoustic mapping.

4.1 A schematic view of the recognition process

Three successive stages can be clearly distinguished in the recognition process. The first stage, which is covered in this chapter, is responsible for the transition from the acoustic representation of the incoming signal to the pseudo-articulatory representation with feature trajectories available as a function of time. The second stage, which is presented in the following chapter, concerns the derivation from the pseudo-articulatory representation of syllable structures and produces a sequence of recovered syllables. The third stage presented in chapter 6, focuses on the transition from the syllable patterns to the phonetic level of description and produces a sequence of phoneme labels. The first and second stage are considerably more interesting from the research point of view, being the final test for this particular new approach to speech recognition. The third stage, nevertheless, is crucial for practical reasons: in speech recognition one is interested in obtaining a phoneme sequence as a result of the recognition process, rather than any

intermediate representations. Finally, derivation of pseudo-articulatory trajectories has to demonstrate whether the articulatory-acoustic mapping works and that the many-to-one mapping problem can be dealt with. Different stages of the recognition process are governed by their own rules. The schematic view of the recognition process is shown in figure 4.1.



acoustic representation PAR trajectories syllable structures phoneme-level description

Figure 4.1 A schematic view of the recognition process

The first stage of the recognition process establishes every 10 ms a set of 18 formant parameters for the incoming speech. A brute search mechanism is used which by gradually reducing the solution space determines four PAR values for each set of 18 formant parameters. As a result of this, an utterance is described with a set of values for high, back, round, tense every 10ms. When plotted, these values present pseudo-articulatory feature trajectories for that utterance.

4.2 Speech data – test set

Since all the utterances available for a single speaker (ID: dr1/mmrp0) in the TIMIT database have been used to provide acoustic data for the mapping (training set [Iskra 2000]), the same speech data are used as a test set in the recognition process making this a closed testing procedure. In a fully speaker independent system it is compulsory to use different speakers for training and testing in order to prevent biasing towards certain voice characteristics. In this case, however, the approach is completely new and limited speech data have been used. Furthermore, it seems important not to introduce too much variance. For these reasons, using the same speaker and the same data for the mapping analysis (training) and testing is justified.

In the mapping analysis, reflection coefficients and cepstral coefficients have been used as the acoustic parameters to provide the phoneme reference [Iskra 2000]. In order to avoid a too closed testing, formant parameters have been chosen to set the system up and they are also used for the subsequent recognition processing in the testing set. However, the syllable patterns mentioned in the second stage of the recognition process are recovered from feature trajectories and are not parts of any mapping analysis (see chapter 3.4). In all the speech material uttered by the male speaker (see section 3.2.1), no single example was found for the following phones: [ax-h], [em], [en], [eng], [hh], [jh]. The first four are actually allophones, instantiations of a phoneme in a particular context. Therefore, it is not a major violation if they are not included in the phone inventory. Instead of the phone [hh], its voiced allophone [hv] was present in the phoneme reference. Finally the only missing phone was [jh].

As in the mapping analysis the same 10 sentences of the male speaker (ID: dr1/mmrp0) from the dialect region of New England were selected as the test set. All the speech material has been analysed every 10ms with a set of 18 formant parameters as the result.

4.3 Brute search

The brute search algorithm has been used to derive the pseudo-articulatory trajectories. The idea of the algorithm is borrowed from Iles [1995], but it is implemented again to suit the current task. The same procedure has been used earlier to obtain feature values for consonants [Iskra 2000]. The following equation is used again:

$$f = a_0 + a_1 h + a_2 b + a_3 r + a_4 t + a_5 hb + a_6 hr + a_7 ht + a_8 br + a_9 bt + a_{10} rt$$

Equation 4.1

where f are the target formant parameters from the test files and a are the regression coefficients. The feature values of h, b, r, t are the results of the search process. For each record in the test files, 18 formant parameters are available, thus forming a set of 18

equations. Brute search is used to solve the equations in a space confined by feature values. The brute search is done by incrementally substituting values between 0 and 100 for *h*, *b*, *r*, *t* and calculating the formant parameters for each hypothetical set of feature values. Then the distance between these formant parameters and the target ones from the test file is calculated. The search is continued until the top value of the range (100) is reached and the minimal distance is remembered. The feature values which have produced the minimal distance are taken to determine a new range of search. The maximum of the range is equivalent to (value + increment), while the minimum is (value-increment). The increment value is then refined and a new search is begun. That continues again until all the value combinations have been tested. The set of feature values which have produced formant parameters with the minimal distance to the target formant parameters from the test file are then regarded as the best matching pseudo-articulatory values for a particular slice of speech signal.

In this way, the target formant parameters for the next speech record are used as the input and the search process starts again. No information from the previous time window is used to constrain the search space and restrict the potential feature values in the current time window, which has a positive influence on the results. If the cepstral coefficients are used as the parametric representation instead of formant parameters, it is argued that this kind of constraint is inherent in the parametric representation itself and is thus reflected in the values of the cepstral coefficients. Moreover, at this stage of the development of the approach this kind of constraint might have a negative influence on the results. If the values found in the preceding time window were used to constrain the search space in the current time, an assumption would be made that these values were correct. If the values were incorrect, the new search would be constrained to the wrong part of the feature space. Hence, the usage of formant parameters instead of cepstral coefficients can avoid such kind of potential negative influence on the results.

On the other hand, the algorithm may seem inefficient because every time a whole range of values has to be tested. It should be pointed out, however, that the increment value is

related to the degree of articulatory activity, and has to demonstrate physiological plausibility. It seems safe to start with a larger increment (5%) which is then refined (to 1%) since articulators move relatively slowly and are unable to jump from one extreme position to another within a short period of time. A larger initial increment value makes, however, no risk of overlooking any articulatory activity, but helps to increase the efficiency of the algorithm.

4.4 Recognition results as pseudo-articulatory trajectories

As a result of the brute search mechanism a series of four pseudo-articulatory values is produced for each 10ms of speech in the test file. This is repeated for a set of 10 sentences. The results are plotted as trajectories for respective features and compared to the idealized ones. The idealized trajectories are produced by ascribing four feature values to every segment in the transcription files. The values for vowels are taken from the vowel model. The values for consonant are taken from the consonant model (see section 3.4 & 3.5).

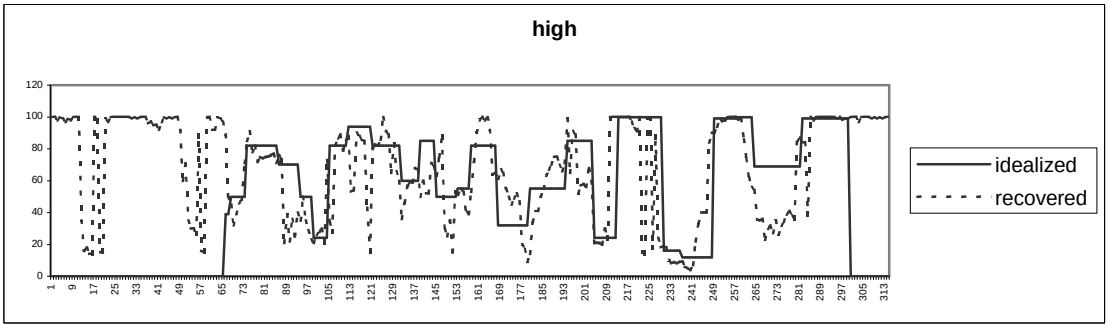


Figure 4.2 Trajectories for feature “high” for one sentence: idealized and recognized

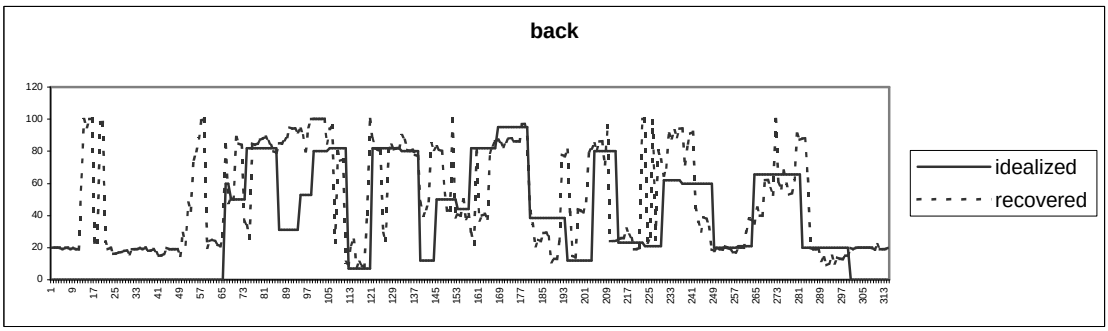


Figure 4.3 Trajectories for feature “back” for one sentence: idealized and recognized

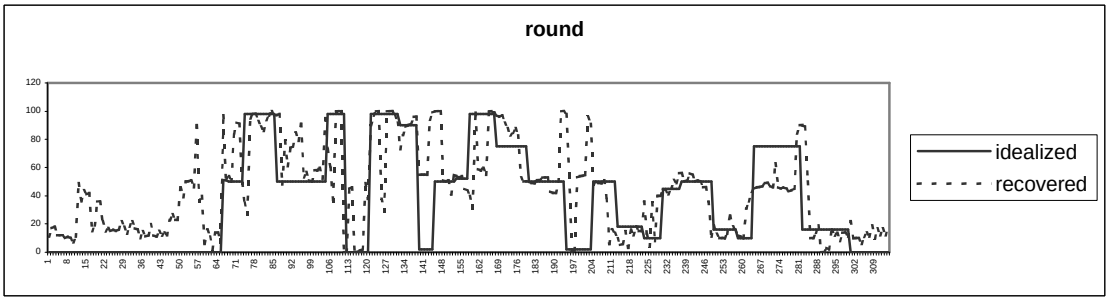


Figure 4.4 Trajectories for feature “round” for one sentence: idealized and recognized

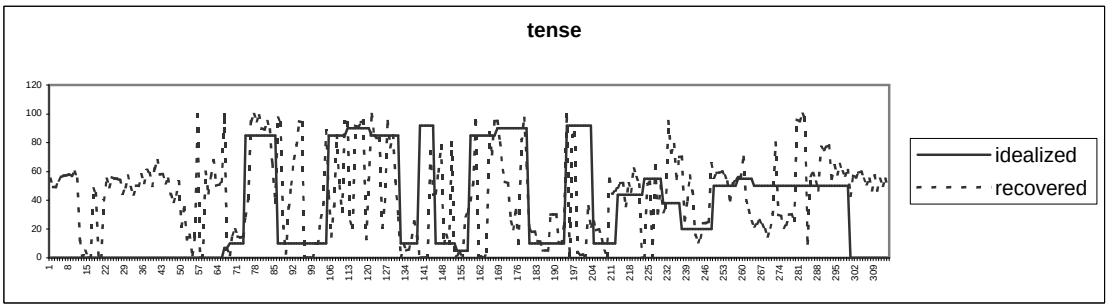


Figure 4.5 Trajectories for feature “tense” for one sentence: idealized and recognized

As can be seen in the above figures, the recovered pseudo-articulatory trajectories are very irregular and extreme minimal and maximal values are often reached within the space of a few records. Therefore, it is hard to create a general picture of how close the

recognized trajectories are to the idealized ones. The least similarities seem to occur in the case of feature “tense”. For features “high”, “back”, “round”, on the other hand, the idealized trajectories seem to be a reasonable match with the recovered ones – at least on average, since the new trajectories clearly contain considerably more peaks and troughs.

4.5 Evaluation by resynthesis

Another way of evaluating pseudo-articulatory trajectories is resynthesis. In order to be able to do it, the whole recognition process has to be inverted: cepstral values have to be calculated on the basis of pseudo-articulatory trajectories and these values have then to be used as input to a speech synthesizer. If the quality of resynthesized speech is close to the original, that will indicate that the articulatory-acoustic mapping is correct and the derived pseudo-articulatory trajectories are satisfactory. Moreover, that will imply that it is possible to deal successfully with the problem of many-to-one mapping.

The pseudo-articulatory trajectories obtained as a result of the recognition process are used to calculate the cepstral values. The mapping equation (see section 4.3) is run backwards: for each record the four feature values are supplied and the mapping coefficients are known from Iskra’s [2000] multiple regression procedure (since she used cepstral coefficients as the acoustic parameters). So the cepstral values can be easily calculated on the basis of the above information. As a result, each record is represented by an 18-dimensional cepstral vector, which is then used for synthesis.

After the cepstral values have been obtained, a standard ESPS synthesis procedure is used which requires the following parameters.

- reflection coefficients – these are calculated on the basis of cepstral coefficients using one of the ESPS conversion program

- F0 – fundamental frequency estimated using the normalized cross correlation function and dynamic programming

- probability of voicing (prob_voice) – frame voicing state which could take on two values: 0 and 1

- rms – local root mean squared measurements

The last three parameters are derived from the original sentences using ESPS programs. Then together with the reflection coefficients they are used as input to a synthesizer.

The ESPS synthesizer uses reflection coefficients (or alternatively linear prediction coefficients) to synthesize a speech signal from an excitation source or a parametric source including F0, rms and prob-voice. The synthesizer uses a Rosenberg-polynomial glottal-flow pulse, open-phase damping, and per sample gain correction [ESPS 1996].

The quality of the synthesized speech is evaluated by listening to it. The listening is done in an informal manner rather than as part of a proper test. All the sentences are comprehensible and clear, and sound natural. The waveforms of the original as well as the synthesized speech are presented in figure 4.6. The sentence is: *The willowy woman wore a muskrat coat.*

It can be seen that the waveforms of both sentences are very similar. The differences are perceived as less amplitude (loudness) in parts of the waveform of the synthesized speech. This is true of the remaining sentences as well. Additionally, the synthesized speech of our work is better than Iskra's [Iskra 2000], since there are some clicks in Iskra's synthesized speech because of some errors in file handling. Thus, the derived pseudo-articulatory trajectories reported here are more reliable than the trajectories recovered by Iskra.

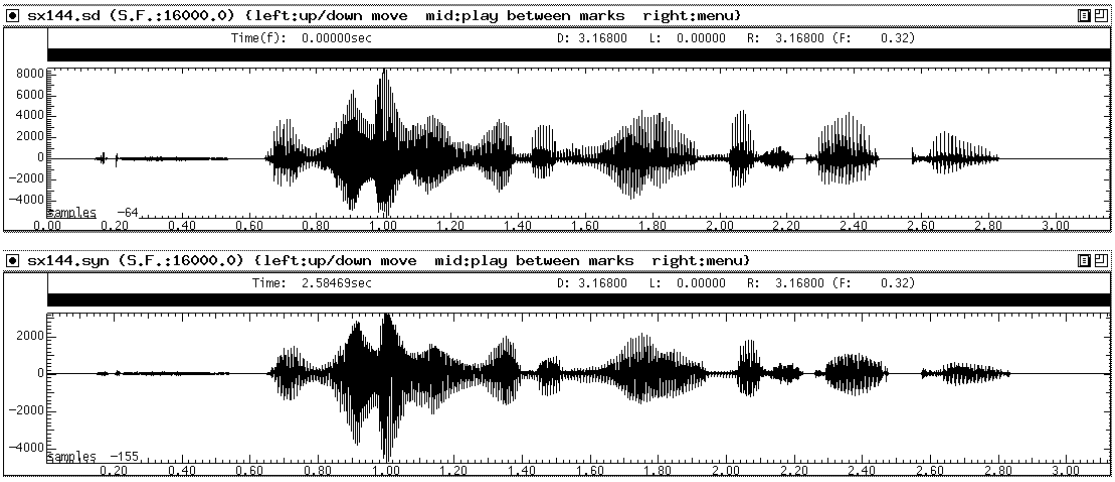


Figure 4.6 Signal waveforms for one sentence: top – original, bottom – synthesized speech

4.6 Discussion and Conclusions

When the pseudo-articulatory trajectories are derived, it is not clear if they support the articulatory-acoustic mapping. The derived trajectories are clearly too "noisy". Although the general trend of four idealized trajectories follows that of the new ones, the feature trajectory of "tense" seems not as good as the others. In the previous work [Iskra 2000], "tense" already proves to be poorly modeled as for very few consonants the calculated values for "tense" fall into the control range[4] in the phoneme reference. It is, therefore, expected to perform less adequately than the other features in the recognition process.

The fact that some features perform better than others may have to do with the inherent nature of these features. "High" and "back" refer to the position of the tongue, an aspect of articulation which is valid for both vowels and consonants. Both vowels and

[4] In previous work [Iskra 2000], in order to evaluate the consonant model, the PAR values obtained for consonant were compared to phonetic feature specifications found in the textbooks. The feature values given in books are always binary, so in order to make the comparison possible [-feature] was assumed to correspond to all the values in the range 0-33, [+feature] to 67-100 and anything in between [± feature] to 34-66. These are the control ranges. If a found PAR value fell within this control range, it was considered to be "the right match".

consonants can be classified in terms of binary values of the feature “back” in English. “Round” corresponds in the same way to the shape of the lips. These features can be interpreted in the classificatory phonological sense with binary values, but also can be applied as phonetic features describing the position of the tongue and the degree of lip rounding with a number of possible values. Moreover, these features also have major acoustic correlates (see section 3.2.2). In brief, the first formant is inversely related to the feature “high”. The front-back dimension is more simply expressed by reference to the difference between the first and the second formant frequencies. Lip rounding enhances the low second formant frequency with a high back tongue position. By using the representation of formant parameters, such correlates can be very well captured. The feature “tense” in the phonological sense has been widely used to classify vowels that occur in stressed open syllables and consonants which carry the phonological feature [+voiced] [Ladefoged 1993]. Its acoustic correlates are however not straightforward, although some trends in the voice source can be measured. These concern, for instance, the sharpness of the glottal closure, whether the vocal folds make contact instantaneously or gradually along their length, the cutoff frequency, the degree of openness which mainly controls the amplitude of the lower components of the spectrum or the skewing of the glottal pulse [Ni Chasaide and Gobl 1997].

In the context of consonants, the phonological feature “tense” is related to the phonological feature [voiced]. If one is present, the other one is absent and vice versa. The phonological feature [+voiced] cannot be, however, taken to mean voiced in the phonetic sense of vocal fold vibration. The sounds phonologically designated as [+voiced] are not necessarily accompanied by the vocal fold vibration. Evidence towards the discrepancy between the phonological feature [voiced] and the actual vibration of the vocal folds is provided by stop sounds [MacKay 1987]. It implies that the closure phase of a [+voiced] plosive can be voiceless or voiced depending on its position. Therefore, the phonological feature [voiced] cannot be directly translated to phonetic activity. If “tense” is assigned to consonants on the basis of the phonological feature [voiced], then “tense” cannot be entirely correlated with the activity of the vocal folds. The lack of clear

unambiguous phonetic/acoustic correlates is more evidence for the poor recognition and modeling of the feature “tense”. However, though it may not have seemed entirely satisfactory, “tense” is important in as much as it provides classifications for vowels and consonants. This feature can be measured in terms of the parameters of the voice source [Ni Chasaide and Gobl 1997], which may provide more reliable modeling reference. Moreover it may be possible to improve “tense” trajectories by using other (or combinations of) acoustic parameters, which can capture enough phonetic/acoustic correlates for this feature.

The derived pseudo-articulatory trajectories were expected to be smoother. Even though the recovered trajectories seem to follow the idealized ones at least on the average, they demonstrate numerous irregularities. At many points, there is a large sudden change in the feature values in the space of a few records. This effect has blurred the actually important articulatory transitions. In order to prevent it, the search mechanism should perhaps have been more controlled by taking into account feature values of the previous records and disallowing implausible transitions. Unfortunately, such constraints may have the risk of sending the new search into the wrong direction if the current values are incorrect since with further processing, the incorrectness will increase. Consequently, a smoothing algorithm is chosen as the solution, which is described in the following chapter. It can act as an additional articulatory constriction, which monitors the articulator movement in time and prevents a magnitude of change which is unfeasible from the physiological point of view.

To summarize, since there are so many peaks and troughs in the newly derived trajectories, it is difficult to evaluate them. A need arises for more evidence to support the mapping. As another means of evaluation, resynthesis was resorted to. The resynthesis results are very satisfactory. The synthesized speech is comprehensible and clear, and sounds natural too. The results are in favor of the mapping procedure.

Moving back and forth between the articulatory and the acoustic space has seemed a major challenge because of the non-unique nature of the relationship. In the earlier work, Iles [1995] has already effectively solved the articulatory-acoustic mapping problem for speech synthesis. The resynthesized work reported here, which has inverted the recognition procedure, also implies that the many-to-one mapping problem can be successfully dealt with. That is why pseudo-articulatory features have been used for the description of this space. They are idealized and effective enough for the abstraction from the acoustic detail into the linguistic domain. Further, the abstraction proves sufficient to achieve a two-way “communication” between the articulatory and acoustic spaces. For the work here, the use of pseudo-articulatory features and the mapping between them and their acoustic parameters provide us the feasibility to move back into the acoustic domain without the loss of speech quality.

5 DERIVATION OF RECOVERED SYLLABLE PATTERNS – RECOGNITION II

This chapter presents the second stage of the recognition process, which concerns the derivation from the pseudo-articulatory representation of syllable structures and the production of a sequence of recovered syllables. An account of syllable structure as the basis for organising articulatory activity is used to demonstrate that working with syllables can provide the basis for linguistically motivated speech recognition using pseudo-articulatory representation. In order to avoid additional problems, the experiments start with idealized pseudo-articulatory trajectories. After having obtained the recovered syllable patterns from the idealized trajectories, computationally derived pseudo-articulatory trajectories are considered. Since the newly derived trajectories clearly contain too many peaks and troughs, a smoothing algorithm is used to provide additional constraints and make them plausible from the physiological point of view. Then a conventional resynthesis procedure is followed to evaluate the smoothed trajectories. The quality of the synthesized speech is judged to be very good proving that the smoothing procedure has been satisfactory. Finally, recognition results of recovered syllables from the derived trajectories are presented.

5.1 Using syllables in speech recognition

Whether the syllable or the segment is the basic unit of articulation is a well-established debate. Sonority has been suggested and discussed as an organizing principle for syllable structure by Bell and Hooper [1980] and Giegerich [1992]. Recently, more research work, which has incorporated syllable structure into phonological representations, has presented the benefits [Kaye 1989]. Additionally, Kaye [1989] has expressed the view that “the phoneme is dead”. In this research, we assume that the syllable can be used as a basic unit or domain for organizing articulatory activity, and we demonstrate that it is a valuable unit when undertaking speech recognition processing.

Several different ways of analyzing syllable structure have been presented in section 2.2. But our main concern here is that which is most useful for automatic speech recognition processing. In conventional way, speech segments are regarded as the basic articulatory units (see section 2.2.1). Then they are organized and patterned as syllables according to phonotactic principles. In another word, vowels and consonants 'constitute' syllables. If an explicit syllable structure model (such as that given in figure 2.1 on page 8) is used in speech recognition system, it must recognize the consonants and vowels first. Structural phonotactic constraints are then provided for patterning them as syllables. Though it is feasible in speech recognition systems to identify candidate consonants and vowels at the beginning (as Iskra did in the previous work [Iskra 2000]), it is a step we try to avoid because of the fact that such processing assumes too much about the articulatory organization of speech and has a poor accuracy rate.

Syllables can be described as larger units with hierarchical structure as well. The most widely used one is the syllable model consisting of Onset and Rhyme, with Rhythm being Nucleus and Coda (shown in figure 2.1).

The onset and the coda are optional in this syllable structure. The three elements are not segments in the conventional sense since onset and coda can both be a cluster of consonants. Additionally, the nucleus is not always a vowel, for example as in the second syllable of the word "button".

Indeed, this analysis is more abstract than the traditional CVC type of sequence. However, when it comes to speech recognition systems, it still cannot offer enough information for syllable recognition processing, though the abstract structure can be used as an organizing constraint ('maximize the onset' and so forth) for syllables. And the recognition must be still attempted independently of the syllables.

A different way of working with the syllable as a unit is to use sonority as the organizing principle. As stated in section 2.2.2, syllables are 'sonority waves'. In brief, the sonority

of the speech sound builds up during the onset, to the peak value at the nucleus, and drops away again in the coda, the whole cycle repeating as syllables are produced in sequence. In this model, each individual speech sound/segment has a sonority value on a scale of 0.5-10 according to Selkirk [1984]. Thus the constraints on sequential arrangements of consonants in the onset and the coda are explained in terms of sonority contours. This provides additional constraints compared with CVC type of models of syllable structure, which can assist recognition.

5.1.1 Articulatory pattern in the syllable

The approach we have taken focuses instead on the notion that a syllable is basically an articulatory unit. As Ladefoged [1993] said, "In many cases, a consonant can only be said to be a particular way of beginning or ending a vowel". And this is a general description of the syllable model that we use in our speech recognition system. Our conjecture [Edmondson and Zhang 2002] is that initially mandible movement cycles of mouth opening and closing produce specified resonant vocalizations – syllable targets (e.g. vowels) with unspecified transitions between them. The unspecified transitions gain specifications as dynamic targets, which will in turn lead to further transitions. Recursively, we can go one layer deeper and invoke transition targets (consonants) to obtain the articulatory layered syllable model illustrated in figure 5.1[5].

The model gives three layers altogether, where 's-tar' means syllable target, 'd-tar' means dynamic target, 'tr-tar' means transition target, 'tr' means transition. The use of bold font in figure 5.2 means that the identified component is marked for a specific 'phonetic' value, normal font means that the component is not identified as marked (it may have a complex specification, or no specification), italic means the component cannot be marked. Clearly, s-tar is always marked in reality (else there would be no syllable).

[5] 'Layer' has no specific theoretical meaning, in this context just being used in figure 5.1 to refer to recursive refinements. Here three layers have been considered sufficient.

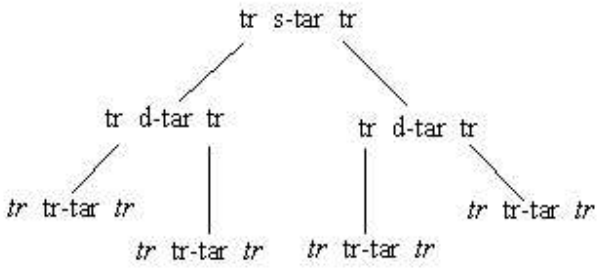


Figure 5.1 The layered syllable model

In this scheme articulatory activity must consist of tr, x-tar, tr, x-tar, tr, x-tar etc. where syllable nuclei are marked by x = s, and where phonetically irrelevant tr are *tr*. Typically, then, a CCCVCCC syllable might look like:

tr, tr-tar, *tr*, d-tar, *tr*, tr-tar, *tr*, s-tar, *tr*, tr-tar, *tr*, d-tar, *tr*, tr-tar, *tr*

An example of how this might be used for the English word ‘apt’, is shown in figure 5.2.

tr,	**s-tar**,	*tr*,	**tr-tar**,	*tr*,	d-tar,	*tr*,	**tr-tar**,	*tr*
	[æ]		[>p]		[pt]		[t<]	

Figure 5.2 An example of the syllable model

Figure 5.2 shows that the articulatory detail can be labeled ‘phonetically’ but this does not equate to phones. The [>p] is shown not as a phone, but rather just as the closure phase; likewise the [t<] is shown as release phase. ‘>’ refers to closure and ‘<’ refers to release. Additionally, complex articulatory activity, without phonetic significance but required for the phonetic string in which it is embedded, can be recorded, as in the case of the change in point of obstruction in the phase, or component, labeled ‘d-tar’.

There are several points to note about this model. First of all, the natural periodicity of syllabic activity is provided by the physiology and can be derived by a task dynamics analysis of mandible movement [Beckman et al. 1992; Hawkins 1992]. Since not all syllables involve a mandible movement, the proposed syllable model provides the underlying periodicity for all syllable sequencing. Secondly, the overall periodicity of the syllable model provides a syllable time frame within which successive linguistically specified articulatory targets have to be reached or approximated, with intervening motion not specified; there is insufficient time to fit in more specified targets. Finally, the mapping between conventional phonetic segments and the articulatory targets (s-tar, d-tar, tr-tar) is not trivially obvious although the number of articulatory events does seem right (see Beckman et al. [1992]). And now we discuss it in more detail.

Alignment

The syllable model mentioned above has two broad categories of targets: syllable nucleus (s-tar) and others (d-tar, tr-tar). The s-tar is characterized as a rather stable articulatory configuration with voicing, yielding sonorant continuous airflow. The transition targets (d-tar, tr-tar) are generally characterized as obstructive in the oral cavity with or without oral airflow (liquids, glides, fricatives, stops and nasals).

A plausible explanation at this stage of the work is that d-tar is like a target which stops the oral cavity, while the tr-tar is described as transitions more conventionally called closure and release, as well as obstruents. Additionally, not every slot of the syllable model outlined in figure 5.1 has to be filled or presented in the articulation of each syllable, except for the syllable target. The unmarked transitions in the model may be more or less brief, ranging from clearly perceptible to inaudible. Whether or not each characterized transition has to be labeled as a value equivalent to a conventional phonetic segment remains in doubt. Usually, the transitions are characterized as audible and this is how we recognize where the oral cavity is closed. The actual configuration in the closed condition is not audible, and can be changed. Thus the d-tar is inferred and it is not

readily recovered directly in the speech recognition process we have developed (see section 5.4).

Here are more examples to show how the elements in the syllable structure are characterized.

‘asks’ -> [aa] [s>k] [k<s]

‘pines’ -> [p<h] [ay] [y>n] [n<z]

‘pints’ -> [p<h] [ay] [y>n] [t<s]

‘spines’ -> [s>p] [p<a] [ay] [y>n] [n<z]

Using the proposed syllable model, it is possible to describe the difference between the aspirated voiceless stops and the unaspirated allophones as a difference in specification for the unmarked transition between stop and vowel.

Finally, there are still several problem cases left unsolved using this model, which require further work.

5.2 Syllable recovery using idealized pseudo-articulatory trajectories

There is a general point which needs to be made before considering the details of syllable recovery. A conventional approach to speech recognition may ignore, for example, information in the signal which can be used to derive syllable timing or structural information independently of phone sequences. Our approach shows that syllable structure information can be directly recovered from the speech stream without reliance on phonetic segment identification. In fact, independent measures of timing information and sonority may provide more independent contributions. Speech recognition can proceed from a much richer base when syllable structures are derived irrespective of

phone recognition as the beginning. This general observation is important regardless of the detail of our approach.

We now return to the problem of syllable recovery in relation to pseudo-articulatory trajectories. We chose to work with idealized pseudo-articulatory trajectories first because we want to determine the feasibility of relating PARs to syllable structure without any additional problems which might arise from the use of PARs computationally derived from speech signals. If we can demonstrate the feasibility of the relationship, we will go on to consider syllable recovery from computationally derived PARs.

As previously stated, the idealized pseudo-articulatory trajectories are produced by ascribing four feature values to every segment in the transcription files. The values for vowels are taken from the vowel model. The values for consonants are taken from the consonant model (see section 3.4 & 3.5). Now, we will show how the details of syllable articulation can be recovered from idealized PARs.

5.2.1 Syllable recovery

In the idealized pseudo-articulatory trajectories, smoothed transitions between ideal targets are presented, as well as the targets themselves. Between targets there is a significant change in the feature values. For any idealized target, especially vowel targets, the trajectories remain stable, and thus feature values as well. By using the articulatory pattern in the syllable, which we have discussed in 5.1.1, as a rule, an algorithm has been created to identify the targets and transitions in the utterance context. The detailed description of the algorithm is below.

```
for 4 feature trajectories
  {
      compare two adjacent records;
      if any two of the four feature values remain the same[6]
          {   It is a target;
              duration =1;
              call target_processing (duration);
          }
      else
          {   It is a transition;
              call tr_processing;
          }
  }
  output results;


target_processing (int duration)
{   continue comparing the next two adjacent records;
    if any two feature values remain the same, increment duration until no two feature
    values remain the same;
    if duration >= min_duration_of _vowel
          {   It is a s_tar;
              record position;
          }
    else
          {   It is a tr_tar;
              record position;
           }
}


tr_processing
{   continue comparing the next two adjacent records until any two feature values remain
    the same;
}
```

Two arbitrary parameters are used in this algorithm. The first is the choice of two of the four features remaining the same for the trajectories to be considered 'stable'. This choice was found in the training data (three male speakers and one female speaker[7]). This gives

[6] In processing the idealized PARs, "the same" means literally identical. As given on page 79, for recovered PARs a threshold of 5 was used.

[7] Speaker ID: *xwaves* example speaker Speech data: There is usually a valve.
Speaker ID: dr8/fclt0 Speech data: sx358, si2068

the best match between the results (recovered syllable structures) of the algorithm and the original transcription files. In a similar way, the minimum duration of a vowel was chosen to be 40 ms. Here is an example for the identification of targets and transitions for the idealized PAR trajectories. At the beginning of the utterance, after the first transition, there will be a target. It has an uncertain specification because in the syllable onset there can be more than one consonant or no consonant at all. The algorithm will read the following data points along the sequences of feature values to recover further information. On the basis of evidence from the following data, the unknown articulatory activity can be marked for a specific articulatory value. The subsequent articulatory activities are marked in the same way, using data even further down the sequences as well as information from the already labeled articulatory activities. After any target has been identified, the transition will be recognized subsequently and vice versa. According to the syllable model we have discussed in 5.1.1, between any two syllable targets, there is a maximum of 6 transition targets to be recovered. In this way the syllable structures are recovered in sequence. Meaningful syllable structures for one utterance have been derived in this way.

Finally, a sequence of tr, x-tar, tr, x-tar, etc, has been derived as well as syllable positions. The syllable positions are represented by record numbers (see section 5.4). As we have discussed in 5.1, syllables are ‘sonority waves’. The sonority of the speech sound builds up during the onset, to the peak value at the nucleus, and drops away again in the coda, the whole cycle repeating as syllables are produced in sequence. Since the sequence of articulatory events and syllable positions are recovered, it would be easy to find the syllable peaks according to the principle mentioned above. Finally, the ‘sonority waves’ are produced in sequence and are shown diagrammatically in figure 5.3.

Speaker ID: dr8/mmws0 Speech data: sx348, si559
Speaker ID: dr1/mmrp0 Speech data: sx144

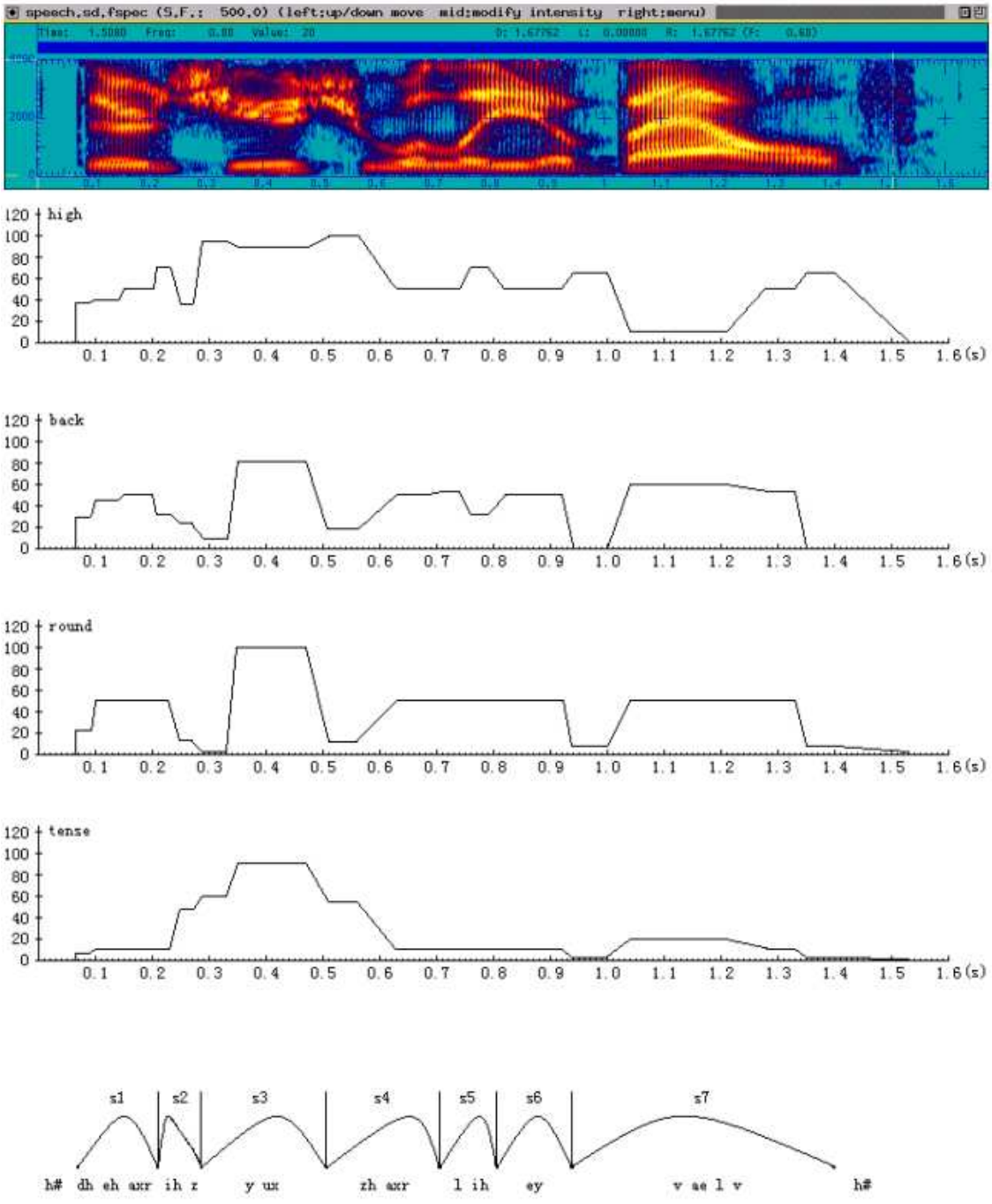


Figure 5.3 The top section shows the spectrogram of the utterance " There is usually a valve. ". The middle 4 traces show the idealized feature trajectories of high, back, round, tense. The bottom section shows in schematic form the recovered syllable positions.

The algorithm seems promising although currently it is based on idealized PARs.

5.3 Syllable recovery using derived pseudo-articulatory trajectories

After having successfully demonstrated the feasibility of relating idealized PAR trajectories to syllable structure, now we will go on to consider syllable recovery from computationally derived PAR trajectories. Since the derived PAR trajectories clearly contain considerable peaks and troughs, a smoothing algorithm is used to make them more plausible. Then a resynthesis evaluation procedure is applied to assess the quality of the smoothed trajectories. The quality of the synthesized speech is also investigated. By using the syllable recovery algorithm, meaningful syllable structures have also been recovered from the smoothed trajectories, which makes this algorithm more promising.

5.3.1 Smoothing derived pseudo-articulatory trajectories

The recognition work is being continued with the focus on such aspects as smoothing the computationally derived PARs. It is a well-established fact that articulator movements are in reality relatively slow so the derived trajectories, illustrated in figures 4.2, 4.3, 4.4, and 4.5, do not equate well with natural articulation. To solve this problem, a two point averaging algorithm is used to constrain the scope of variation for the movement of each articulator. Averaging is applied repeatedly until for any feature the changing feature value between adjacent two records reaches maximum 30. The results are illustrated in figures 5.4 and 5.5.

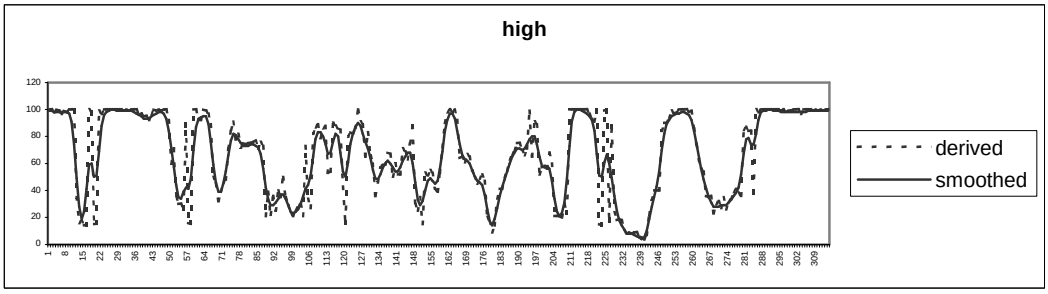


Figure 5.4 Trajectories for feature “high” for one sentence: derived and smoothed

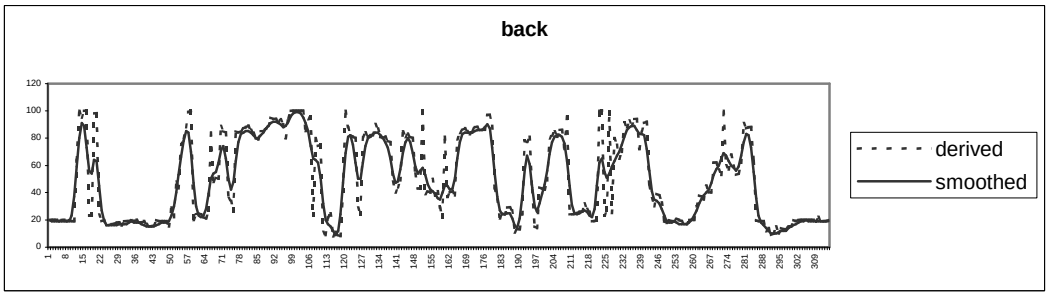


Figure 5.5 Trajectories for feature “back” for one sentence: derived and smoothed

The value of 30 was chosen as follows. In previous work, in order to evaluate the consonant model, the PAR values obtained for consonant were compared to phonetic feature specifications found in the textbooks. The feature values given in books are always binary, so in order to make the comparison possible [-feature] was assumed to correspond to all the values in the control range 0-33, [+feature] to 67-100 and anything in between [± feature] to 34-66 [Iskra 2000]. This suggests a threshold for changes of approximately 32 (i.e. 34-66). Initially, if a found PAR value fell within this control range, it was considered to be “the right match”. Then 3 different sentences from 3 different speakers (speech data: sx358 [speaker ID: dr8/fclt0], sx348 [speaker ID: dr8/mmws0], “There is usually a valve.” [speaker ID: *xwaves* example speaker]) have been chosen to adjust the threshold value. Several different threshold values around 32 were tried. 30 was chosen as the final threshold value, since the smoothed trajectories using 30 as the threshold value are closer to the idealized ones but still carry enough useful information.

Then we assume for any feature any change (in 10 ms) between adjacent two records within 30 is a reasonable articulatory transition for one segment or two segments. If any change is more than 30 in 10 ms, the transition will probably cover two ranges, which is implausible to the natural articulation since it is a well established fact that articulator movements are in reality relatively slow. On the other hand, if the threshold value is less

than 30, then it is probably too sensitive to possible transitions within 10ms for any feature so that the trajectories smoothed in this way cannot capture all the necessary and crucial acoustic information, which may be important for the next step processing. Consequently, 30 was chosen as the reasonable threshold range value. Finally, smoothed trajectories have been obtained for all the 10 utterances. The next step is to evaluate the smoothed trajectories.

5.3.2 Evaluation by resynthesis

A standard ESPS synthesis procedure is used to evaluate the smoothed pseudo-articulatory trajectories. It is the same procedure mentioned in section 4.5. The smoothed trajectories obtained are used to calculate new cepstral values. The mapping equation is run backwards. Then the cepstral values can easily be calculated. As a result, each record is again represented by an 18-dimensional cepstral vector which is then used for synthesis.

Then reflection coefficients are calculated by the conversion program provided by ESPS using the new cepstral values. Together with the other three parameters: F0, prob-voice, and rms, which are derived from the original sentences using ESPS programs, the synthesized speech is produced.

The quality of the synthesized speech is evaluated by informally listening to it. All the sentences are comprehensible, clear, and sound natural. Additionally, The quality of the synthesized speech from the smoothed trajectories is slightly better than the quality of the synthesized speech from the unsmoothed ones[8]. At this point, it shows that the smoothing procedure is satisfactory and we consider the PAR data are good quality and suitable for running the syllable recovery algorithm.

[8] For examples see the CD attached to this thesis or http://www.cs.bham.ac.uk/~lxz/speech.htm.

5.3.3 Syllable recovery

Previously we have considered idealized PAR trajectories as the basis for the syllable recovery. Here we are using smoothed trajectories recovered from speech. The next step in our account is to demonstrate how the details of syllable articulation can be recovered.

The same algorithm was used as given on page 73. However, since the PAR values are now real numbers, a further parameter is needed, namely a threshold to consider values identical. In each of the four recovered feature trajectories, if the absolute difference in value between two adjacent records is no more than 5, then we regard them as two consecutive identical values. The threshold value 5 is obtained from the training of the same 6 different sentences from the same 4 different speakers (see page 73) but based on the computationally derived feature trajectories. In the smoothed trajectories, smoothed transitions between smoothed targets are presented. Between targets there is a significant change in the feature values. For any smoothed target, especially vowel targets, the trajectories remain stable, and thus the feature values as well. Based on the rule of the articulatory pattern in the syllable, the slightly changed algorithm with new definition for two consecutive identical values runs to identify the targets and transitions in the utterance context. By using data even further down the sequences as well as information from the already labeled articulatory activities, the unknown articulatory activities are marked for specific articulatory values. It was quite exciting when meaningful syllable structures were recovered for the first sentence. Meaningful syllable structures have been recovered for all the 10 utterances. The transition targets (tr_tar) and the syllable targets (s_tar) are very well recognized. The results are very encouraging and explained in detail in the following section.

5.4 Recognition results of recovered syllable patterns

Before moving on to a discussion of the recognition results, the result format is explained in order to help the reader to understand the example.

5.4.1 Result format

In order to explain the recognition results of recovered syllables here is one invented example.

Record number (10ms intervals)	TIMIT phone symbol	Recovered syllabic details
97	A	s_tar, tr
98	B	tr_tar
99	B	tr_tar
100	B	tr_tar

tr_tar3

The waveform of every utterance is processed in 10ms samples, and the processing result of every 10ms is called a record. The record number is used to mark the records from the beginning to the end. In the TIMIT database, phone boundaries are presented by time. According to the relationship between the record number and the time, phone boundaries are presented by the corresponding record numbers instead. A or B represents the phone symbol as found in the TIMIT transcription file. In this example, phone B starts at record 98, and ends at record 100. If the phone boundary in the TIMIT original transcription file does not coincide with the target boundary or the transition boundary in the recognized sequence, a number is placed after the target symbol or the transition symbol. This number refers to the number of records aligned with the particular original label. In the example, tr_tar3 in the recovered syllabic detail shows 3 records of transition targets are aligned with the original label B.

5.4.2 Recognition results

Our approach to speech recognition focuses on the use of syllable structure. The second stage of this approach concerns the derivation from the pseudo-articulatory representations of syllabic details and eventually produces a sequence of recovered

syllables. Since this part is the core of the new approach, we have presented below some examples of the results of recovered syllable patterns for the processing of 10 utterances. The transition targets (tr_tar) and syllable targets (s_tar) are very well recognized. The average accuracy rate for all the targets (duration of correctly labelled syllable targets and transition targets divided by total duration of phonemes) is 42.1%, which is very promising.

We present one example from the processing in figure 5.6. On the left-hand side there are record numbers and original phone symbols as found in the TIMIT transcription files. Following the colon there are the recognized syllabic details. A crude time alignment has been attempted here. If the phone boundary, however, does not coincide with the target boundary or the transition boundary in the recognized sequence, a number is placed after the target symbol or the transition symbol. This number refers to the number of records aligned with the particular original label. This is why the numbers can be found only at boundaries. The correctly recognized targets (in accord with TIMIT) are printed in bold, i.e. where both the target and the time overlap. The complete recognition results of recovered syllable patterns for the processing of 10 utterances are presented in Appendix D.

5.5 Discussion and conclusions

This chapter presents promising work on the development of an account of syllabic structures in speech which can be used in processing speech for recognition. Speech processing for recognition is conventionally concerned to recover a string of phones from the acoustic waveform. We have chosen here to explore the idea that it might be easier to recover strings of phonetically unlabeled syllables, and to use this information to recover phonetic detail without requiring that this detail be expressed in terms of phones. The theoretical work blends articulatory phonetics with phonology, in a way which is not reliant on conventional phonetic segmentation, to yield the basis for an articulatory model of syllables which facilitates direct recovery of syllables in speech processing.

Speaker ID: dr1/mmrp0
phonetically-diverse sentence
si2034: Make it come off all right.

si2034

15	h#:	s_tar, tr, tr_tar, tr
19	m:	**tr_tar**, tr
32	ey:	**s_tar**, tr, s_tar6
36	kcl:	s_tar4
40	k:	s_tar4
44	ix:	**s_tar4**
49	tcl:	s_tar5
50	t:	s_tar1
54	kcl:	tr, **tr_tar**, tr
59	k:	**tr_tar**, tr
65	ah:	tr_tar, tr
69	m:	**tr_tar**, tr1
87	ao:	tr3, **s_tar**, tr3
99	f:	tr1, **tr_tar**, tr, s_tar2
109	ao:	**s_tar8**, tr
115	l:	tr_tar, tr, **tr_tar**
124	r:	tr, **tr_tar**, tr, s_tar1
140	ay:	**s_tar12**, tr4
148	tcl:	tr1, **tr_tar**, tr
156	h#:	s_tar

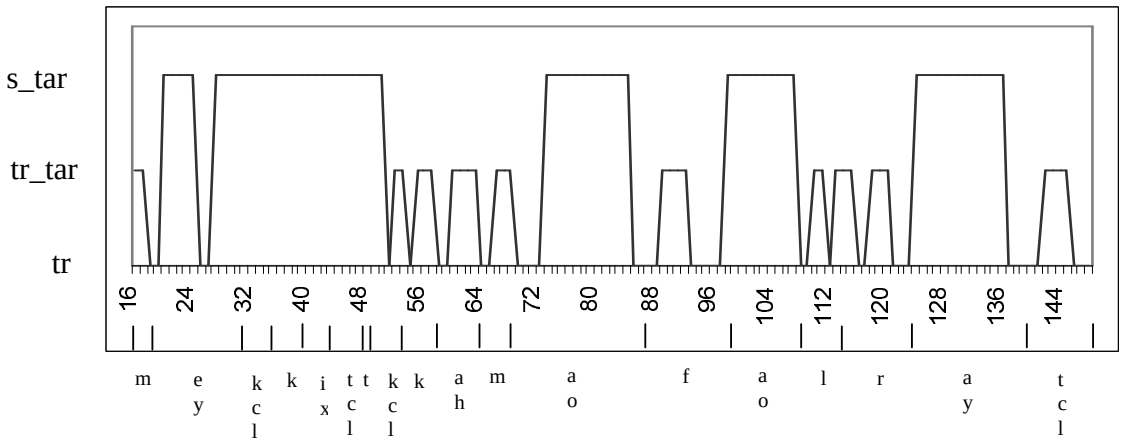


Figure 5.6 One example result of recovered syllable patterns

We have also shown that it is in fact possible to recover the desired details of syllables from speech without resorting to statistical models of phone sequences, or to models of

the syllable as a sequence of phones. This suggests that the syllable is a valuable articulatory unit for speech recognition processing. Additionally, the work demonstrates for the first time the potential of processing speech to yield independent structures and characteristics, each of which can be assessed separately in terms of linguistic and articulatory plausibility, before being combined in a speech recognition system (see figure 5.7). We believe that speech recognition systems must exploit as many interpretations of the incoming signal as possible. Other independent sources of information should also be considered to further enrich the 'multithreaded' base, for example, timing and sonority, and future work will consider these possibilities.

In figure 5.7, source 1 is the conventional route via HMM. Source 2 is the work reported in this thesis. Sources 3 and 4 are possible future techniques. Alignment of recovered details from sources 1 and 2 has not been attempted but is possible (dashed arrows). The value of deriving multiple independent "low-level" representations of speech from the waveform lies in the reliability given to subsequent recovery and complex linguistic levels of representations.

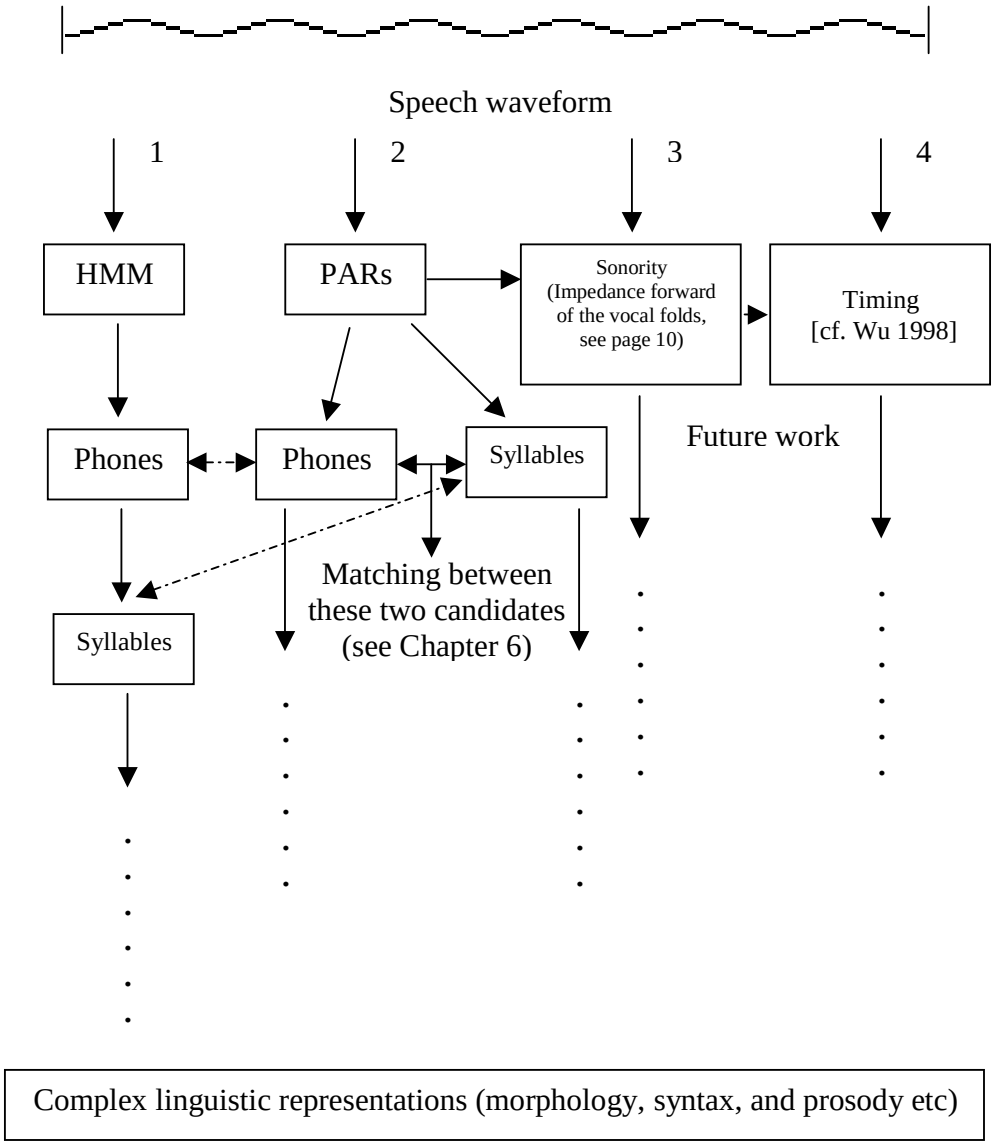


Figure 5.7 The derivation of multiple independent sources of information

6 FINDING A PHONEME SEQUENCE – RECOGNITION III

This chapter presents the third stage of the recognition process, which focuses on the transition from the syllable patterns to the phonetic level of description and produces sequences of phonemes. Having obtained the recovered syllable patterns, the next task is to derive phoneme sequences based on these articulatory descriptions. Since speech recognition is in most cases a speech-to-text conversion, one is interested in the orthographic or phonetic transcription as a result of the recognition process. Therefore, the three-stage recognition system described here ends up with phoneme sequences, since this type of output is uniform for most recognizers regardless of the approach. Dynamic programming is used to find phoneme sequences for the recovered syllabic targets. The phoneme recognition results are presented in the form of confusion matrix. As expected, the results differ for different sound classes. Vowels perform reasonably well, whereas some of the fricatives are rarely matched with the correct labels. Then syllable and phoneme recognition were repeated on another speaker (ID: dr1/mdac0) from the TIMIT corpus. Evaluation results are discussed as well. On the whole, the results are regarded as promising and the approach can be seen as a viable method of incorporating more phonetic and phonological knowledge into the recognition process.

6.1 Dynamic programming

In order to deal with time scale difference recognition systems use dynamic programming to achieve alignment between the incoming speech and the stored models [Holmes 1988]. The Viterbi algorithm is a dynamic programming algorithm, too, which is applied to probabilities rather than distances [Mariani 1989]. Dynamic programming is a mathematical technique which applies the optimal non-linear time-scale distortion to achieve the best match at all points [Holmes 1988]. It guarantees determination of the cumulative distance along the optimal path without having to calculate the distance along all the possible paths. It deals with time scale differences by applying penalties for time distortions. The total difference is calculated by summing all the distances (or

probabilities) between the individual pairs of frames along whichever path from the beginning to the end gives the smallest distance (or highest probability).

The idea of dynamic programming as developed by Holmes is used in the recognition system described here. It provides the final alignment between the pseudo-articulatory trajectories corresponding to the recovered syllabic targets (syllable targets and transition targets) and phonemes. An algorithm introduced by Sakoe and Chiba [1978] seems most appropriate in this case. Originally applied to word recognition it is extended here to continuous speech recognition. In equation 6.1, at every point in time t^*, the distance D is calculated between a new incoming vector, characterized by four feature values of high, back, round, tense (h, b, r, t), and every phone model p, characterized by another set of four feature values ($p[h_i]$, $p[b_i]$, $p[r_i]$, $p[t_i]$). The distance is further modified by the phone duration d:

$$D(t^*, p, d) = \left(|h - p[h_i]| + |b - p[b_i]| + |r - p[r_i]| + |t - p[t_i]|\right) + \frac{(d-\mu)^2}{\delta^2}$$

Equation 6.1

where μ is the mean duration of a phone and δ^2 the variance, $d \leq 2\mu$. Since the aim is to compute the total distance for the whole target along the best model sequence, the best paths up to time $s = t^* - d$ are taken into account and the cost of extending them to model p at time t^* is calculated in equation 6.2:

$$Cum(t^*, p) = \min_{\substack{s<t^* \\ q}} Cum(s, q) + D(t^*, p, d)$$

Equation 6.2

where q is the best phone model sequence which leads to the best paths up to time $s = t^* - d$. Finally, the path resulting in the smallest total distance is chosen by finding the minimum cumulative distance over all the phones at time t^* in equation 6.3:

$$DistCum = \min_{p} Cum(t^*, p)$$

Equation 6.3

At that point the best path is recovered by tracing back step-by-step along the local distances and thus retrieving the phoneme sequence.

A tool which implements this algorithm was written by Iskra [2000]. It requires pseudo-articulatory trajectories (corresponding to the recovered syllabic targets) in the form of feature values as input and produces a sequence of phonemes as output. In this way, phoneme candidates are derived for all the targets in one utterance. This has been done for all the 10 utterances (speaker ID: dr1/mmrp0). The recognition results are discussed in detail in the following section.

6.2 Phoneme recognition results

Dynamic programming is applied to the pseudo-articulatory trajectories via the recovered syllabic targets (syllable targets and transition targets as explained on page 68) to derive phoneme sequences. The phoneme models have been established for vowels by providing feature descriptions from phonetic textbooks, and for consonants by providing feature descriptions from the mapping procedure of Iskra [2000]. But we have no models for diphthongs, pause, silence and the begin/end marker[9] (see Appendix A.1). The distance is calculated between the incoming pseudo-articulatory vectors and the reference values, and further modified by the duration statistics (the mean and the variance) for each phoneme. Finally, the path leading to the smallest distance is chosen as the optimal one

[9] Nor for some allophones (e.g. the American tongue flap /d/ (see Appendix A.1)) either because the data is not available or because they are rare in the dataset.

and the phoneme sequence for one syllabic target is determined by backtracking along it. Consequently, the final recognition results for one utterance are presented as a sequence of phonemes with reasonable transitions between them.

In reality, there will be transitions between adjacent phones. The recognition results we have presented try to reflect what happens in the real articulation. Since the TIMIT database provides phones and phone durations one after another without any transitions, in order to make the comparison between the original transcription files in the TIMIT and our recognition results possible, we need to relabel the original transcription files in the TIMIT to provide the phone sequence with reasonable transitions between them. In the results of recovered syllable patterns presented in Appendix D, the duration of transitions divided by the duration of syllabic targets (syllable targets and transition targets) for the 10 utterances ranges from 0.29 to 0.41 (0.29, 0.30, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.41). Hence 0.40 is chosen as the ratio to label the TIMIT original transcription files. For example, if the duration of a phone in TIMIT is 10 records, then the beginning 2 records and the ending 2 records will be transitions and the rest are the targets of the phones. The complete final recognition results are presented in Appendix E.1. The statistical results are produced based on this new standard (the transition imposed transcription files). Not only are the statistical results of phonemes presented, but also the statistical results of transitions.

In order to illustrate the phoneme recognition results graphically, we expand the phonemes over their duration. Therefore, if a phoneme is labelled to last 60ms (whether it is the original utterance or the recognized one), it will be counted as 6 “occurrences” of the same phoneme (10ms each). This is meant to evaluate not only the recognition of the phoneme, but to take into account its duration as well. Then a percentage is calculated by dividing the number of correctly/wrongly recognized phonemes by the number of all the occurrences of this phoneme in the newly created “original” TIMIT transcription files. The total recognition statistical results for all the phoneme classes are presented in the figure 6.1. The higher the recognition percentage, the darker the shading. Only some of

the phoneme labels are visible. They are ordered in sound classes with silence/noise, plosives, affricates, fricatives, nasals, approximants, vowels and transition from left to right and bottom to up.

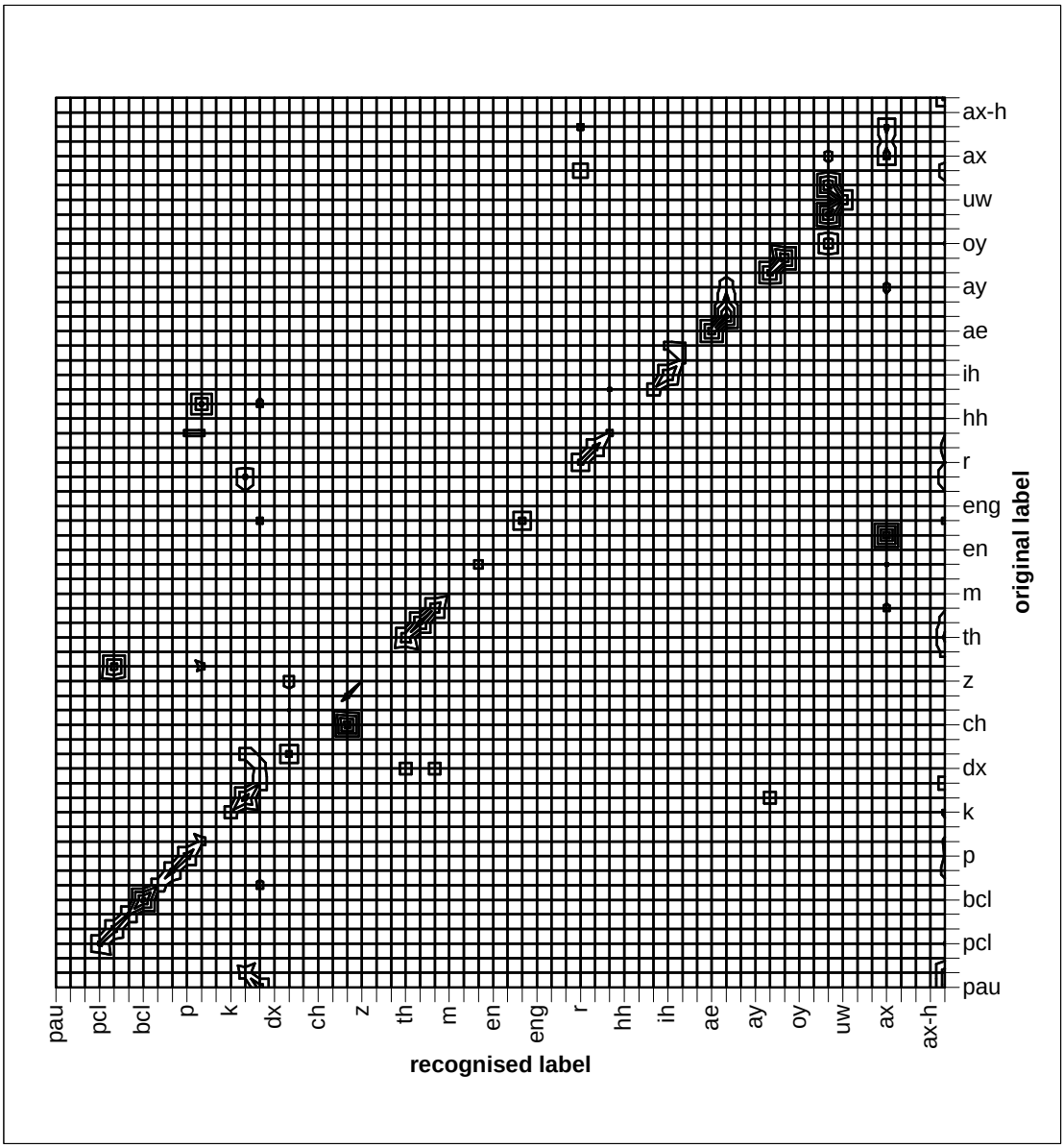


Figure 6.1 Phoneme recognition results

(Space does not allow every label to be shown. Refer to Appendix E.2)

With 100% correct recognition figure 6.1 should display a clear diagonal line from the lower left to the upper right corner. The recognition percentages are very consistent and stable within every group of phonemes. However, there is also some confusion. The greatest confusion seems to take place among fricatives and plosives. They are recognized incorrectly within their own groups. But also fricatives, and even approximants are labelled as plosives in the recognition process. Moreover, the begin and the end marker /h#/, which is not modelled separately, but has a high rate of occurrence, is most frequently recognised as one of the plosives. It has to be kept in mind that some sounds such as diphthongs, some allophones or silence are not accounted for by separate models, therefore, they have to be ascribed to some other available modelled sounds.

Table 6.1 contains a fragment of the confusion matrix for some vowels, approximants and transition. The complete matrix is in Appendix E.2. The rows depict the phonemes as found in the original transcription files whilst the columns represent the recognized phonemes. The numbers on the diagonal, which are printed in bold, multiplied by 100 represent the correct recognition percentages for every phoneme. In the complete matrix, adding up all the numbers in a row results in 1.

The recognition percentages for vowels are highest of all, and among them the long vowel with 89% recognized correctly for /aa/. Improvements have been made for plosives too, since they are usually short and a match over a short duration is easier to establish, e.g., /bcl/ - 86%. The nasals and approximants follow with, e.g., 50% for /ng/. Some of the fricatives are recognized pretty well, e.g., 67% for /v/. On the whole, the fricatives and affricates do not do very well.

Since all the phones within one class present many similarities, the biggest of which is their manner of articulation, the confusion can be observed within various sound classes. For example for plosives, /dcl/ and /pcl/ are often recognized as /d/. Since there are no separate models for diphthongs, diphthongs are often recognized as the vowels which are parts of the signals of the diphthongs and are adjacent to them in the confusion matrix.

For nasals, /n/ is often recognized as /ng/ as well. Some of the other confusions are caused by limited speech data so that not enough examples are available to produce a representative model, such as /ch/, /m/, /l/, and /f/. Since earlier, we choose a comparatively large ratio (0.4) to impose transitions into the original TIMIT transcription files, transitions are often wrongly "recognized" as phonemes.

Orig/rec	r	w	y	iy	ih	ae	aa	ah	ao	uh	uw	ax	tr
r	**0.48**	0.00	0.00	0.00	0.00	0.02	0.05	0.04	0.00	0.08	0.00	0.03	0.20
w	0.00	**0.47**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.27
y	0.00	0.00	**0.25**	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.19
iy	0.00	0.00	0.22	**0.35**	0.05	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.02
ih	0.00	0.00	0.00	0.00	**0.60**	0.00	0.00	0.00	0.08	0.00	0.00	0.03	0.08
ae	0.09	0.00	0.00	0.00	0.00	**0.77**	0.04	0.04	0.00	0.00	0.00	0.04	0.02
aa	0.00	0.00	0.00	0.00	0.00	0.00	**0.89**	0.00	0.00	0.00	0.00	0.00	0.11
ah	0.00	0.00	0.00	0.00	0.00	0.00	0.14	**0.73**	0.00	0.00	0.00	0.09	0.04
ao	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.82**	0.06	0.00	0.00	0.00
uh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.92**	0.00	0.00	0.08
uw	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	**0.58**	0.00	0.05
ax	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.27	0.00	**0.50**	0.09
tr	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.01	0.04	**0.47**

Table 6.1 A fragment of the confusion matrix

In general, it is clear that some classes of sounds are recognized better than others, which is not unexpected. Vowels, approximants and nasals have very stable recognition scores. These are the classes of sounds well known for their consistency, clarity and steadiness in their phonetic realization. These are also the sounds which can be described most adequately with the features selected earlier (high, back, round, tense). Not surprisingly, the fricatives and affricates pose major problems, which is a case well known in automatic speech recognition and is due to the acoustic nature of these sounds. Therefore, future efforts to improve the recognition results will concentrate on this class of sounds.

It might be argued that these results are biased because the training and testing data are from the same 10 utterances (speaker ID: dr1/mmrp0). To determine whether this is the case, syllable and phoneme recognition were repeated on another speaker (ID: dr1/mdac0) from the TIMIT corpus. 8 different utterances have been used to do the experiments. One sentence example of the recognition results for syllables is presented in Appendix D.

For syllables, the same evaluation method was used. The average accuracy rate for all the targets (duration of correctly labelled syllable targets and transition targets divided by total duration of phonemes) is 41.5%, which is effectively the same as the result of previous experiments (42.1%). This suggests that syllable recovery performs very stably.

Using this new dataset, the insertion rate during syllable recovery can be analyzed. For example, we might have the following syllable recovery results for one utterance.

End position of segment	Duration	Actual	Expected
10	10	t: tr4, tr_tar2, tr4	t: tr4, tr_tar2, tr4
17	7	f: tr1, tr_tar3, tr1, *tr_tar1*, tr1	f: tr1, tr_tar3, tr3
27	10	ah: tr1, s_tar6, tr1, *tr_tar1*, tr1	ah: tr1, s_tar6, tr3

Table 6.2 One pseudo example of recovered syllable results

The italics will be counted as insertion since they are unexpected. The insertion rate is the sum of the durations of the insertion divided by the total duration of these phonemes. In this example, the insertion rate is 2/27. In this way, insertion rates were produced for all 8 utterances. The average insertion rate is 13.1%. The insertion in the recovered syllable structures may be introduced by the syllable recovery algorithm. It may also be introduced by the smoothing procedure. Future efforts will be made to solve the high insertion rate problem for syllable recovery.

The opportunity was taken to evaluate the phoneme results for this data in a more standard way. The HResults tool from HTK [Young, Evermann, Kershaw, Moore, Odell, Ollason, Povey, Valtchev and Woodland 2002] was used to produce various evaluation results, such as a confusion matrix and overall error rates in terms of correct (Corr), accuracy (Acc), insertions (Ins), deletions (Del), substitutions (Sub), errors (Err), etc. The detailed results are presented below.

```
------------------------ Overall Results --------------------------
WORD: %Corr=61.99, Acc=21.18 [H=398, D=7, S=237, I=262, N=642]
=================================================================
  |=============================================================|
  |              Corr    Sub    Del    Ins    Err    S. Err     |
  |-------------------------------------------------------------|
  | Sum/Avg      61.99   36.92  1.09   40.81  78.82  100.00     |
  '-------------------------------------------------------------'
```

Table 6.3 Evaluation results for the new speaker (ID: dr1/mdac0)

The above evaluation results are obtained without counting the beginning and ending noise (h#). According to the HTK's manual [Young et al. 2002], "The first line is the word accuracy based on the dynamic programming matches between the label files and the transcriptions. H is the number of correct labels, D is the number of deletions, S is the number of substitutions, I is the number of insertions and N is the total number of labels in the defining transcription files". Other evaluation results obtained using HResults for the 8 utterances, such as a confusion matrix, are displayed in Appendix F.

In order to compare the phoneme recognition results of the new speaker (ID: dr1/mdac0) with the results of the previous speaker (ID: dr1/mmrp0), we also use HResults to evaluate the previous recognition results. The overall error rates are listed in table 6.4.

The data presented in table 6.3 and table 6.4 show that a recognition model developed on the basis of a single speaker (ID: dr1/mmrp0) can be successfully applied to recognition

of speech produced by a different speaker (ID: dr1/mdac0), though the recognition results of the new speaker (ID: dr1/mdac0) are worse than those of the previous speaker (ID: dr1/mmrp0). The two speakers were chosen from TIMIT corpus of the same dialect region. Thus they are not strongly different in accent. The data therefore demonstrate the potential for speaker independent recognition.

```
----------------------- Overall Results -------------------------
  WORD: %Corr=66.67, Acc=34.17 [H=478, D=19, S=220, I=233, N=717]
=================================================================
```

	Corr	Sub	Del	Ins	Err	S. Err
Sum/Avg	66.67	30.68	2.65	32.50	65.83	100.00

Table 6.4 Evaluation results for the previous speaker (ID: dr1/mmrp0)

On the whole, the recognition results can be regarded as promising. The approach offers a viable alternative for incorporating more phonetic knowledge into the recognition process, which is worthwhile pursuing further.

6.3 Evaluation

An appropriate way to evaluate the current approach to recognition is to compare it with another approach, e.g., the feature-based two-stage system developed by Iskra [2000]. In these two approaches, the same speech material (10 utterances from speaker ID: dr/mmrp0) has been used for training and testing. Iskra's system was to derive phoneme sequences directly from recovered pseudo-articulatory trajectories without any smoothing procedure, after having obtained feature trajectories from the incoming signal. Since there are considerable peaks and troughs in the recovered pseudo-articulatory trajectories, this kind of direct derivation of phonemes may provide a lot of "noise" data, even mistakes. Additionally, without knowing any other detail, this approach provided a very coarse mapping between the phoneme reference and the feature trajectories. In the new approach, we recover syllabic details from smoothed pseudo-articulatory trajectories.

Then the recovered syllabic details provide additional information for the derivation of phoneme sequences. Moreover it is very valuable that the syllabic targets identify the positions to be scrutinized for the identification of candidate phonemes. All of these give this new approach a finer mapping between the phoneme reference and the incoming feature vectors. Moreover, it also provides statistical data for transitions at the phone boundaries. By comparing the final results of these two approaches in figure 6.1 and 6.2, the new approach outperforms Iskra's system. Looking in detail at the confusion matrix, for every sound class, recognition percentages have been improved greatly by this new approach. Altogether, the results are encouraging and the syllable-based pseudo-articulatory approach is worth pursuing further.

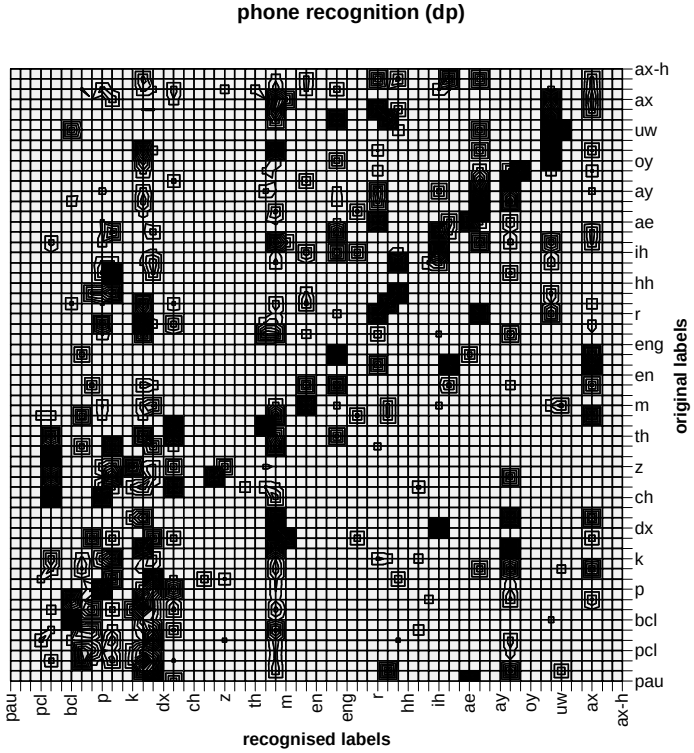


Figure 6.2 Iskra's recognition results

6.4 Discussion and conclusions

The phoneme recognition results determined by the alignment between the pseudo-articulatory trajectories corresponding to the syllabic targets and the phoneme models are very diverse and range from very good for some vowels to very poor for some fricatives. Considerable confusion can be observed within various sound classes. All the phones within a class present many similarities, the biggest of which is their manner of articulation. They usually differ with regard to the place of articulation. That refers not only to the point in the vocal tract where the sound is produced, but also indicates which articulators are primarily involved in the production of that particular sound. The recognition errors in the data suggest that perhaps the four features used to describe sounds are not powerful enough to capture sufficient acoustic detail to discriminate between them reliably. Four features are not many to describe the whole phone inventory. Although they are perfectly adequate for the description of vowels and it was expected that ascribing them with continuous values would create enough space for consonant discrimination, perhaps some additional features should be used. These will have to be chosen with the purpose of extracting finer acoustic characteristics in order to increase discriminative power within consonant classes in particular.

It should be pointed out that the original TIMIT database is not an ideal database for our approach. In reality and from linguistic point of view, there will be a transition between adjacent phones. But the TIMIT database provides phones and phone durations one after another without any transitions, which conflicts with our approach. In our approach, the processing takes transitions between adjacent phones into account and can often find transitions at phone boundaries, which is plausible. In order to evaluate our work, we have to impose transitions into the TIMIT transcriptions. Since this is not an optimal procedure, it can easily account for some of the mistakes in the recognition where phoneme labels in the recognition files are encountered in the preceding or following transition in the transition imposed original alignment.

Additionally, the original TIMIT transcription is not always reliable. TIMIT is a very large speech database, hence the manual labelling of it proved too time consuming. The speech waveform was, therefore, aligned automatically with the acoustic-phonetic sequence using an alignment programme developed at MIT [Zue and Seneff 1988]. Although the boundaries were then corrected manually, errors still exist. Moreover, some of the boundary assignment criteria will always remain subjective. Since the original transcription files work as the basis of the imposition of transitions and the frame accuracy evaluation, this can easily contribute to the errors in the confusion matrix.

The use of recovered syllable structures to find phonetic segments has never been attempted before and the success reported here is very rewarding. The processing based on syllable structures yields promising phoneme recognition results. Introducing a few changes, such as improving phoneme models or correcting the time alignment problem, has the potential for ensuring more satisfactory recognition results, which can then serve as the basis for introducing more speech data and more speakers. The approach at this stage cannot compete with the results obtained with statistical models, but offers a viable alternative for incorporating more phonetic knowledge into the recognition process, which is worthwhile pursuing further.

7 CONCLUSIONS

This chapter presents a brief summary of the work presented in this thesis and also directions for future work. Since the current PAR descriptions are not sufficient for the description of consonants, a new PAR description is introduced as one possible direction for future work. Additionally, the refinement of our syllable model is also discussed as one line of the work to be pursued.

7.1 Summary and conclusions

The purpose of designing the recognition system described in this thesis is to create a speech recognition approach which incorporates more phonetic and linguistic knowledge into recognition processing. In order to achieve this, the idea of pseudo-articulatory trajectories is first introduced as the intermediate level of description during the recognition processing. The pseudo-articulatory trajectories are based on the concept of phonological and phonetic features, with values ranging from 0 to 100 in order to comply more with phonetic reality and account for such effects as coarticulation. Then vowels and consonants are described by using four pseudo-articulatory features: high, back, round, and tense. Phone models for vowels are taken from phonetic textbooks, while phone models for consonants are taken from previous work [Iskra 2000]. A set of vowels is used to perform the mapping between the four features and their formant parameters based on the speech data (speaker ID: dr1/mmrp0) from TIMIT. This is done in order to set the system up. Then everything is ready to run recognition processing.

The first stage of the recognition processing concerns the transition from the acoustic representation of the incoming signal to the pseudo-articulatory space and for every incoming formant vector four feature values are found using a brute search algorithm. The newly derived trajectories clearly contain many peaks and troughs, though the general trend coincides with the idealized PAR trajectories. In order to evaluate them, the recovered trajectories are converted back into cepstral values, which are subsequently synthesized. The quality of synthesized speech is very good proving that the mapping has

been satisfactory. The success of this part of the research implies that the articulatory-acoustic mapping problem concerning its many-to-one nature can be successfully solved.

In the first stage of the processing, the waveforms are analyzed every 10ms continuously. Thus there is no rigid segmentation for the derivation of pseudo-articulatory trajectories. Consequently, the information recovered is not segmental and it is possible to directly recover both syllabic information (intermediate level) and even phoneme sequences from the recovered trajectories. The latter is a step we try to avoid because of its poor recognition results (see page 95) and the coarse mapping. With additional syllabic details recovered from the signal, the former provides us a plausible articulatory description with transitions and targets. Hence, we choose syllabic details as another intermediate level to bridge the recovered trajectories and phoneme sequences.

The next stage of the processing is the derivation of syllable structures from the pseudo-articulatory representation, which produces a sequence of recovered syllables. In this process, an unconventional articulatory syllable model is used (see page 68). Since the newly derived trajectories clearly contain too many peaks and troughs, a smoothing algorithm is used to provide additional constrains and make them plausible from the physiological point of view. Then the conventional resynthesis procedure is used again to evaluate the smoothed trajectories. The quality of the synthesized speech is judged to be very good proving that the smoothing procedure has been satisfactory. After running the syllable recovery algorithm, meaningful syllabic details are derived for all the 10 utterances (speaker ID: dr1/mmrp0) from smoothed trajectories. The transition targets (tr_tar) and syllable targets (s_tar) are very well recognized. The average accuracy rate for all the targets (duration of correctly labelled syllable targets and transition targets divided by total duration of phonemes) is 42.1%, which is very promising. In general, we have shown that it is in fact possible to recover the desired details of syllables from speech without resorting to statistical models of phone sequences, or to models of the syllable as a sequence of phones.

The third stage of the recognition process focuses on the transition from the syllable patterns to the phonetic level of description and produces sequences of phonemes. Having obtained the recovered syllable patterns, the next task is to derive phoneme sequences based on these articulatory descriptions. Dynamic programming is used to provide the best path between the pseudo-articulatory trajectories corresponding to the syllabic targets and phoneme models. The recognition percentages (see Appendix E.2) are very consistent and stable within every group of phonemes, which is very important for a speech recognition system. Using HResults for evaluation, we obtained 34.2% word accuracy rate for the testing of the 10 utterances for the speaker ID: dr1/mmrp0.

The overall experiments have been conducted on another speaker (ID: dr1/mdac0) from the TIMIT corpus. We obtained 21.2% word accuracy rate for the testing of 8 different utterances. Although the results of these two speakers cannot compete with the state-of-the-art recognition systems, the accuracy rates are very promising. The syllable-based, pseudo-articulatory approach presented in this thesis shows its potential for speaker independent recognition. In general, despite some clear flaws, the approach is worth pursuing further as a viable method of incorporating more phonetic and phonological knowledge into the recognition process.

Finally, the whole work is summarized diagrammatically in figure 7.1. It illustrates that recovered syllable structures D are used to target portions of A to identify (via PARs) phoneme candidates to match labeled segments in B (Appendix E.2) and also to identify transitions in D to match imposed transitions in C (Appendix E.2).

7.2 Future directions

In our approach, three different sets of recognition results are obtained as the outcome of the three stages of the recognition process: the derived pseudo-articulatory trajectories, the recovered syllable patterns and the phoneme sequences with transitions at the phoneme boundaries. We have identified four fundamental questions to be addressed,

arising from the analysis of the three-stage recognition process. The first two research questions will be discussed in detail.

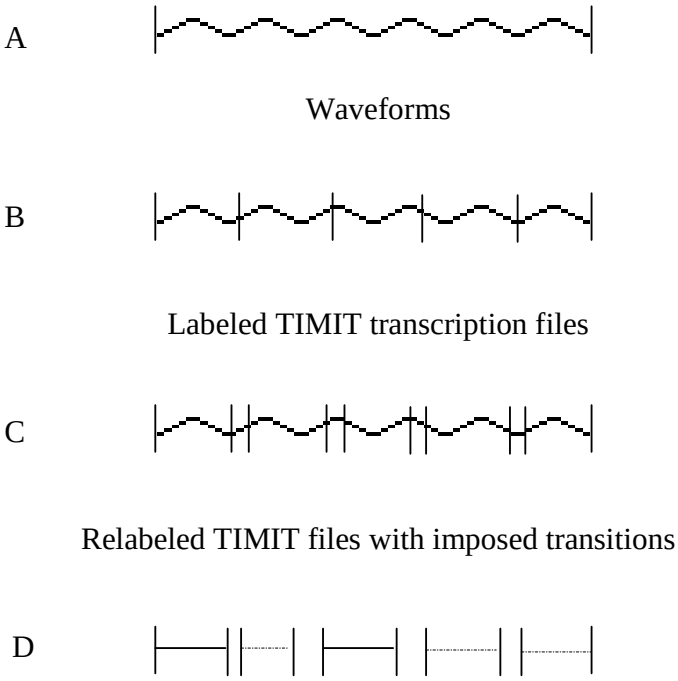


Figure 7.1 Summary of work done

Firstly, in order to improve the recovery of PAR trajectories extra features might be needed for the descriptions of all the sounds. Despite their continuous values the current four features seem too few to describe all the sounds adequately. Especially in the case of consonants, they seem not efficient enough to capture all the important acoustic detail, which contributes to the number of recognition errors. Additionally, it seems that the problem also concerns the choice of features. The features used so far are perfectly adequate for the description of vowels and it was expected that they will be sufficient for

the description of consonants. However, major confusion within some consonant classes indicates that some extra features should probably be used in order to distinguish between these sounds. The division of consonants into classes such as stops, fricatives, affricates etc, is based on the manner of articulation, while within one class they differ primarily with the place of articulation and the primary articulator involved in the production of that sound. Consequently, additional features should be analyzed and evaluated with a view to capturing this kind of distinction.

The recognition results of recovered syllable patterns also suggest that the smoothed pseudo-articulatory trajectories sometimes lose important information, and sometimes are too sensitive. So far, only one averaging algorithm has been used to smooth the trajectories. The utilisation of different smoothing algorithms in different conditions can probably solve this problem.

Some of the confusions are also caused by inadequate speech data so that not enough examples are available to produce a representative model (PAR). Especially for diphthongs, there are no separate models for them at all. Future efforts will concentrate on improving phone models, which will probably improve the recognition results significantly. Additionally, PARs are derived from *linguistic* specifications of articulatory activity. Such linguistically approximate accounts of articulatory activity make PARs suitable for speaker independent and language independent tasks. Preliminary experiments show that the PAR approach can go beyond single speakers. We believe that more data and more speakers can be handled in the PAR approach without many difficulties and can also contribute to the improvements of phone models. Nevertheless, though PARs are abstract and idealized accounts of articulatory activity, it is possible that the PAR space is still rich enough to account for details which could be speaker or language dependent. For example, some aspects of speaker characterization (such as accent) may be specified in linguistically relevant terms in PAR space, which provides valuable information for speaker identification. The full implications of working with PARs remain to be explored.

Secondly, our research work is based on recovery of syllable structural information directly from the speech stream, i.e. using pseudo-articulatory trajectories without resorting to phone segment identification. This independently recovered syllable information has contributed significantly to speech recognition processing [Zhang and Edmondson 2002a; 2003]. One line of work to be pursued is refinement of the syllabic template model expressed in terms of targets and transitions [Edmondson and Zhang 2002]. Additionally, this thread can be augmented by recovery of other related material from the speech stream. In particular, independent measures of timing information and sonority should provide independent contributions to complement the derived syllabic structure and also enrich the 'multithreaded' base (see figure 5.7). We believe that speech recognition systems must exploit in combination as many interpretations of the incoming signal as possible.

Thirdly, it has been pointed out in the discussion of phoneme recognition results that many errors may be due to the non-optimal imposition of transitions into the TIMIT transcription files and the faulty time alignment in the TIMIT transcription files. Since the TIMIT database provides phones and phone durations one after another without transitions between adjacent phones, which conflicts with our approach, this makes TIMIT transcription files less than ideal for the evaluation of our approach. Additionally, the imposition of transitions into the original transcription files is a rather crude device to limit errors at the boundaries. Furthermore, the accuracy in time alignment in TIMIT transcription files is very important as the basis of the imposition and the frame accuracy used as the recognition measure. Any inaccuracy in time alignment will result in recognition errors. Hence, it would be interesting to run the same experiments using a different speech database, one which provided transitions at phone boundaries and careful manual transcriptions of the speech data. This is expected to contribute to the increase of correct recognition.

Finally, we are concerned to see how our system performs when using more different speech data in the training and testing set. The prediction is that the recognition results will deteriorate, as they usually do when more speakers are modelled [Lee 1989a], although the pseudo-articulatory representation is in principle capable of providing for generalization across speakers and languages. With more speakers extra inter-speaker variability is imposed on the system, but more speech data provide more contexts and contribute to the training of more representative phone models – this balance needs to be explored. In addition, we can also try to record our own data for training and testing set. The specific goal of this aspect of the investigation is to provide the system with the capability for speaker identification.

7.2.1 One possible choice of new features

In the current work, the PAR description is obtained by selecting four articulatory features: high, back, round, tense, and ascribing a value between 0 and 100 to every vowel based on the data provided by Ladefoged [1993]. In one area for the future work a new PAR description will probably be used. Depending on the position of the tongue, whether it is in the front or the back of the mouth, up or down, vowels can be represented as located on a quadrilateral; i.e. the Cardinal Vowel System. We will adapt the vowel quadrilateral by using the lower left corner as the origin of a polar coordinate system. Then we will use 'r' (called 'distance' below) as a measure of the degree of constriction and 'θ' (called 'theta' below) as a measure of the vocal tract length. The new description will probably be obtained by replacing the features high and back with the new features theta and distance. Finally, the new PAR description will be obtained by selecting four features – theta, distance, round, tense – and ascribing a value between 0 and 120 for theta, a value between 0 and 150 for distance.

In the current work the consonants are located on the vowel quadrilateral – or more precisely, just outside the periphery. In order to determine new PAR values for consonants the diagram suggested by Ball [2002], which is closer to physical reality, will be adopted. As in the current work this permits the account of vowels to be extended to

cover consonants (in terms of target configurations, but not dynamic aspects). Our interest in Ball's proposal is that it offers the possibility that the acoustic parameters recovered from speech can be more easily linked to the palatal, velar, uvular and pharyngeal places of articulation, and thus to articulatory behaviour and phonetic categories. We hope the new features can provide enough space for consonants' discrimination, especially within one class.

After the establishment of a new vowel model and consonant model with the new features, we will repeat the earlier experiments with these new models. In addition, we will explore the opportunity for deriving a new pseudo-articulatory feature based on the linguistic phonological feature 'coronal' which describes the deployment of the tongue in a blade-like configuration. This too will lead to new vowel and consonant models and further experimentation.

7.2.2 Refinement of syllable model

The syllable template model needs improvements in the predictions of sonority and timing information. A description of the link between syllable templates and sonority is provided by Durand [1990]. This is a valuable starting point for our future theoretical work, despite being based on segments rather than phonetic/articulatory events. As described in chapter 5.2.1, syllables are 'sonority waves'. Our derived syllable structure information can be used to produce basic sonority waves, which may serve as the base for the prediction of sonority. There are also various issues to be analyzed related to timing, such as durations of transitions, durations of phonetic events, and time alignment with conventional HMM phone sequences. The syllable model needs to be refined not just to enhance its inherent value as one of the threads in the system, but because it will serve as the base to which will be anchored the other related interpretations of the data (timing, sonority, phonetic values).

APPENDICES

Appendix A TIMIT data

This appendix contains data related to the TIMIT database.

Appendix A.1 Phone representation

This section contains all the phone symbols used in the TIMIT database together with example words. The phones are divided into following classes: plosives, fricatives, nasals, affricates, semivowels and glides, vowels and diphthongs, and other symbols. The phones which have not been modeled separately are marked with an asterisk.

symbol	*example word*	*possible phonetic transcription*
b	bee	BCL B iy
d	day	DCL D ey
g	gay	GCL G ey
p	pea	PCL P iy
t	tea	TCL T iy
k	key	KCL K iy
dx*	muddy, dirty	m ah DX iy, dcl d er DX iy (flap)
q*	bat	bcl b ae Q (glottal stop)

Table A.1 Plosives (stops)

symbol	*example word*	*possible phonetic transcription*
s	sea	S iy
sh	she	SH iy
z	zone	Z ow n
zh	azure	ae ZH er
f	fin	F ih n
th	thin	TH ih n
v	van	V ae n
dh	then	DH eh n
m	mom	M aa M
n	noon	N uw N
ng	sing	s ih NG
em*	bottom	b aa dx EM
en*	button	b ah q EN
eng*	washington	w aa sh ENG tcl t ax n
nx*	winner	w ih NX axr (nasal flap)

Table A.2 Fricatives and nasals

symbol	*example word*	*possible phonetic transcription*
jh*	joke	DCL JH ow kcl k
ch	choke	TCL CH ow kcl k

Table A.3 Affricates

symbol	*example word*	*possible phonetic transcription*
l	lay	L ey
r	ray	R ey
w	way	W ey
y	yacht	Y aa tcl t
hh*	hay	HH ey
hv	ahead	ax HV eh dcl d
el*	bottle	bcl b aa dx EL

Table A.4 Semivowels and glides

symbol	*example word*	*possible phonetic transcription*
iy	beet	bcl b IY tcl t
ih	bit	bcl b IH tcl t
eh	bet	bcl b EH tcl t
ey*	bait	bcl b EY tcl t
ae	bat	bcl b AE tcl t
aa	bott	bcl b AA tcl t
aw*	bout	bcl b AW tcl t
ay*	bite	bcl b AY tcl t
ah	but	bcl b AH tcl t
ao	bought	bcl b AO tcl t
oy*	boy	bcl b OY
ow*	boat	bcl b OW tcl t
uh	book	bcl b UH kcl k
uw	boot	bcl b UW tcl t
ux*	toot	tcl t UX tcl t
er*	bird	bcl b ER dcl d
ax	about	AX bcl b aw tcl t
ix*	debit	dcl d eh bcl b IX tcl t
axr*	butter	bcl b ah dx AXR
ax-h*	suspect	s AX-H s pcl p eh kcl k tcl t

Table A.5 Vowels and diphthongs

symbol	*description*
pau*	pause
epi*	epenthetic silence
h#*	begin/end marker (non-speech events)
1	primary stress
2	secondary stress

Table A.6 Others

Appendix A.2 Speech data transcription

This section contains all the sentences used in the training and test set. The original sentence transcriptions in the TIMIT are also provided. The sentences are divided into three groups: dialect, phonetically-diverse and phonetically-compact sentences.

speaker ID: dr1/mmrp0

dialect sentences

sa1: She had your dark suit in greasy wash water all year.

h# sh iy hv ae dcl d y axr dcl d aa r kcl k s ux tcl ih n gcl g r iy s iy w ao sh epi w ao dx axr q ao l y ih axr h#

sa2: Don't ask me to carry an oily rag like that.

h# d ow nx ae s kcl k m iy dx ix kcl k eh r iy ix n q oy l iy r ae gcl g l ay kcl k dh ae tcl h#

phonetically-diverse sentences

si2034: Make it come off all right.

h# m ey kcl k ix tcl t kcl k ah m ao f ao l r ay tcl h#

si717: Now here is truly a marvel.

h# n aw hv ih r ix z tcl t r uw l ih ey m aa r v el h#

si774: Place work on a flat surface and smooth out.

h# p l ey s epi w er kcl k q ao nx ey f l ae tcl t s er f ix s pau ae n dcl s epi m uw dh q aw tcl h#

phonetically-compact sentences

sx144: The willowy woman wore a muskrat coat.

h# dh ax w ih l ah w iy w uh m ix n w ao ey m ah s kcl k r ae tcl k ow tcl h#

sx234: The paper boy bought two apples and three ices.

h# dh ax pcl p ey pcl p axr bcl b oy bcl b ao tcl t uw q ae pcl p el z ae n dcl th r iy q ay s ix z h#

sx324: The local drugstore was charged with illegally dispensing tranquillisers.

h# dh ax l ow kcl k el dcl d r ah gcl g s tcl t ao ax w ax z tcl ch aa r z dcl d w ix th ih l iy gcl el iy dcl d ix s pcl p eh n tcl s ix ng tcl t r ae ng kcl k w el ay z ix z h#

sx414: Irish youngsters eat fresh kippers for breakfast.

h# ay r ix sh y ah ng gcl g s tcl t ix z q iy tcl t f r eh sh kcl k ih pcl p ix z f ax bcl b r eh kcl f ix s tcl t h#

sx54: The eastern coast is a place for pure pleasure and excitement.

h# dh ih q iy s tcl t ix n kcl k ow s tcl t ih z ey pcl p l ey s f ax pcl p y uh pcl p l eh zh axr ix n ix kcl k s ay tcl m ix n tcl h#

speaker ID: dr1/mdac0

phonetically-diverse sentences

si1261: In two cases, airplanes only were indicated.

h# q ix n tcl t ux kcl k ey s ih z pau q eh r pcl p l ey n z q ow n l iy w axr ih n dcl d ix kcl k ey dx ix dcl h#

si1837: Her hum became a gurgle of surprise.

h# hh axr hv ah m bcl b ix kcl k ey m ix gcl g er gcl el ax v s ax-h pcl p r ay z h#

si631: Program note reads as follows: take hands; this urgent visage beckons us.

h# p r ow gcl g r ih m n ow tcl r iy dcl d z ax-h z f aa l ow z pau t ey kcl k hh ae n dcl z pau dh ih s q er dcl jh en tcl v ih z ih dcl jh bcl b eh kcl k en z ah s h#

phonetically-compact sentences

sx181: Rich looked for spotted hyenas and jaguars on the safari.

h# r ih tcl ch epi l uh kcl t f axr s pcl p aa dx ih dcl d hv ay y iy n ah z pau hh ix n dcl jh ae gcl w aa z en dh s ax f aa r iy h#

sx271: Be careful not to plow over the flower beds.

h# b ix kcl k eh f el n aa tcl t ax-h pcl p l aw ow v axr dh ix f l aw axr bcl b eh dcl d z h#

sx361: The speech symposium might begin Monday.

h# dh ix s pcl p iy tcl ch s em pcl p ow z iy ax m pau m ay tcl b ax-h gcl g ih n m ah n dcl d ey h#

sx451: The thick elm forest was nearly overwhelmed by Dutch Elm Disease.

h# dh ix th ih kcl k q eh l m f ao r ih s epi w ax z n ih l iy q ow v axr w eh l m dcl b ay dcl d ah sh q eh l m dcl d ix z iy z h#

sx91: The misprint provoked an immediate disclaimer.

h# dh ix m ih s pcl p r ix n tcl p ax-h v ow kcl t ix n em iy dcl d iy ix tcl d ix s kcl k l ey m axr h#

Appendix B Regression coefficients

The text concerning this analysis can be found in Chapter 3, section 3.4.

	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
f1	1324.412	-12.6422	-7.48402	0.80659	1.419374	-0.031	0.077635	0.018874	0.018129	-0.02359	-0.00249
f2	2355.62	-3.23688	-16.7005	0.720341	2.402089	-0.15412	0.167824	0.031391	0.008956	-0.01488	-0.0207
f3	3119.687	2.517716	0.598372	-9.70967	-2.66603	-0.05824	0.016528	-0.00681	0.033043	0.058161	-0.00448
f4	4447.684	-5.28086	-1.40161	-7.19231	-3.37093	-0.06211	0.067955	0.01057	-0.03083	-0.00508	0.046524
f5	6784.247	-14.0304	-20.1792	18.04175	-4.97166	-0.05296	-0.05555	0.107504	-0.02921	-0.15905	0.261896
f6	6707.43	6.906682	-4.58879	11.76722	-1.82831	0.042113	-0.20295	0.017488	-0.01437	-0.07135	0.12113
a1	49.44774	-0.07061	0.002645	0.061247	-0.02978	0.000939	-0.00073	-0.00011	0.000131	0.000137	-0.00019
a2	65.45606	-0.0925	-0.04986	0.100072	-0.0095	0.000621	-0.00081	-0.00022	-0.00035	0.00042	0.000165
a3	71.4319	0.006275	0.033965	-0.16962	-0.0606	-0.00185	0.002589	0.00079	1.19E-06	-0.00093	0.000947
a4	68.99002	-0.05449	0.12922	-0.29294	-0.02228	-0.00426	0.006437	0.000417	-0.0004	-0.0003	0.000154
a5	13.47493	0.561385	0.497421	-0.05586	-0.21928	0.001028	-0.00318	-0.00045	-0.00121	0.000132	0.001877
a6	36.00435	0.157543	0.335084	-0.32674	-0.20641	-0.00261	0.003351	0.001154	-0.00057	-0.00173	0.002769
b1	70	-8.28E-14	-1.24E-13	-4.28E-14	-1.93E-14	5.79E-16	3.68E-16	2.16E-16	8.56E-16	6.56E-16	-5.29E-16
b2	343.971	-3.02935	-4.36232	2.549435	0.440257	-0.00948	0.008175	0.007036	0.002631	-0.00652	0.00392
b3	141.648	0.692462	0.907116	-1.10127	-0.68466	-0.00592	-0.00123	-0.00013	-0.00552	-0.00315	0.011061
b4	217.3282	-0.3133	1.917032	-2.21659	-1.67583	-0.00878	0.007322	0.006306	-0.00574	-0.01439	0.026138
b5	212.409	-0.53337	-0.43085	0.171256	0.150265	-0.00125	0.002012	0.00149	0.001684	-0.00289	0.000818
b6	102.7559	1.395924	1.057708	-0.6092	-0.27285	0.003984	-0.00555	-0.00338	-0.00306	0.006655	-0.00293

Table B.1 Regression coefficients

a_0 to a_{10} in the column headings are the regression coefficients and refer to equation 3.2 on page 46. They should not be confused with a1 to a6 in the row labels which are values of y_i (formant amplitudes).

Appendix C Consonant feature values

This appendix contains mapping results for the experiments in the form of pseudo-articulatory values for consonants. In table C.1, 'myData' refers to the consonant feature values obtained by our experiment (see section 3.5). 'No. correct' in last row refers to the number of consonant feature values falling into the control range.

consonant	high		back		round		tense	
	myData	control	myData	control	myData	control	myData	control
w	0	67-100	68	34-66	57	67-100	77	0-33
l	75	34-66	72	0-33	83	0-33	51	0-33
r	94	34-66	80	0-33	100	0-33	98	0-33
y	93	67-100	6	34-66	12	0-33	100	0-33
m	35	34-66	22	0-33	25	0-33	1	0-33
n	36	34-66	1	0-33	32	0-33	38	0-33
ng	99	67-100	42	67-100	45	0-33	100	0-33
p	2	34-66	52	0-33	10	0-33	1	67-100
pcl	8	34-66	49	0-33	14	0-33	47	67-100
b	100	34-66	14	0-33	22	0-33	71	0-33
bcl	85	34-66	0	0-33	32	0-33	100	0-33
t	29	34-66	33	0-33	3	0-33	84	67-100
tcl	65	34-66	0	0-33	22	0-33	100	67-100
d	57	34-66	32	0-33	21	0-33	52	0-33
dcl	55	34-66	0	0-33	15	0-33	41	0-33
k	76	67-100	11	67-100	6	0-33	67	67-100
kcl	84	67-100	0	67-100	39	0-33	55	67-100
g	98	67-100	0	67-100	6	0-33	66	0-33
gcl	98	67-100	29	67-100	60	0-33	0	0-33
hv	76	0-33	59	34-66	64	0-33	29	0-33
f	43	34-66	3	0-33	29	0-33	47	67-100
v	14	34-66	19	0-33	39	0-33	22	0-33
th	40	34-66	2	0-33	27	0-33	47	67-100
dh	57	34-66	41	0-33	28	0-33	82	0-33
s	83	34-66	0	0-33	59	0-33	95	67-100
z	70	34-66	0	0-33	56	0-33	72	0-33
sh	100	67-100	0	0-33	25	0-33	97	67-100
zh	39	67-100	0	0-33	25	0-33	44	0-33
ch	44	67-100	0	0-33	25	0-33	0	67-100
No. correct	15		17		19		9	

Table C.1 Consonant feature values

Appendix D Recognition results – recovered syllable patterns

This appendix contains recognition results of recovered syllable patterns for individual sentences. All results are given for speaker ID: dr1/mmrp0. One example is given for speaker ID: dr1/mdac0. On the left-hand side there are record numbers and original phone symbols as found in the TIMIT transcription files. Following the colon there are the recognized syllabic details. A crude time alignment is used here. If the phone boundary, however, does not coincide with the target boundary or the transition boundary in the recognized sequence, a number is placed after the target symbol or the transition symbol. This number refers to the number of records aligned with the particular original label. This is why the numbers can be found only at boundaries. The correctly recognized targets (in accord with TIMIT) are printed in bold, i.e. where both the target and the time overlap.

Speaker ID: dr1/mmrp0

sa1

50 h#: s_tar, tr, tr_tar, tr, s_tar, tr, tr_tar, tr
60 sh: s_tar10
67 iy: **s_tar7**
76 hv: s_tar7, tr
88 ae: **s_tar**, tr
92 dcl: s_tar4
94 d: s_tar2
97 y: tr, **tr_tar**
105 axr: tr, tr_tar, tr, **s_tar**, tr1
112 dcl: tr1, **tr_tar**, tr
113 d: **tr_tar1**
128 aa: tr_tar2, tr, **s_tar**

```
132    r:   tr, tr_tar
138  kcl:   tr, tr_tar, tr, tr_tar
140    k:   tr, tr_tar
153    s:   tr, s_tar, tr
169   ux:   tr_tar, tr, s_tar, tr4
177  tcl:   tr2, tr_tar, tr
187   ih:   s_tar10
192    n:   s_tar3, tr, tr_tar1
198  gcl:   tr_tar4, tr
201    g:   tr_tar
208    r:   tr, tr_tar, tr, tr_tar, tr
215   iy:   s_tar7
225    s:   s_tar8, tr
231   iy:   s_tar, tr
242    w:   tr_tar, tr, tr_tar, tr, tr_tar, tr1
260   ao:   tr2, s_tar, tr
271   sh:   tr_tar, tr, tr_tar, tr3
274  epi:   tr1, tr_tar, tr
280    w:   s_tar6
290   ao:   s_tar9, tr1
292   dx:   tr1, tr_tar
300  axr:   tr, s_tar, tr1
305    q:   tr1, s_tar4
318   ao:   s_tar13
330    l:   tr, tr_tar, tr, tr_tar, tr
343    y:   s_tar
354   ih:   tr, s_tar
361  axr:   tr, s_tar
388   h#:   tr, tr_tar, tr, tr_tar, tr, tr_tar, tr, s_tar,
             tr, tr_tar
```

sa2

```
127   h#:   s_tar, tr, tr_tar, tr, s_tar, tr, tr_tar, tr,
             tr_tar, tr, s_tar, tr, tr_tar, tr, s_tar, tr,
             tr_tar, tr, s_tar, tr1
128    d:   tr1
145   ow:   tr3, tr_tar, tr, tr_tar, tr
146   nx:   tr_tar
162   ae:   tr, s_tar, tr
168    s:   tr_tar, tr, s_tar2
175  kcl:   s_tar5, tr
177    k:   tr_tar
184    m:   tr, tr_tar, tr, tr_tar1
192   iy:   tr_tar1, tr, s_tar, tr1
196   dx:   tr1, s_tar3
200   ix:   s_tar4
206  kcl:   tr, s_tar4
213    k:   s_tar5, tr
221   eh:   tr_tar, tr, tr_tar, tr
228    r:   tr_tar, tr
235   iy:   s_tar7
240   ix:   s_tar5
243    n:   tr, tr_tar
249    q:   tr, s_tar4
266   oy:   s_tar5, tr, tr_tar, tr, tr_tar, tr
273    l:   tr_tar
280   iy:   tr, s_tar5
```

```
291    r:   s_tar8, tr
309   ae:   s_tar, tr, tr_tar, tr1
312  gcl:   tr2, tr_tar
316    g:   tr, tr_tar
320    l:   tr, tr_tar1
333   ay:   tr_tar5, tr, s_tar, tr, tr_tar1
336  kcl:   tr_tar2, tr1
339    k:   tr2, tr_tar
342   dh:   tr, s_tar1
365   ae:   s_tar20, tr
370  tcl:   tr_tar, tr1
376   h#:   tr2, tr_tar
```

si2034

```
15    h#:   s_tar, tr, tr_tar, tr
19     m:   tr_tar, tr
32    ey:   s_tar, tr, s_tar6
36   kcl:   s_tar4
40     k:   s_tar4
44    ix:   s_tar4
49   tcl:   s_tar5
50     t:   s_tar1
54   kcl:   tr, tr_tar, tr
59     k:   tr_tar, tr
65    ah:   tr_tar, tr
69     m:   tr_tar, tr1
87    ao:   tr3, s_tar, tr3
```

```
99     f:   tr1, tr_tar, tr, s_tar2
109   ao:   s_tar8, tr
115    l:   tr_tar, tr, tr_tar
124    r:   tr, tr_tar, tr, s_tar1
140   ay:   s_tar12, tr4
148  tcl:   tr1, tr_tar, tr
156   h#:   s_tar
```

si717

```
64    h#:   s_tar, tr, tr_tar, tr, s_tar, tr, tr_tar, tr,
             s_tar, tr, tr_tar, tr, tr_tar, tr, tr_tar, tr
67     n:   tr_tar, tr, s_tar1
81    aw:   s_tar3, tr, s_tar, tr, tr_tar
92    hv:   tr, s_tar, tr1
101   ih:   tr1, s_tar
111    r:   tr, tr_tar
117   ix:   tr, s_tar5
127    z:   s_tar5, tr, s_tar4
130  tcl:   s_tar3
139    t:   s_tar8, tr
143    r:   tr_tar, tr, s_tar2
149   uw:   s_tar5, tr
158    l:   tr_tar, tr, s_tar1
164   ih:   s_tar6
174   ey:   s_tar10
185    m:   tr, tr_tar, tr, s_tar, tr, tr_tar, tr
198   aa:   s_tar13
```

205 r: s_tar4, tr, **tr_tar**, tr
210 v: tr_tar, tr, **tr_tar2**
218 el: tr_tar2, tr, tr_tar, tr
248 h#: s_tar

si774

23 h#: s_tar, tr, s_tar15
30 p: s_tar5, tr2
36 l: tr1, **tr_tar**, tr1
48 ey: tr2, **s_tar10**
57 s: s_tar6, tr, **tr_tar**
60 epi: tr, tr_tar1
70 w: **tr_tar3**, tr, s_tar4
85 er: **s_tar7**, tr, tr_tar, tr, s_tar1
89 kcl: s_tar4
92 k: s_tar1, tr
95 q: **tr_tar**, tr
105 ao: **s_tar**
108 nx: tr, s_tar2
118 ey: **s_tar7**, tr, **s_tar2**
132 f: s_tar3, tr, s_tar, tr
140 l: **tr_tar**, tr, s_tar4
153 ae: **s_tar8**, tr, tr_tar3
157 tcl: **tr_tar1**, tr, tr_tar2
160 t: **tr_tar1**, tr
170 s: tr_tar, tr, **tr_tar**, tr, tr_tar, tr, tr_tar
186 er: tr, **s_tar**, tr, tr_tar1

```
195    f:   tr_tar2, tr, tr_tar, tr, tr_tar
202   ix:   tr, tr_tar, tr, s_tar1
212    s:   s_tar10
219  pau:   tr, tr_tar, tr, s_tar2
228   ae:   s_tar9
231    n:   tr, tr_tar
234  dcl:   tr, tr_tar, tr
247    s:   tr_tar, tr, tr_tar, tr, tr_tar, tr, s_tar
251  epi:   tr, s_tar1
256    m:   s_tar5
272   uw:   s_tar14, tr
278   dh:   tr_tar, tr, tr_tar1
284    q:   tr_tar2, tr, s_tar3
307   aw:   s_tar17, tr, s_tar
318  tcl:   tr, tr_tar, tr, s_tar6
350   h#:   s_tar
```

sx144

```
65    h#:   s_tar, tr, tr_tar, tr, tr_tar, tr, tr_tar, tr,
             s_tar, tr, tr_tar, tr, tr_tar, tr, tr_tar
67    dh:   tr, tr_tar1
73    ax:   tr_tar1, tr, tr_tar, tr2
85     w:   tr3, s_tar
93    ih:   tr, s_tar5
98     l:   s_tar4, tr
104   ah:   s_tar, tr
111    w:   tr_tar, tr, tr_tar, tr, tr_tar, tr
```

120 iy: **s_tar**
131 w: tr, tr_tar, tr, tr_tar, tr, **tr_tar**
138 uh: tr, **s_tar**
144 m: tr, **tr_tar**, tr
152 ix: **s_tar**
157 n: tr, s_tar4
167 w: s_tar2, tr, **tr_tar**, tr, s_tar2
179 ao: **s_tar8**, tr, tr_tar2
193 ey: tr_tar1, tr, **s_tar**, tr, **s_tar**, tr1
203 m: tr2, **tr_tar**, tr, s_tar1
213 ah: **s_tar5**, tr, s_tar2
222 s: s_tar7, tr, **tr_tar1**
226 kcl: tr_tar1, tr, **tr_tar2**
229 k: tr_tar1, tr, **tr_tar**
236 r: tr, **tr_tar**, tr, s_tar3
248 ae: **s_tar5**, tr, tr_tar, tr
257 tcl: s_tar9
263 k: s_tar3, tr
281 ow: **s_tar**, tr, **s_tar**, tr
299 tcl: tr_tar, tr, **tr_tar**, tr, s_tar11
316 h#: s_tar17

sx234

60 h#: s_tar, tr, s_tar, tr, tr_tar, tr, tr_tar3
63 dh: **tr_tar1**, tr, tr_tar1
68 ax: tr_tar2, tr, tr_tar2
76 pcl: **tr_tar1**, tr, s_tar6

```
80     p:   s_tar2, tr, s_tar1
92    ey:   s_tar10, tr, tr_tar1
98   pcl:   tr_tar1, tr, tr_tar3
100    p:   tr_tar1, tr
105  axr:   s_tar, tr
111  bcl:   s_tar, tr
113    b:   tr_tar, tr
130   oy:   s_tar, tr, s_tar
138  bcl:   tr, s_tar, tr
139    b:   s_tar1
155   ao:   s_tar12, tr, tr_tar
170  tcl:   tr, tr_tar, tr, s_tar3
177    t:   s_tar7
185   uw:   s_tar2, tr, s_tar5
191    q:   s_tar2, tr, s_tar1
206   ae:   s_tar15
214  pcl:   tr, tr_tar, tr, tr_tar1
216    p:   tr_tar2
228   el:   tr, s_tar, tr, s_tar2
238    z:   s_tar7, tr, s_tar1
246   ae:   s_tar4, tr, tr_tar2
250    n:   tr_tar2, tr
257  dcl:   s_tar, tr1
260   th:   tr3
269    r:   tr1, s_tar
279   iy:   tr, s_tar, tr
283    q:   tr_tar, tr, tr_tar1
301   ay:   tr_tar1, tr, s_tar, tr, s_tar, tr
313    s:   tr_tar, tr, s_tar, tr
320   ix:   s_tar
```

```
334    z:   tr, tr_tar, tr
348   h#:   tr_tar, tr, tr_tar, tr, tr_tar, tr, tr_tar
```

sx324

```
60    h#:   s_tar, tr, tr_tar, tr, s_tar, tr, s_tar, tr,
             tr_tar, tr, s_tar, tr
62    dh:   tr_tar2
66    ax:   tr_tar1, tr, s_tar1
75     l:   s_tar3, tr, tr_tar, tr
88    ow:   s_tar, tr, s_tar, tr
93   kcl:   tr_tar, tr, tr_tar2
96     k:   tr_tar2, tr1
104   el:   tr2, tr_tar, tr, tr_tar, tr
110  dcl:   tr_tar, tr, tr_tar2
113    d:   tr_tar1, tr
118    r:   s_tar5
129   ah:   s_tar9, tr, tr_tar1
133  gcl:   tr_tar1, tr, tr_tar1
135    g:   tr_tar1, tr
143    s:   s_tar8
147  tcl:   s_tar3, tr1
149    t:   tr1, s_tar1
158   ao:   s_tar6, tr, s_tar2
162   ax:   s_tar4
171    w:   tr, tr_tar, tr, tr_tar
177   ax:   tr, s_tar5
184    z:   s_tar3, tr, s_tar3
```

```
189  tcl:   s_tar5
197   ch:   s_tar6, tr2
209   aa:   tr1, s_tar, tr1
217    r:   tr1, tr_tar, tr, s_tar3
221    z:   s_tar4
223  dcl:   tr, tr_tar1
227    d:   tr_tar2, tr, tr_tar1
230    w:   tr_tar1, tr, s_tar1
235   ix:   s_tar4, tr
246   th:   tr_tar, tr, tr_tar, tr, tr_tar3
254   ih:   tr_tar1, tr, s_tar6
261    l:   s_tar3, tr, tr_tar, tr2
270   iy:   tr2, s_tar, tr2
274  gcl:   tr1, tr_tar3
282   el:   tr_tar1, tr, tr_tar, tr1
290   iy:   tr2, s_tar6
295  dcl:   s_tar4, tr
296    d:   tr_tar1
300   ix:   tr_tar1, tr, s_tar2
310    s:   s_tar10
317  pcl:   s_tar3, tr, tr_tar, tr1
318    p:   tr1
327   eh:   s_tar
329    n:   tr, tr_tar1
332  tcl:   tr_tar1, tr, tr_tar1
342    s:   tr_tar2, tr, tr_tar, tr, s_tar1
347   ix:   s_tar5
351   ng:   s_tar1, tr, tr_tar
358  tcl:   tr, tr_tar2
367    t:   tr_tar2, tr, tr_tar, tr, s_tar1
```

371 r: s_tar4
381 ae: **s_tar8**, tr2
383 ng: tr1, **tr_tar1**
388 kcl: tr_tar2, tr, s_tar2
392 k: s_tar2, tr2
396 w: tr1, **tr_tar**, tr1
402 el: tr4, s_tar2
421 ay: **s_tar8**, tr, tr_tar, tr, tr_tar, tr, tr_tar1
427 z: **tr_tar3**, tr, s_tar2
438 ix: **s_tar11**
448 z: s_tar10
462 h#: s_tar6, tr, tr_tar

sx414

67 h#: tr_tar, tr, s_tar, tr, tr_tar, tr, s_tar, tr, s_tar, tr, s_tar, tr, s_tar4
81 ay: **s_tar14**
88 r: s_tar5, tr2
95 ix: tr2, **s_tar5**
104 sh: s_tar9
109 y: s_tar1, tr, **tr_tar**, tr2
117 ah: tr2, **s_tar**, tr
120 ng: **tr_tar**, tr, tr_tar1
124 gcl: **tr_tar1**, tr, tr_tar2
126 g: **tr_tar1**, tr
130 s: **tr_tar**
135 tcl: tr, **tr_tar3**

```
138    t:   tr_tar1, tr
147   ix:   tr_tar, tr, s_tar4
153    z:   s_tar6
159    q:   s_tar6
168   iy:   tr, s_tar
172  tcl:   tr, tr_tar3
175    t:   tr_tar1, tr, tr_tar1
186    f:   tr_tar1, tr, s_tar
190    r:   tr, tr_tar
197   eh:   tr, tr_tar, tr, s_tar2
206   sh:   s_tar9
212  kcl:   tr, tr_tar, tr
218    k:   tr_tar, tr, tr_tar1
223   ih:   tr_tar2, tr3
229  pcl:   tr1, tr_tar, tr
231    p:   tr_tar
239   ix:   tr, tr_tar, tr, s_tar2
246    z:   s_tar7
256    f:   s_tar4, tr, tr_tar3
263   ax:   tr_tar1, tr, s_tar
272  bcl:   tr, tr_tar, tr, tr_tar1
274    b:   tr_tar1, tr
277    r:   s_tar3
284   eh:   s_tar6, tr1
290  kcl:   tr1, tr_tar, tr3
299    f:   tr2, tr_tar, tr, tr_tar, tr2
307   ix:   tr2, s_tar
318    s:   tr, s_tar6
321  tcl:   s_tar3
326    t:   s_tar5
```

342 h#: s_tar3, tr, s_tar, tr

sx54

79 h#: s_tar, tr, s_tar, tr, tr_tar, tr, tr_tar, tr, tr_tar, tr, s_tar, tr, tr_tar, tr, tr_tar, tr
82 dh: **tr_tar**
87 ih: tr, tr_tar, tr
92 q: s_tar5
103 iy: **s_tar11**
112 s: s_tar2, tr, s_tar
114 tcl: tr, **tr_tar1**
117 t: tr_tar2, tr
123 ix: **s_tar**, tr
128 n: **tr_tar**, tr, s_tar1
134 kcl: s_tar6
140 k: s_tar2, tr, **tr_tar**
155 ow: tr, **s_tar**, tr, tr_tar, tr
161 s: **tr_tar**, tr, tr_tar1
164 tcl: **tr_tar3**
167 t: tr, **tr_tar1**
174 ih: tr_tar2, tr, **s_tar3**
178 z: s_tar3, tr1
187 ey: tr1, **s_tar**, tr, tr_tar1
193 pcl: **tr_tar1**, tr, tr_tar2
198 p: **tr_tar2**, tr
203 l: **tr_tar**, tr, tr_tar
211 ey: tr, tr_tar, tr, **s_tar2**

```
216    s:   s_tar5
224    f:   tr, tr_tar, tr, s_tar2
230   ax:   s_tar6
238  pcl:   tr, s_tar7
245    p:   s_tar1, tr, tr_tar, tr, tr_tar2
250    y:   tr_tar1, tr, s_tar2
260   uh:   s_tar5, tr, tr_tar, tr1
268  pcl:   tr2, tr_tar, tr
272    p:   tr_tar, tr, tr_tar2
278    l:   tr_tar1, tr, s_tar3
287   eh:   s_tar4, tr, s_tar3
293   zh:   s_tar6
305  axr:   tr, tr_tar, tr, s_tar, tr, s_tar1
312   ix:   s_tar5, tr, s_tar1
315    n:   s_tar3
321   ix:   s_tar3, tr
326  kcl:   tr_tar, tr, tr_tar
328    k:   tr, tr_tar
340    s:   tr, tr_tar, tr
356   ay:   s_tar, tr, s_tar
364  tcl:   tr, tr_tar, tr, tr_tar, tr, tr_tar1
366    m:   tr_tar1, tr
370   ix:   s_tar
373    n:   tr3
378  tcl:   tr2, tr_tar, tr1
385   h#:   tr2, s_tar
```

Speaker ID: dr1/mdac0

si1261

13 h#: tr_tar, tr, s_tar, tr, tr_tar, tr, s_tar1
15 q: s_tar2
23 ix: tr, **s_tar**, tr, tr_tar2
31 n: **tr_tar2**, tr, s_tar5
34 tcl: s_tar3
37 t: tr, **tr_tar**, tr
50 ux: **s_tar**, tr, s_tar1
52 kcl: s_tar2
54 k: s_atr2
68 ey: tr, **s_tar**, tr, **s_tar1**
78 s: s_tar5, tr, **tr_tar**, tr, s_tar1
87 ih: **s_tar9**
99 z: tr, s_tar, tr, **tr_tar**, tr
114 pau: tr_tar, tr, **tr_tar**, tr, tr_tar, tr
116 q: **tr_tar**, tr
125 eh: **s_tar**, tr
132 r: s_tar
134 pcl: tr
140 p: **tr_tar**, tr3
145 l: tr1, **tr_tar**, tr
154 ey: **s_tar**, tr
160 n: **tr_tar**, tr, tr_tar1
164 z: **tr_tar3**, tr1
168 q: tr1, s_tar3
182 ow: **s_tar6**, tr, s_tar6
188 n: s_tar3, tr3

```
191    l:   tr1, tr_tar
195   iy:   tr, tr_tar, tr2
199    w:   tr1, tr_tar, tr
211  axr:   s_tar, tr
218   ih:   s_tar, tr
224    n:   tr_tar, tr2
226  dcl:   tr1, tr_tar1
227    d:   tr_tar1
235   ix:   tr_tar, tr
237  kcl:   tr_tar2
240    k:   tr_tar1, tr, s_tar1
252   ey:   s_tar4, tr, s_tar, tr, s_tar2
255   dx:   s_tar3
258   ix:   s_tar2, tr
264  dcl:   s_tar6
270   h#:   s_tar6
```

Appendix E Recognition results – phoneme candidates and transitions

This appendix contains the final recognition results of phonemes and transitions for individual sentences. And the results are also evaluated using a confusion matrix.

Appendix E.1 Phoneme candidates and transitions for individual sentences

This section contains recognition results of recovered phoneme candidates and reasonable transitions at phoneme boundaries. As presented in Appendix D, on the left-hand side there are record numbers, original phone symbols (as found in the TIMIT transcription files) and the imposed transitions. Following the colon there are the recognized phoneme sequence details with reasonable transitions. A crude time alignment has been attempted here as well. There is a number after each recognized symbol. And it refers to the number of records aligned with the particular original label. The correctly recognized phonemes and transitions (in accord with TIMIT) are printed in bold, i.e. where both the recognized symbol and the time overlap.

sa1

50 h#: tcl17, tr1, p2, tr1, d5, tcl5, d12, tr1, p2, tr1
52 tr: p1, g1
58 sh: g3, t3
61 tr: t2, d1
66 iy: d5
68 tr: d2
73 hv: d3, t2
77 tr: t1, **tr2**, ae1
84 ae: **ae7**
88 tr: **tr4**
92 dcl: **dcl3**, d1
93 tr: gcl1

```
94     d:   gcl1
95    tr:   tr1
97     y:   y2
98    tr:   tr1
102  axr:   ax2, tr1, ax1
106   tr:   ax2, tr2
109  dcl:   w3
112   tr:   tr3
113    d:   p1
117   tr:   dh2, tr2
127   aa:   tr3, aa7
130   tr:   aa1, tr2
132    r:   r2
134   tr:   tr2
135  kcl:   g1
139   tr:   tr1, dh2, tr1
140    k:   k1
142   tr:   tr2
150    s:   tr2, dcl4, g2
158   tr:   g1, tr2, ax3, tr2
165   ux:   tr1, uh6
170   tr:   tr5
174  tcl:   tr1, tcl3
179   tr:   tcl1, tr2, ih2
184   ih:   ih4, ng1
188   tr:   ng4
190    n:   ng1, n1
193   tr:   tr1, w2
195  gcl:   w2
199   tr:   w1, tr2, b1
```

```
201    g:   b2
202   tr:   tr1
205    r:   r2, tr1
209   tr:   tr1, dh1, tr1, d1
212   iy:   d3
217   tr:   d2, p3
222    s:   pcl5
225   tr:   pcl1, tr2
228   iy:   th3
233   tr:   th1, tr2, w1, tr1
238    w:   w3, tr2
244   tr:   ax3, tr3
255   ao:   ao11
261   tr:   tr5, p1
268   sh:   p2, tr2, ah3
272   tr:   tr4
273  epi:   b1
274   tr:   tr1
277    w:   w3
282   tr:   w3, ao2
287   ao:   ao5
292   tr:   ao2, tr2, dh1
293   dx:   tr1
295   tr:   tr2
298  axr:   ax2, r1
301   tr:   r1, tr2
304    q:   ao3
307   tr:   ao3
317   ao:   ao10
320   tr:   ao1, tr2
```

325 l: b3, tr2
331 tr: tr2, ax1, **tr2**, p1
339 y: p3, t4, p1
347 tr: p3, d1, **tr4**
353 ih: **ih6**
355 tr: ih1, **tr1**
359 axr: ax4
363 tr: ax2, **tr2**
388 h#: tr1, ax3, tr2, ah1, tr2, p1, d3, tr1, tcl5, tr1, g5

sa2

127 h#: dcl9, kcl5, k5, dcl16, d6, tcl5, k5, tcl14, tr1, p2, tr1, sh1, t4, tcl5, k5, d2, tr2, g2, tr1, g2, tr1, ih7, tr1, g1, tr1, tcl4, tr1, g2, tr1, tcl14, tr1
128 d: tr1
131 tr: **tr3**
139 ow: dh3, tr2, uh3
145 tr: uh3, **tr3**
146 nx: ax1
149 tr: **tr2**, ae1
157 ae: **ae8**
162 tr: **tr5**
165 s: b2, tr1
169 tr: **tr1**, ch3
172 kcl: ch3

```
176   tr:   ch1, tr2, dh1
177    k:   dh1
179   tr:   tr2
181    m:   b1, m1
184   tr:   tr2, m1
191   iy:   m1, tr1, y5
194   tr:   tr2, d1
196   dx:   d1, th1
198   tr:   th2
200   ix:   th2
202   tr:   tr2
204  kcl:   kcl2
208   tr:   sh4
211    k:   sh3
215   tr:   tr2, k2
218   eh:   k1, tr1, y1
221   tr:   y2, tr1
225    r:   r2, tr2
228   tr:   tr3
232   iy:   uh4
235   tr:   uh3
238   ix:   uh2, ax1
241   tr:   ax2, tr1
243    n:   th1, n1
245   tr:   tr2
247    q:   ax2
252   tr:   ax2, n3
261   oy:   n2, tr1, uh3, tr3
266   tr:   tr2, uh2, tr1
271    l:   l5
```

```
275   tr:   l2, tr2
280   iy:   ih3, iy2
282   tr:   iy2
288    r:   uh4, r2
295   tr:   tr3, r1, ae3
304   ae:   ae9
311   tr:   tr2, p1, g1, tr3
312  gcl:   gcl1
314   tr:   tr2
316    g:   g2
318   tr:   tr2
320    l:   tr1, l1
323   tr:   ax3
329   ay:   ax2, tr2, ax2
333   tr:   ax1, tr2, p1
336  kcl:   p2, tr1
338   tr:   tr2
339    k:   b1
341   tr:   tr2
342   dh:   dh1
346   tr:   dh2, ah2
359   ae:   ah3, ae10
366   tr:   ae3, tr3, dh1
369  tcl:   dh1, gcl2
372   tr:   tr3
376   h#:   d4
```

si2034

```
14    h#:   tcl1, g2, gcl4, tr2, bcl3, tr2
15    tr:   tr1
17     m:   r2
19    tr:   tr2
28    ey:   ih5, tr2, tcl2
32    tr:   kcl4
36   kcl:   kcl3, d1
37    tr:   d1
40     k:   d3
41    tr:   d1
44    ix:   tcl3
45    tr:   tcl1
47   tcl:   tcl1, t1
49    tr:   t2
50     t:   t1
51    tr:   tr1
53   kcl:   p1, kcl1
55    tr:   tr1, kcl1
57     k:   k2
60    tr:   tr2, ah1
62    ah:   ah2
66    tr:   ah1, tr2, b1
69     m:   b2, tr1
72    tr:   tr3
84    ao:   ao12
89    tr:   tr4, d1
94     f:   g3, tr2
99    tr:   tr3, ao2
105   ao:   ao6
```

```
110   tr:   ao2, tr2, b1
114    l:   b1, tr1, l2
117   tr:   l1, tr2
120    r:   r3
127   tr:   tr3, r4
136   ay:   ax9
141   tr:   tr5
145  tcl:   tcl3, th1
148   tr:   tr3
156   h#:   dcl3, pcl5
```

si717

```
63    h#:   k10, g1, tr1, dh2, tr1, d4, gcl4, tr2, p2, tr1,
             d22, tr1, dh2, tr2, t3, tr1, p4
64    tr:   tr1
65     n:   sh1
67    tr:   tr1, eh1
77    aw:   ax3, tr4, r3
83    tr:   r2, tr1, ah1, tr2
90    hv:   tr1, t6
93    tr:   t1, tr2
99    ih:   th2, ih4
103   tr:   ih2, tr2
110    r:   tr5, uh1, dh1
112   tr:   dh1, tr1
115   ix:   uh3
119   tr:   uh2, s1, tcl1
```

```
125    z:   tcl3, tr1, d2
128   tr:   d3
130  tcl:   d1, tcl1
131   tr:   tcl1
137    t:   tcl3, t1, y2
139   tr:   y1, tr1
143    r:   m1, tr1, r2
144   tr:   r1
148   uw:   b1, ao3
150   tr:   tr1, b1
156    l:   b2, ah1, tr3
159   tr:   tr1, ih2
163   ih:   ih4
166   tr:   ih3
172   ey:   ih3, ng3
177   tr:   ng2, tr1, ih1, tr1
182    m:   ng4, tr1
187   tr:   dh1, tr2, aa2
195   aa:   aa8
199   tr:   aa2, r2
203    r:   r3, tr1
206   tr:   ao1, tr1, th1
209    v:   v1, tr1, v1
211   tr:   v1, b1
215   el:   b1, tr2, p1
218   tr:   p2, tr1
248   h#:   g6, k5, pcl19
```

si774

```
23    h#:   dh5, tr3, g1, ch8, dcl4, p2
24    tr:   p1
28     p:   p4
31    tr:   tr3
35     l:   l2, dh2
38    tr:   tr3
45    ey:   d2, sh5
49    tr:   sh4
56     s:   sh2, g3, tr1, s1
58    tr:   s1, tr1
60   epi:   tr1, b1
62    tr:   b1, th1
68     w:   g1, tr3, b1, uh1
73    tr:   uh5
81    er:   uh2, dh2, tr1, r3
86    tr:   tr3, th1, kcl1
88   kcl:   kcl1, k1
89    tr:   k1
90     k:   k1
92    tr:   tr2
94     q:   dh2
97    tr:   tr1, ao2
104   ao:   ao7
106   tr:   ao1, tr1
108   nx:   ax2
110   tr:   ax2
115   ey:   ng5
120   tr:   tr1, g2, p2
```

```
129    f:   p1, tr1, f2, v5
132   tr:   tr3
137    l:   b1, ao1, b1, tr1, ae1
141   tr:   ae4
148   ae:   ae7
154   tr:   tr2, th1, p3
156  tcl:   tr1, dh1
157   tr:   dh1
159    t:   dh1, tr1
161   tr:   tr1, b1
167    s:   b1, tr1, b2, tr1, v1
173   tr:   tr2, th1, tr3
183   er:   tr5, dh1, r4
187   tr:   tr2, p2
192    f:   p1, tr2, dh2
196   tr:   tr1, p2, tr1
199   ix:   ax3
203   tr:   tr2, p1, g1
210    s:   g7
214   tr:   g2, tr2
216  pau:   d1, tr1
220   tr:   tr1, eh1, ae2
226   ae:   ae6
230   tr:   ae2, tr2
231    n:   ax1
232   tr:   tr1
234  dcl:   g1, tr1
237   tr:   b2, tr1
244    s:   d1, tr1, ax3, tr1, d1
248   tr:   d3, tr1
```

```
251  epi:   tr2, uw1
252   tr:   uw1
255    m:   uw3
259   tr:   uw4
268   uw:   uw6, w3
273   tr:   w2, tr2, dh1
276   dh:   dh2, tr1
279   tr:   tr1, d2
282    q:   d1, tr1, r1
288   tr:   r3, aa3
301   aw:   aa11, b2
308   tr:   tr1, uh5, tr1
315  tcl:   w1, d2, tr1, tcl3
318   tr:   tcl2, dcl1
350   h#:   dcl32
```

sx144

```
65    h#:   tcl6, k4, tr3, dh2, tr1, dh2, tr1, dh1, tr2,
             sh12, dcl4, ch8, p3, tr3, dh2, tr1, dh2, tr3,
             pcl5
66    tr:   tr1
67    dh:   ax1
69    tr:   ax1, tr1
71    ax:   ax2
75    tr:   tr4
83     w:   tr1, w7
88    tr:   b2, tr3
```

```
91    ih:   ao3
94    tr:   ao3
97     l:   ao2, dh1
99    tr:   tr1, ah1
102   ah:   ah3
106   tr:   tr2, ao1, tr1
109    w:   dh1, tr1, w1
112   tr:   tr2, gcl1
118   iy:   gcl6
122   tr:   b2, tr1, b1
128    w:   b1, tr1, th1, g1, tr1, w1
132   tr:   w2, b1, tr1
137   uh:   uh5
139   tr:   uh1, tr1
141    m:   ax2
144   tr:   tr3
150   ix:   dh1, ax5
153   tr:   ax2, tr1
156    n:   ax3
160   tr:   ax3, tr1
164    w:   w3, tr1
169   tr:   tr1, w2, b2
175   ao:   b2, uh4
181   tr:   tr2, r3, tr1
191   ey:   eh6, tr1, ih3
195   tr:   ih1, tr3
200    m:   g2, m2, tr1
204   tr:   tr2, ah2
209   ah:   ah4, tr1
213   tr:   tr2, g2
```

```
220    s:   g7
224   tr:   tr1, dh2, tr1
226  kcl:   l1, ax1
228   tr:   ax1, tr1
229    k:   k1
230   tr:   tr1
234    r:   r2, tr1, aa1
238   tr:   aa4
245   ae:   aa3, tr1, ax3
248   tr:   tr3
255  tcl:   dcl4, tcl3
258   tr:   tcl2, dcl1
260    k:   dcl2
265   tr:   tr3, eh2
278   ow:   eh2, tr1, ae3, r7
285   tr:   tr3, w1, tr3
295  tcl:   tcl2, tr1, d2, y5
299   tr:   d4
316   h#:   k5, dcl12
```

sx234

```
60    h#:   hv7, tcl4, tr1, p4, d2, sh31, tr3, eh1, dh1,
             tr3, pcl1, p2
61    tr:   p1
63    dh:   tr1, ax1
64    tr:   ax1
67    ax:   ax1, tr1, b1
```

```
70    tr:   b2, tr1
74   pcl:   d2, pcl2
77    tr:   p3
80     p:   p1, tr1, ih1
82    tr:   ih2
90    ey:   ih4, ax4
94    tr:   tr1, d2, tr1
96   pcl:   tr1, ah1
98    tr:   dh2
99     p:   dh1
100   tr:   tr1
104  axr:   ah4
105   tr:   tr1
109  bcl:   bcl4
111   tr:   bcl1, tr1
112    b:   b1
115   tr:   tr1, uh2
129   oy:   uh6, tr1, ah3, uh4
132   tr:   uh1, tr2
136  bcl:   bcl4
138   tr:   b1, tr1
139    b:   ah1
141   tr:   ah1, dh1
151   ao:   dh1, ao9
157   tr:   tr1, ah1, dh2, tr2
165  tcl:   tr1, tcl4, tr3
170   tr:   tr2, t3
174    t:   t2, d2
177   tr:   y3
183   uw:   uw2, tr1, uw3
```

```
185   tr:   uw2
188    q:   uw2, tr1
192   tr:   tr2, ae2
204   ae:   ae7, r5
209   tr:   r2, tr3
212  pcl:   pcl1, p1, tr1
213   tr:   tr1
216    p:   b1, ax2
218   tr:   tr1, b1
224   el:   b6
228   tr:   tr2, g2
235    z:   g7
238   tr:   tr2, ae1
243   ae:   ae2, r2, tr1
246   tr:   tr1, ah2
248    n:   dh2
250   tr:   tr2
256  dcl:   gcl2, d2, dcl2
258   tr:   tr2
260   th:   tr2
261   tr:   tr1
268    r:   b3, r4
272   tr:   r1, tr3
277   iy:   iy5
279   tr:   iy1, tr1
282    q:   ax2, tr1
286   tr:   dh2, tr2
298   ay:   tr1, aa2, r2, tr1, ih6
301   tr:   tr3
310    s:   v1, b2, tr3, p3
```

```
313   tr:   p2, tr1
318   ix:   gcl4, g1
323   tr:   g2, tr3
330    z:   tr2, p3, tr2
334   tr:   tr4
348   h#:   b2, tr1, p2, tr1, eh1, dh2, tr2, d3
```

sx324

```
59    h#:   dcl1, k10, d8, tcl5, d3, tr1, sh1, g1, tr1, d9,
             tr1, p6, tr2, eh1, dh2, tr2, tcl5
60    tr:   tr1
62    dh:   dh1, ax1
64    tr:   ax1, tr1
66    ax:   tr1, ax1
67    tr:   ax1
73     l:   b2, tr1, ax3
77    tr:   tr2, ao2
84    ow:   ao1, b1, tr1, b4
89    tr:   tr4, kcl1
92   kcl:   tr2, kcl1
94    tr:   kcl1, k1
96     k:   k1, tr1
98    tr:   tr2
102   el:   b2, tr1, dh1
106   tr:   dh1, tr1, dh1, tr1
109  dcl:   tr2, dcl1
110   tr:   d1
```

```
112    d:   d1, tr1
114   tr:   tr1, ah1
117    r:   ah3
119   tr:   ah2
127   ah:   ah3, aa3, ax2
130   tr:   tr1, b2
132  gcl:   tr2
133   tr:   g1
134    g:   g1
136   tr:   tr1, p1
141    s:   tcl5
143   tr:   tcl2
146  tcl:   tcl2, d1
148   tr:   tr2
149    t:   ng1
150   tr:   ng1
155   ao:   ng5
159   tr:   tr1, ax3
162   ax:   ax1, uh2
164   tr:   tr2
169    w:   w3, tr1, ax1
172   tr:   ax2, tr1
175   ax:   ax3
178   tr:   g2, tcl1
182    z:   tcl2, tr1, tcl1
184   tr:   tcl2
187  tcl:   tcl2, sh1
190   tr:   sh3
195   ch:   sh5
198   tr:   tr3
```

```
208   aa:   aa10
210   tr:   tr2
215    r:   ax2, tr2, p1
218   tr:   p3
221    z:   g2, d1
222   tr:   tr1
223  dcl:   dcl1
224   tr:   dcl1
226    d:   d1, tr1
227   tr:   th1
230    w:   g1, tr1, eh1
231   tr:   eh1
234   ix:   r3
236   tr:   tr1, th1
243   th:   th3, tr1, th2, tr1
248   tr:   ng4, tr1
252   ih:   ih3, ng1
255   tr:   ng3
259    l:   ng2, tr1, th1
263   tr:   tr4
268   iy:   y2, iy3
271   tr:   tr3
274  gcl:   gcl3
276   tr:   gcl1, tr1
280   el:   tr3, b1
284   tr:   b1, tr3
288   iy:   y4
290   tr:   dcl2
294  dcl:   dcl1, d3
295   tr:   tr1
```

```
296    d:   d1
298   tr:   ax1, tr1
300   ix:   s2
302   tr:   s1, tcl1
308    s:   s6
311   tr:   s1, tcl2
316  pcl:   tcl2, tr1, pcl2
317   tr:   tr1
318    p:   tr1
319   tr:   uh1
325   eh:   uh4, dh2
328   tr:   dh2, tr1
329    n:   th1
331   tr:   d1, tr1
332  tcl:   b1
335   tr:   b2, tr1
339    s:   tr1, p3
343   tr:   tr2, ax2
346   ix:   ng3
349   tr:   ng2, tr1
351   ng:   tr1, ng1
354   tr:   tr3
357  tcl:   tr2, tcl1
359   tr:   tcl2
365    t:   tcl1, tr3, dh2
368   tr:   tr1, r2
371    r:   r2, ae1
373   tr:   ae2
379   ae:   ae6
382   tr:   tr3
```

```
383   ng:   d1
384   tr:   d1
387  kcl:   d1, tr1, k1
388   tr:   k1
392    k:   k2, tr2
393   tr:   tr1
395    w:   b2
398   tr:   tr3
402   el:   tr2, ao1, aa1
405   tr:   aa3
418   ay:   aa5, tr1, aa1, r2, tr1, ah1, dh2
421   tr:   tr2, z1
424    z:   z2, p1
429   tr:   tr1, p2, g2
435   ix:   g6
440   tr:   g4, z1
446    z:   z3, g3
448   tr:   g2
462   h#:   g6, tr2, dh6
```

sx414

```
67    h#:   dh2, tr2, f1, d2, pcl5, tr2, dh2, tr2, d13,
             gcl14, tr2, eh1, dh5, tr1, ah8, tr1, aa4
69    tr:   aa2
79    ay:   aa10
82    tr:   aa2, r1
86     r:   r4
```

```
90    tr:   tr4
93    ix:   ax3
96    tr:   p3
102   sh:   p2, d4
106   tr:   d3, tr1
108    y:   y1, tr1
111   tr:   tr3
115   ah:   ah4
117   tr:   tr2
118   ng:   ng1
120   tr:   tr1, gcl1
123  gcl:   gcl1, tr1, d1
124   tr:   d1
125    g:   d1
127   tr:   tr1, gcl1
130    s:   gcl1, pcl2
132   tr:   tr2
134  tcl:   tcl2
135   tr:   tcl1
137    t:   tcl1, tr1
139   tr:   tr1, ng1
145   ix:   ng2, tr2, p1, g1
148   tr:   g1, pcl2
151    z:   pcl3
154   tr:   t3
158    q:   t1, p3
160   tr:   p1, tr1
167   iy:   iy7
169   tr:   iy1, tr1
172  tcl:   tcl3
```

```
173   tr:   t1
175    t:   tr1, d1
178   tr:   d1, tr2
184    f:   tr2, ah1, dh3
187   tr:   r2, tr1
190    r:   r3
191   tr:   tr1
194   eh:   ax2, tr1
197   tr:   tr1, p2
204   sh:   p1, sh6
207   tr:   sh2, tr1
211  kcl:   kcl4
213   tr:   tr1, k1
216    k:   k1, tr2
218   tr:   tr1, eh1
221   ih:   eh2, tr1
224   tr:   tr3
227  pcl:   pcl1, d2
229   tr:   d1, tr1
231    p:   d1, ax1
232   tr:   tr1
236   ix:   n3, tr1
240   tr:   tr1, z3
244    z:   z4
248   tr:   z1, dcl3
254    f:   dcl1, d1, tr3, r1
258   tr:   r3, tr1
262   ax:   uh4
264   tr:   uh1, tr1
270  bcl:   bcl4, tr2
```

```
272   tr:   tr1, b1
273    b:   b1
275   tr:   tr1, aa1
277    r:   aa2
278   tr:   aa1
283   eh:   aa1, eh4
285   tr:   tr2
288  kcl:   kcl1, g1, tr1
292   tr:   tr4
297    f:   p3, tr1, g1
300   tr:   tr3
305   ix:   tr1, ax4
310   tr:   b2, tr3
316    s:   tr2, g2, k2
319   tr:   k2, t1
321  tcl:   t2
322   tr:   t1
325    t:   t3
326   tr:   t1
342   h#:   p3, tr2, k5, d1, tr5
```

sx54

```
78    h#:   dcl8, gcl4, tr1, t3, dcl3, tr1, p3, tr1, s1,
             tr1, dh2, tr2, sh1, d4, gcl30, d1, tr3, dh2,
             tr3, g3, tr1
79    tr:   tr1
82    dh:   dh3
```

```
84    tr:   tr2
86    ih:   p2
88    tr:   tr1, d1
91     q:   d3
94    tr:   d1, y2
102   iy:   y3, iy5
106   tr:   iy3, tr1
110    s:   sh4
113   tr:   tcl2, tr1
114  tcl:   tcl1
115   tr:   t1
116    t:   t1
118   tr:   tr1, ng1
122   ix:   ng4
124   tr:   tr1, n1
126    n:   n2
129   tr:   tr1, ch1, d1
133  kcl:   d1, kcl3
135   tr:   kcl2
139    k:   kcl1, tr2, dh1
143   tr:   dh1, tr3
153   ow:   aa7, tr1, dh2
155   tr:   tr2
159    s:   dh2, tr2
161   tr:   tr1, tcl1
163  tcl:   tcl2
165   tr:   tcl1, tr1
167    t:   tr1, ax1
168   tr:   ax1
173   ih:   ax1, tr2, ih2
```

```
174   tr:   ih1
177    z:   ih1, g2
179   tr:   tr2
185   ey:   eh6
187   tr:   tr1, pcl1
190  pcl:   pcl1, tr2
192   tr:   tr1, p1
196    p:   p3, tr1
199   tr:   tr2, dh1
202    l:   dh1, tr1, dh1
204   tr:   dh1, tr1
208   ey:   eh2, tr2
211   tr:   tr1, g2
215    s:   s4
217   tr:   s1, tr1
221    f:   t3, tr1
223   tr:   tr1, ax1
228   ax:   ax3, gcl2
231   tr:   gcl2, tr1
236  pcl:   pcl4, d1
240   tr:   d3, tr1
243    p:   p2, tr1
245   tr:   f1, y1
249    y:   y1, tr2, uh1
252   tr:   uh3
259   uh:   uh3, tr1, uh3
262   tr:   tr3
266  pcl:   pcl2, d1, tr1
268   tr:   tr2
271    p:   p1, tr1, p1
```

```
272   tr:   p1
277    l:   p1, tr2, ah2
279   tr:   ah2
286   eh:   ah3, tr2, eh2
288   tr:   tcl2
292   zh:   tcl3, t1
294   tr:   t1, tr1
303  axr:   th1, r3, tr1, ax1, r3
306   tr:   tr1, dh2
310   ix:   ax4
312   tr:   tr1, n1
314    n:   n1, ng1
316   tr:   ng1, th1
319   ix:   th2, tr1
323   tr:   tr2, t1, tr1
325  kcl:   tr1, ah1
327   tr:   dh1, tr1
328    k:   k1
331   tr:   tr3
336    s:   tr1, s4
342   tr:   tr4, aa2
353   ay:   aa4, tr3, ax2, ih2
357   tr:   ih3, tr1
362  tcl:   dh1, tr1, tcl3
364   tr:   tr1, p1
365    m:   p1
367   tr:   tr1, ax1
370   ix:   ax3
371   tr:   tr1
373    n:   tr2
```

```
375   tr:   tr2
377  tcl:   tcl1, p1
378   tr:   tr1
385   h#:   tr2, pcl5
```

Appendix E.2 Confusion matrix

This section contains a confusion matrix of all the phoneme recognition results and transition recognition results. The rows depict the phonemes/transition as found in the original transcription files whilst the columns represent the recognized phonemes/transition. The percentages of correctly recognized phonemes and transition are printed in bold along the diagonal. They are calculated by dividing the number of correctly recognized phonemes/transitions by the number of all the occurrences of this phoneme/transition in the transition imposed “original” TIMIT transcription files. The empty rows and columns signify that a particular label is not found in the original or the recognized sequence respectively.

Orig/rec	pau	epi	h#	pcl	tcl	kcl	bcl	dcl	gcl	p	t	q	k	b	d	dx	g	jh	ch	s
pau				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.50		0.00		0.00	0.00
epi				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.33	0.00		0.00		0.00	0.00
h#				0.05	0.12	0.01	0.00	0.12	0.07	0.05	0.01		0.07	0.00	0.13		0.04		0.02	0.00
pcl				**0.45**	0.07	0.00	0.00	0.00	0.00	0.03	0.00		0.00	0.00	0.21		0.00		0.00	0.00
tcl				0.00	**0.49**	0.00	0.00	0.05	0.03	0.01	0.04		0.00	0.01	0.08		0.00		0.00	0.00
kcl				0.00	0.00	**0.43**	0.00	0.00	0.00	0.08	0.00		0.05	0.00	0.08		0.05		0.08	0.00
bcl				0.00	0.00	0.00	**0.86**	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
dcl				0.00	0.00	0.00	0.00	**0.35**	0.09	0.00	0.00		0.00	0.00	0.26		0.04		0.00	0.00
gcl				0.00	0.00	0.00	0.00	0.00	**0.45**	0.00	0.00		0.00	0.00	0.09		0.00		0.00	0.00
p				0.00	0.00	0.00	0.00	0.00	0.00	**0.50**	0.00		0.00	0.04	0.04		0.00		0.00	0.00
t				0.00	0.17	0.00	0.00	0.00	0.00	0.00	**0.27**		0.00	0.00	0.10		0.00		0.00	0.00
q				0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.04		0.00	0.00	0.17		0.00		0.00	0.00
k				0.00	0.00	0.03	0.00	0.07	0.00	0.00	0.00		**0.34**	0.03	0.10		0.00		0.00	0.00
b				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	**0.67**	0.00		0.00		0.00	0.00
d				0.00	0.00	0.00	0.00	0.00	0.12	0.13	0.00		0.00	0.00	**0.38**		0.00		0.00	0.00
dx				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.33		0.00		0.00	0.00
g				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.33	0.17		**0.50**		0.00	0.00
jh																				
ch				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		**0.00**	0.00
s				0.07	0.05	0.00	0.00	0.04	0.01	0.06	0.00		0.02	0.07	0.02		0.21		0.00	**0.15**
sh				0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.11		0.00	0.00	0.15		0.11		0.00	0.00
z				0.07	0.13	0.00	0.00	0.00	0.00	0.09	0.00		0.00	0.00	0.07		0.30		0.00	0.00
zh				0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.25		0.00	0.00	0.00		0.00		0.00	0.00
f				0.00	0.00	0.00	0.00	0.02	0.00	0.13	0.07		0.00	0.00	0.02		0.10		0.00	0.00
th				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
v				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
dh				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
m				0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00		0.00	0.13	0.00		0.09		0.00	0.00
em																				
n				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
en																				
nx				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
ng				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.25		0.00		0.00	0.00
eng																				
l				0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00		0.00	0.19	0.00		0.00		0.00	0.00
el				0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00		0.00	0.45	0.00		0.00		0.00	0.00
r				0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00		0.00	0.04	0.00		0.00		0.00	0.00
w				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.09	0.00		0.07		0.00	0.00
y				0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.25		0.00	0.00	0.00		0.00		0.00	0.00
hh																				
hv				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67		0.00	0.00	0.25		0.00		0.00	0.00
iy				0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00		0.00	0.00	0.13		0.00		0.00	0.00
ih				0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00		0.00	0.00	0.00		0.00		0.00	0.00
eh				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.04	0.00	0.00		0.00		0.00	0.00
ey				0.00	0.04	0.00	0.00	0.00	0.00	0.02	0.00		0.00	0.00	0.04		0.00		0.00	0.00
ae				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
aa				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
aw				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.09	0.00		0.00		0.00	0.00
ay				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
ah				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
ao				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.03	0.00		0.00		0.00	0.00
oy				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
ow				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.13	0.00		0.00		0.00	0.00
uh				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
uw				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.05	0.00		0.00		0.00	0.00
ux				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
er				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
ax				0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00		0.00	0.05	0.00		0.00		0.00	0.00
ix				0.00	0.04	0.00	0.00	0.00	0.06	0.01	0.00		0.00	0.00	0.00		0.11		0.00	0.03
axr				0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00		0.00	0.00
ax-h																				
tr				0.00	0.02	0.01	0.00	0.01	0.01	0.04	0.01		0.01	0.03	0.04		0.03		0.00	0.00

Orig/rec	sh	z	zh	f	th	v	dh	m	em	n	en	nx	ng	eng	l	el	r	w	y	hh
pau	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
epi	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
h#	0.06	0.00		0.00	0.00	0.00	0.05	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
pcl	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
tcl	0.01	0.00		0.00	0.01	0.00	0.04	0.00		0.00			0.00		0.00		0.00	0.01	0.06	
kcl	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.02		0.00	0.00	0.00	
bcl	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
dcl	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.13	0.00	
gcl	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.18	0.00	
p	0.00	0.00		0.00	0.00	0.00	0.04	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
t	0.00	0.00		0.00	0.00	0.00	0.10	0.00		0.00			0.03		0.00		0.00	0.00	0.07	
q	0.00	0.00		0.00	0.00	0.00	0.09	0.00		0.00			0.00		0.00		0.04	0.00	0.00	
k	0.10	0.00		0.00	0.00	0.00	0.07	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
b	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
d	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
dx	0.00	0.00		0.00	0.33	0.00	0.33	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
g	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
jh																				
ch	1.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
s	0.06	0.00		0.00	0.00	0.02	0.02	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
sh	**0.23**	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
z	0.00	**0.20**		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
zh	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
f	0.00	0.00		**0.06**	0.00	0.13	0.13	0.00		0.00			0.00		0.00		0.02	0.00	0.00	
th	0.00	0.00		0.00	**0.56**	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
v	0.00	0.00		0.00	0.00	**0.67**	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
dh	0.00	0.00		0.00	0.00	0.00	**0.58**	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
m	0.00	0.00		0.00	0.00	0.00	0.00	**0.13**		0.00			0.17		0.00		0.09	0.00	0.00	
em																				
n	0.06	0.00		0.00	0.11	0.00	0.11	0.00		**0.28**			0.11		0.00		0.00	0.00	0.00	
en																				
nx	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
ng	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			**0.50**		0.00		0.00	0.00	0.00	
eng																				
l	0.00	0.00		0.00	0.02	0.00	0.10	0.00		0.00			0.03		**0.19**		0.00	0.00	0.00	
el	0.00	0.00		0.00	0.00	0.00	0.04	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
r	0.00	0.00		0.00	0.00	0.00	0.01	0.02		0.00			0.00		0.00		**0.48**	0.00	0.00	
w	0.00	0.00		0.00	0.02	0.00	0.02	0.00		0.00			0.00		0.00		0.00	**0.47**	0.00	
y	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	**0.25**	
hh																				
hv	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
iy	0.00	0.00		0.00	0.05	0.00	0.00	0.02		0.00			0.00		0.00		0.00	0.00	0.22	
ih	0.00	0.00		0.00	0.05	0.00	0.00	0.00		0.00			0.05		0.00		0.00	0.00	0.00	
eh	0.00	0.00		0.00	0.00	0.00	0.08	0.00		0.00			0.00		0.00		0.00	0.00	0.04	
ey	0.09	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.15		0.00		0.00	0.00	0.00	
ae	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.09	0.00	0.00	
aa	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
aw	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.13	0.00	0.00	
ay	0.00	0.00		0.00	0.00	0.00	0.03	0.00		0.00			0.00		0.00		0.07	0.00	0.00	
ah	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
ao	0.00	0.00		0.00	0.00	0.00	0.01	0.00		0.00			0.07		0.00		0.00	0.00	0.00	
oy	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.09			0.00		0.00		0.00	0.00	0.00	
ow	0.00	0.00		0.00	0.00	0.00	0.13	0.00		0.00			0.00		0.00		0.18	0.00	0.00	
uh	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
uw	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.16	0.00	
ux	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
er	0.00	0.00		0.00	0.00	0.00	0.17	0.00		0.00			0.00		0.00		0.39	0.00	0.00	
ax	0.00	0.00		0.00	0.00	0.00	0.00	0.00		0.00			0.00		0.00		0.00	0.00	0.00	
ix	0.00	0.00		0.00	0.06	0.00	0.01	0.00		0.04			0.13		0.00		0.04	0.00	0.00	
axr	0.00	0.00		0.00	0.04	0.00	0.00	0.00		0.00			0.00		0.00		0.25	0.00	0.00	
ax-h																				
tr	0.01	0.01		0.00	0.01	0.00	0.03	0.00		0.00			0.02		0.00		0.02	0.01	0.01	

Orig/rec	hv	iy	ih	eh	ey	ae	aa	aw	ay	ah	ao	oy	ow	uh	uw	ux	er	ax	ix	axr	ax-h	tr
pau	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.50
epi	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.17			0.00				0.50
h#	0.01	0.00	0.01	0.01		0.00	0.01			0.01	0.00			0.00	0.00			0.00				0.13
pcl	0.00	0.00	0.00	0.00		0.00	0.00			0.03	0.00			0.00	0.00			0.00				0.21
tcl	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.14
kcl	0.00	0.00	0.00	0.00		0.00	0.00			0.02	0.00			0.00	0.00			0.02				0.16
bcl	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.14
dcl	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.13
gcl	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.27
p	0.00	0.00	0.04	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.13				0.21
t	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.03				0.23
q	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.13			0.00	0.09			0.17				0.13
k	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.24
b	0.00	0.00	0.00	0.00		0.00	0.00			0.33	0.00			0.00	0.00			0.00				0.00
d	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.37
dx	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.00
g	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.00
jh																						
ch	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.00
s	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.03				0.17
sh	0.00	0.00	0.00	0.00		0.00	0.00			0.11	0.00			0.00	0.00			0.00				0.08
z	0.00	0.00	0.01	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.13
zh	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.00
f	0.00	0.00	0.00	0.00		0.00	0.00			0.02	0.00			0.00	0.00			0.00				0.30
th	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.44
v	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.33
dh	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.25				0.17
m	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.13			0.09				0.13
em																						
n	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.22				0.11
en																						
nx	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			1.00				0.00
ng	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.25
eng																						
l	0.00	0.00	0.00	0.00		0.02	0.00			0.06	0.06			0.00	0.00			0.06				0.19
el	0.00	0.00	0.00	0.00		0.00	0.05			0.00	0.05			0.00	0.00			0.00				0.36
r	0.00	0.00	0.00	0.00		0.02	0.05			0.04	0.00			0.08	0.00			0.03				0.20
w	0.00	0.00	0.00	0.02		0.00	0.00			0.00	0.00			0.02	0.00			0.02				0.27
y	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.06	0.00			0.00				0.19
hh																						
hv	**0.00**	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.00	0.00			0.00				0.08
iy	0.00	**0.35**	0.05	0.00		0.00	0.00			0.00	0.00			0.06	0.00			0.00				0.02
ih	0.00	0.00	**0.60**	0.05		0.00	0.00			0.00	0.08			0.00	0.00			0.03				0.08
eh	0.00	0.00	0.00	**0.25**		0.00	0.04			0.13	0.00			0.17	0.00			0.08				0.17
ey	0.00	0.00	0.27	0.25		0.00	0.00			0.00	0.00			0.00	0.00			0.07				0.07
ae	0.00	0.00	0.00	0.00		**0.77**	0.04			0.04	0.00			0.00	0.00			0.04				0.02
aa	0.00	0.00	0.00	0.00		0.00	**0.89**			0.00	0.00			0.00	0.00			0.00				0.11
aw	0.00	0.00	0.00	0.00		0.00	0.48			0.00	0.00			0.00	0.00			0.13				0.17
ay	0.00	0.00	0.13	0.00		0.00	0.36			0.02	0.00			0.00	0.00			0.25				0.14
ah	0.00	0.00	0.00	0.00		0.00	0.14			**0.73**	0.00			0.00	0.00			0.09				0.04
ao	0.00	0.00	0.00	0.00		0.00	0.00			0.00	**0.82**			0.06	0.00			0.00				0.00
oy	0.00	0.00	0.00	0.00		0.00	0.00			0.13	0.00			0.57	0.00			0.00				0.21
ow	0.00	0.00	0.00	0.05		0.08	0.18			0.00	0.03			0.08	0.00			0.00				0.13
uh	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			**0.92**	0.00			0.00				0.08
uw	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.16			0.00	**0.58**			0.00				0.05
ux	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.86	0.00			0.00				0.14
er	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.11	0.00			0.00				0.33
ax	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.27	0.00			**0.50**				0.09
ix	0.00	0.00	0.00	0.00		0.00	0.00			0.00	0.00			0.07	0.00			0.32				0.07
axr	0.00	0.00	0.00	0.00		0.00	0.00			0.17	0.00			0.00	0.00			0.46				0.08
ax-h																						
tr	0.00	0.01	0.02	0.01		0.02	0.02			0.02	0.02			0.03	0.01			0.04				**0.47**

Appendix F Evaluation results obtained by HResults for new speech data

The text concerning this section can be found in Chapter 6, section 6.2. The following table contains the evaluation results for 8 different utterances of a new speaker (ID: dr1/mdac0).

```
---------------------- Sentence Scores ----------------------
====================== HTK Results Analysis ======================

------------------------ File Results --------------------------
  si1261.rec:   58.02( 40.74)  [H=  47, D=  3, S= 31, I= 14, N= 81]
  si1837.rec:   56.36( 12.73)  [H=  31, D=  0, S= 24, I= 24, N= 55]
  si631.rec:    67.52(  8.55)  [H=  79, D=  1, S= 37, I= 69, N=117]
  sx181.rec:    64.84( 13.19)  [H=  59, D=  1, S= 31, I= 47, N= 91]
  sx271.rec:    60.32( 17.46)  [H=  38, D=  1, S= 24, I= 27, N= 63]
  sx361.rec:    62.69( 26.87)  [H=  42, D=  0, S= 25, I= 24, N= 67]
  sx451.rec:    62.89( 20.62)  [H=  61, D=  1, S= 35, I= 41, N= 97]
  sx91.rec:     57.75( 35.21)  [H=  41, D=  0, S= 30, I= 16, N= 71]
---------------------- Overall Results ----------------------
               SENT: %Correct=0.00 [H=0, S=8, N=8]
  WORD: %Corr=61.99, Acc=21.18 [H=398, D=7, S=237, I=262, N=642]
================================================================
```

Table F.1 Evaluation results for 8 utterances

HResults is also used to produce a confusion matrix for the experiments on the 8 different utterances of a new speaker (ID: dr1/mdac0). The rows depict the phonemes/transition as found in the original transcription files whilst the columns represent the recognized phonemes/transition. According to the HTK's manual [Young et al. 2002], "%c indicates the number of correct instances divided by the total number of instances in the row. %e is the number of incorrect instances in the row divided by the total number of instances". The whole confusion matrix is listed in the following two pages.

---------- Confusion Matrix ----------

	aa	ae	ah	ax	b	bcl	d	dcl	dh	eh	epi	g	gcl	ih	iy	k	kcl	l	m	n	ng	p	pau	pcl	r	s	sh	t	tcl	th	tr	uh	v	w	y	z	Del [%c / %e]
aa	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ae	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ah	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0 [60.0/0.3]
ao	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.2]
aw	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.3]
ax	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [75.0/0.2]
ax-h	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0 [0.0/0.8]
axr	0	0	0	3	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/1.1]
ay	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.6]
b	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0 [50.0/0.5]
bcl	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [33.3/0.3]
ch	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.3]
d	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0 [77.8/0.3]
dcl	0	0	0	0	0	0	3	5	0	0	1	1	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [35.7/1.4]
dh	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [33.3/0.6]
dx	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.3]
eh	0	0	0	0	0	0	0	0	1	4	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [57.1/0.5]
el	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.3]
em	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.3]
en	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.5]
epi	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [50.0/0.2]
er	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.3]
ey	0	0	0	1	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/1.1]
f	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0 [0.0/0.9]
g	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0 [0.0/0.5]
gcl	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0 [20.0/0.6]
hh	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.5]
hv	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1 [0.0/0.2]
ih	0	0	0	0	0	0	1	0	0	0	0	0	0	10	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [76.9/0.5]
ix	0	0	0	1	1	0	3	0	2	0	0	0	0	1	0	0	0	1	1	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0 [0.0/2.5]
iy	0	0	0	0	0	0	2	0	0	0	0	1	0	0	3	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0 [30.0/1.1]

jh	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.5]
k	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	4	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [50.0/0.6]
kcl	0	0	0	0	0	0	3	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0 [10.0/1.4]
l	0	0	0	0	3	0	0	0	1	0	1	0	0	0	0	0	0	4	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0 [36.4/1.1]
m	0	0	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0 [45.5/0.9]
n	1	0	0	1	4	0	2	0	0	0	0	0	0	0	1	0	0	0	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1 [15.4/1.7]
ow	0	0	0	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0 [0.0/1.2]
p	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0 [44.4/0.8]
pau	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0 [20.0/0.6]
pcl	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0 [14.3/0.9]
q	0	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1 [0.0/0.9]
r	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0 [22.2/1.1]
s	0	0	0	0	0	0	2	0	1	0	0	1	0	0	0	0	0	0	4	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0 [9.1/1.6]
sh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.2]
t	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	0	0	0 [60.0/0.3]
tcl	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	2	0	0	0	0	0	0	0 [11.1/1.2]
th	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
tr	0	0	0	0	1	0	2	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	314	1	0	0	0	0	4 [97.8/1.1]
uh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ux	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.2]
v	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0 [0.0/0.8]
w	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0 [0.0/0.6]
y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
z	0	0	0	2	1	0	4	0	0	0	0	0	1	1	0	0	0	0	1	0	0	2	0	0	1	0	0	0	0	1	0	0	0	0	0	2	0 [12.5/2.2]
Ins	3	2	2	8	11	0	36	0	5	7	10	10	11	7	2	2	0	1	15	1	5	16	2	0	10	1	5	2	2	15	56	5	1	1	8	0	

BIBLIOGRAPHY

Atkinson, M., Kilby, D. and Roca, I. 1988. Foundations of General Linguistics. Unwin Hyman Ltd, London.

Ball, M. J. 2002. Teaching Vowels in Practical Phonetics: The Auditory or Articulatory Route? http://www.phon.ucl.ac.uk/home/johnm/ball.htm. University of Ulster.

Beckman, M., Edwards, J., and Fletcher, J. 1992. Prosodic Structure and Tempo in a Sonority Model of Articulatory Dynamics. Chapter 3: Papers in Laboratory Phonology II – Gesture, Segment, Prosody. Edited by G. J. Docherty and D. R. Ladd. Cambridge University Press.

Bell, A. and Hooper, J. B. 1980. Issues and Evidence in Syllabic Phonology. Syllables and Segments. Edited by A. Bell and J. B. Hooper, Amsterdam: North Holland, pp. 1-22.

Chang, S. 2002. A Syllable, Articulatory-Feature and Stress-Accent Model of Speech Recognition. Ph.D. thesis. Computer Science Division, Department of EECS, University of California, Berkeley.

Chomsky, N. and Halle, M. 1968. The Sound Pattern of English. Harper, New York.

De Mori, R. and Brugnara, F. 1995. HMM Methods in Speech Recognition. Survey of the State of the Art in Human Language Technology. Edited by R.A. Cole. Online Version: CSLU OGI, pp. 24-35.

Denes, P. B. and Pinson, E. N. 1993. The Speech Chain: The Physics and Biology of Spoken Language. W.H. Freeman and Company.

Durand, J. 1990. Generative and Non-Linear Phonology. Longman.

Edmondson, W., Iles, J. and Iskra, D. 1996. Pseudo-Articulatory Representations in Speech Synthesis and Recognition. Proceedings of International Conference on Spoken Language Processing (ICSLP 1996), 4:2215-2218.

Edmondson, W., Iskra, D. and Kienzle, P. 1999. Pseudo-Articulatory Representations: Promise, Progress and Problems. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1999), 3:1435-1438.

Edmondson, W. and Zhang, L. 2001. Pseudo-Articulatory Representations and the Recognition of Syllable Patterns in Speech. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2001), 1:595-598.

Edmondson, W. and Zhang, L. 2002. The Use of Syllable Structure for Speech Recognition. University of Birmingham. School of Computer Science. Technical Report CSRP-02-7.

ESPS Manual. 1996. Version 5.1. Entropic Research Laboratory.

Ewen, C. and van der Hulst, H. 2001. The Phonological Structure of Words. Cambridge University Press.

Excel Manual. 1997. Microsoft Corporation.

Ferguson, G. and Takane, Y. 1989. Statistical Analysis in Psychology and Education. McGraw Hill Book Company.

Finke, M., Geutner, P., Hild, H., Kemp, T., Ries, K., and Westphal, M. 1997. The Karlsruhe Verbmobil Speech Recognition Engine. Proceedings of International Conference on Acoustics, Speech and Signal Processing.

Fromkin, V. and Rodman, R. 1998. An Introduction to Language. Harcourt Brace College Publishers.

Fujimura, O. 1992. Commentary on Prosodic Structure and Tempo in a Sonority Model of Articulatory Dynamics. Chapter 3: Papers in Laboratory Phonology II – Gesture, Segment, Prosody. Edited by G. J. Docherty and D. R. Ladd. Cambridge University Press.

Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchoff, K., Ordowski, M. and Wheatley, B. 1997. Syllable – A Promising Recognition Unit for LVCSR. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, pp. 207-214, Santa Barbara, California, USA.

Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G. and Picone, J. 2001. Syllable-Based Large Vocabulary Continuous Speech Recognition. IEEE Transactions on Speech and Audio Processing.

Garolfo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D. and Dahlgren, N. 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology, NISTIR 4930.

Giegerich, H. 1992. English Phonology. Cambridge University Press.

Greenberg. S. 1997. The Switchboard Transcription Project. 1996 LVCSR Summer Research Workshop, Research Notes 24, CLSP, John Hopkins University.

Greenberg, S. 1999. Speaking in Shorthand – A Syllable-Centric Perspective for Understanding Pronunciation Variation. Speech Communication, 29:159-176.

Goldsmith, J. A. 1990. Autosegmental and Metrical Phonology. Basil Blackwell.

Hamaker, J., Ganapathiraju, A., and Picone, J. 1997. Syllable-Based Speech Recognition. Technical Report. Prepared for Speech Research Group, Personal System Laboratory. Texas Instruments, lnc. Texas. Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University.

Hatzis, A. and Green, P. D. 2001. A Two Dimensional Kinematic Mapping between Speech Acoustics and Vocal Tract Configurations. Workshop on Innovation in Speech Processing.

Hawkins, S. 1992. An Introduction to Task Dynamics. Chapter 1: Papers in Laboratory Phonology II – Gesture, Segment, Prosody. Edited by G. J. Docherty and D. R. Ladd. Cambridge University Press.

Holmes, J. 1988. Speech Synthesis and Recognition. Van Nostrand Reinhold Co. Ltd.

Iles, J. 1995. Text-to-Speech Conversion Using Feature-Based Formant Synthesis in a Non-Linear Framework. Ph.D. thesis. School of Computer Science, University of Birmingham. UK.

Iskra, D. 2000. Feature-Based Approach to Speech Recognition. Ph.D. thesis. School of Computer Science, University of Birmingham. UK.

Jakobson, R., Fant, C. and Halle, M. 1952. Preliminaries to Speech Analysis. Technical Report No. 13, Acoustics Laboratory, Cambridge MA; Published under the Same Title in 1963 by MIT Press, Cambridge MA.

Jones, R., Downey, S., and Mason, J. 1997. Continuous Speech Recognition Using Syllables. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1997), pp. 1171-1174, Rhodes, Greece.

Junqua, J. 1991. A Two-Pass Hybrid System Using a Low Dimensional Auditory Model for Speaker-Independent Isolated-Word Recognition. Speech Communication. 10(1): 33-44.

Kaye, J. 1989. Phonology: A Cognitive View. New Jersey: Lawrence Earlbaum Associates.

King, S., Taylor, P., Frankel, J. and Richmond, K. 2000. Speech Recognition via Phonetically-Featured Syllables. In *PHONUS*, 5: 15-34, Institute of Phonetics, University of the Saarland.

Ladefoged, P. 1993. A Course in Phonetics. Harcourt Brace, Fort Worth, TX, 3rd Edition.

Ladefoged, P. 2001. Vowels and Consonants: An Introduction to the Sounds of Languages. Blackwell Publishing.

Laver, J. 1994. Principles of Phonetics. Cambridge University Press.

Lee, K. 1989a. Automatic Speech Recognition: the Development of the SPHINX System. Kluwer Academic Publishers, Boston.

Lee, K. 1989b. Hidden Markov Models: Past, Present, and Future. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1989), pp. 148-154.

Lee, L. 1997. Voice Dictation of Mandarin Chinese. IEEE Signal Processing Magazine, (97): 63-101.

Leung, K. and Siu, M. 2002. Speech Recognition Using Combined Acoustic and Articulatory Information with Retraining of Acoustic Model Parameters. Proceedings of International Conference on Spoken Language Processing (ICSLP 2002). Denver, USA, pp. 2117-2120.

Levelt, W., Roelofs, A. and Meyer, A. 1999. A Theory of Lexical Access in Speech Production. Behav. and Brain Sci., 22:1-38.

Lippmann, R. 1989. Review of Neural Networks for Speech Recognition. Neural Computation. Vol. 1, pp. 1-38.

MacKay, I. 1987. Phonetics, the Science of Speech Production. Boston: Little, Brown.

Makhoul, J. and Schwartz, R. 1994. State of the Art in Continuous Speech Recognition. Voice Communication Between Humans and Machines. Edited by D. Roe and J. Wilpon. National Academy Press, pp. 165-198.

Mariani, J. 1989. Recent Advances in Speech Processing. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 429-440.

Markov, K., Dang, J., Iizuka, Y. and Nakamura, S. 2003. Hybrid HMM/BN ASR System Integrating Spectrum and Articulatory Features. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, pp. 965-968.

Matlab Manual. 2001. Microsoft Corporation.

Metze, F. and Waibel, A. 2002. A Flexible Stream Architecture for ASR Using Articulatory Features. Proceedings of International Conference on Spoken Language Processing (ICSLP 2002). Denver, USA, pp. 2133-2136.

McCullough, W.S. and Pitts, W. H. 1943. A Logical Calculus of Ideas Immanent in Nervous Activity. Bull Math Biophysics, 5:115-133.

Nakagawa, S., Hanai, K., Yamamoto, K. and Minematsu, N. 1999. Comparison of Syllable-Based HMMs and Triphone-Based HMMs in Japanese Speech Recognition. Proceedings of the International Workshop on Automatic Speech Recognition and Understanding, pp. 197-200, Keystone, CO.

Ni Chasaide, N. and Gobl, C. 1997. Voice Source Variation. The Handbook of Phonetic Sciences. Edited by W.J. Hardcastle and J. Laver. Blackwell Publishers.

Noel, M. 1997. "Alphadigits," http://cslu.cse.ogi.edu/corpora/alphadigit. Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA.

O'Conner, J. 1974. Phonetics. Pelican Books.

Ohala, J. 1992. The Segment: Primitive or Derived? Chapter 7: Papers in Laboratory Phonology II – Gesture, Segment, Prosody. Edited by G. J. Docherty and D. R. Ladd. Cambridge University Press.

Okadome, T. and Honda, M. 2001. Generation of Articulatory Movements by Using a Kinematic Triphone Model. JASA, pp. 453-463.

Ostendorf, M. et al. 1997. Modeling Systematic Variations in Pronunciations via a Language-Dependent Hidden Speaking Mode. 1996 LVCSR Summer Research Workshop, Research Notes 24, CLSP, John Hopkins University.

Rabiner, L. and Juang, B. 1986. An Introduction to Hidden Markov Models. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4-16.

Rabiner, L. and Juang, B. 1993. Fundamentals of Speech Recognition. Prentice-Hall International.

Ramsay, G. and Deng, L. 1995a. Articulatory Synthesis Using a Stochastic Target Model of Speech Production. Proceedings of International Conference on Phonetic Sciences. 2:338-341.

Ramsay, G. and Deng, L. 1995b. Maximum-Likehood Estimation for Articulatory Speech Recognition Using a Stochastic Target Model. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 1995), pp. 1401-1404.

Robinson, T., Hochberg, M. and Renals, S. 1996. The Use of Recurrent Neural Networks in Continuous Speech Recognition. Automatic Speech and Speaker

Recognition. Edited by C.-H. Lee, F.K. Song and K.K. Paliwal. Klewer Academic Publishers, Norwell, MA.

Sakoe, H. and Chiba, S. 1978. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol. ASSP-26, no.1, pp. 43-49.

Selkirk, E. 1984. On the Major Class Features and Syllable Theory. Language Sound Structure. Edited by M. Aronoff and R. T. Oehrle. The MIT Press, Cambridge, Massachusetts, London, pp. 107-136.

Stewart, D., Ming, J., Hanna, P. and Smith, F. 2002. A State-Tying Approach to Building Syllable HMMs. Proceedings of International Conference on Spoken Language Processing (ICSLP 2002), 4:2649-2652.

Stuker, S., Metze, F., Schultz, T. and Waibel, A. 2003. Integrating Multilingual Articulatory Features into Speech Recognition. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, pp. 1033-1036.

Wu, S. 1998. Incorporating Information from Syllable-Length Time Scale into Automatic Speech Recognition. Ph.D. thesis. Department of EECS, University of California, Berkeley.

Young, S. 1995. Large Vocabulary Continuous Speech Recognition: a Review. Proceedings of IEEE Automatic Speech Recognition Workshop, pp. 3-28.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. 2002. The HTK Book (for HTK Version 3.2).

Zhang, L. and Edmondson, W.H. 2001. Pseudo-Articulatory Representations in Speech Recognition. Proceedings of PREP 2001. Keele University, UK.

Zhang, L. and Edmondson, W.H. 2002a. Speech Recognition Using Syllable Patterns. Proceedings of International Conference on Spoken Language Processing (ICSLP 2002). 2:1237-1240. Denver, USA.

Zhang, L. and Edmondson, W.H. 2002b. Pseudo-Articulatory Basis in Speech Recognition. Proceedings of WSEAS International Conference on Acoustics, Music, Speech and Language Processing (former Acoustics and Music: Theory and Applications). Puerto De La Cruz, Tenerife, Canary Islands, Spain.

Zhang, L. and Edmondson, W.H. 2002c. Pseudo-Articulatory Representations and the Use of Syllable Structure for Speech Recognition. Advances in Communications and Software Technologies. Edited by Nikos E. Mastorakis and Vitaliy V. Kluev. WSEAS Press, pp. 259-264.

Zhang, L. and Edmondson, W.H. 2002d. Feature-Based Approach to Speech Recognition. Advances in System Engineering, Signal Processing and Communications. Edited by Diego Andina de la Fuente and Nikos E. Mastorakis. WSEAS Press, pp. 132-136.

Zhang, L. and Edmondson, W.H. 2003. Speech Recognition Based on Syllable Recovery. Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland.

Zhang, L. 2003. A Syllable-Based Pseudo-Articulatory Approach To Speech Recognition. Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China.

Zue, V. W. and Seneff, S. 1988. Transcription and Alignment of the TIMIT Database. Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language.

INDEX

A

ACF 14, 15
acoustic correlates ..... 34, 40, 41, 63, 64, Table 3.1
acoustic feature 13, 14
AF 14, 15, 18, 24, 25
allophone 5, 6, 28, 56, 71, 90
alphadigit 22, 23, 171
ambisyllabics 23
articulatory configuration .. 31, 33, 34, 70
articulatory event 70, 74, 105
articulatory feature . 5, 12, 13, 14, 15, 16, 17, 19, 40, 104
articulatory model 17, 81
articulatory speech recognition 12
articulatory targets 70
articulatory-acoustic mapping 33, 34, 54, 55, 60, 65, 99
artificial neural network 14, 26, 47
atemporal 29
Atkinson et al. 43, 47

B

Ball 104, 105
Bayesian network 12
beads-on-a-string 20, 24
Beckman et al. 10, 70
Bell and Hooper 66
BN 12, 13
brute search 53, 54, 55, 56, 57, 58, 98

C

cepstral coefficients 30, 33, 40, 49, 52, 56, 57, 60
Chang 24
coarticulation 3, 11, 31, 43, 98
co-articulatory 2
coda .....8, 9, 67, 68, 74, Figure 2.1
coil 12
confusion matrix... 85, 90, 91, 93, 95, 97, 131, 158, 162, 163, Table 6.1
consonant model..... 4, 33, 51, 52, 53, 58, 72, 77, 105
coronal 16, 105
cross-lingual 18, 19

D

Denes and Pinson 1, 10, 37
diphones 30
discrete 5, 11
discriminative model combination 18
distinctive features..... 6, 29, 35, 43, 51
Durand 105
dynamic programming 60, 85, 86, 87, 93 100

E

Edmondson and Zhang..... 3, 36, 68, 103
Edmondson et al. 29, 35, 43
electromagnetic midsagittal articulographic 12
esps *xwaves* 49
Ewen and van der Hulst 8
excel 45

F

Finke et al. 16
formants 30, 37, 38, 39, 40, 42, 44, Table 3.1
Fourier 26, 40
Fromkin and Rodman..... 43, 52

G

Ganapathiraju et al. 3, 22, 23
Garolfo et al..... 27, 45
gemination 29
Giegerich 8, 9, 66
Goldsmith 25
Greenberg 21, 22, 25

H

Hamaker et al. 3, 21, 23
Hatzis and Green. 47
Hawkins..... 70
hidden Markov modeling 2
hidden Markov models..... 10, 30
Holmes 85, 86

hybrid HMM 12, 14
hyperarticulated 17

I

idealized trajectories 54, 58, 60, 62, 66
Iles .... 2, 3, 30, 43, 46, 47, 50, 51, 56, 65, Table 3.3
impedance ..... 10, Figure 5.7
imposed transitions 100, 101, 131, Figure 7.1
Iskra .. 2, 3, 30, 33, 36, 40, 43, 44, 45, 46, 47, 48, 49, 52, 53, 55, 56, 60, 61, 62, 67, 77, 87, 94, 95, 98, Figure 6.2

J

Janus 16
Jones et al. 25

K

Kaye 66
King et al. 11

L

Ladefoged ... 7, 38, 39, 40, 41, 44, 46, 49, 63, 68, 104, Table 3.1
language models 11, 18
Laver 6, 43
lax 44, 52
Lee 11, 20, 104
Leung and Siu 14
Levelt et al. 7
Lippmann 26
LVCSR 20, 21

M

MacKay 34, 43, 46, 47, 48, 63
Mariani 26, 85
Markov et al. 12
McCullough and Pitts 26
mel-frequency cepstral coefficients 12
Metze and Waibel 15, 17, 18
MFCC 12, 13
midsagittal 12
minimal pair 6
monolingual 17, 18, 19
morpho-phonemic 2
multi-layer perceptrons 14
multilingual 17, 18, 19
multilingual mixed 18
multiple regression analysis 33, 45, 50
multithreaded 3, 83, 103

N

Nakagawa et al. 20
Ni Chasaide and Gobl 44, 63, 64
Noel 22
non-segment 36
nucleus ..... 8, 9, 67, 68, 70, 74, Figure 2.1

O

obstruent 29
obstruents 70
OGI alphadigit corpus 22
Okadome and Honda 12
Ostendorf et al. 21

P

palatal 16, 105
periodicity 70
phone ..... 5, Figure 5.7
phoneme 5
phoneme reference 33, 53, 56, 62, 94, 95
phonetic features . 5, 6, 31, 34, 43, 63, 98
phonetically-compact .. 27, 109, 110, 111
phonetically-diverse 27, 109, 110
phonetics 5
phonology 5
phonotactic 7, 8, 67
pitch 8
probability 18, 61, 86
pseudo-articulatory features ... 31, 33, 34, 37, 65, 98
pseudo-articulatory representations 2, 30, 35, 80

R

Rabiner and Juang 2, 26
read Broadcast News task 16
readBN 16
regression equation 33, 49, 50

resynthesis 54, 60, 64, 66, 76, 78, 99
rhyme ..... 8, 67, Figure 2.1
Robinson et al. 26

S

Sakoe and Chiba 86
Selkirk 8, 68
sonorant 8, 9, 29, 70
sonority ..... 8, 9, Figure 2.2, Figure 5.7
sonority scale 8, 9
sonority theory 8, 9, 10
sonority waves 9, 67, 74, 105
speaker-independent 22, 26
speaking rate 25
speech recognition techniques 5, 10
Stewart et al. 11, 23
stream architecture 16, 17
stress-accent 24, 25
Stuker et al. 17
SWB 17, 21, 22, 23
Switchboard 17, 21
syllabary 21
syllabic targets 85, 86, 87, 88, 95, 96, 100
syllable model 68, 69, 70, 71, 74, 98, 99, 105, Figurer 5.1, Figure 5.2
syllable-based HMM systems 20
syllable-based HMMs 11
syllables ..... 7, Figure 5.7

T

time-aligned phonetic transcription 28
TIMIT ..... 26, 106, Figure 7.1
triphone 20, 21, 22, 30

V

viterbi algorithm 85
vowel model 4, 33, 49, 52, 53, 58, 72, 105

W

Wu ..... 25, Figure 5.7

Y

Young 93, 162

Z

Zhang and Edmondson 3, 103
Zue and Seneff 28, 97