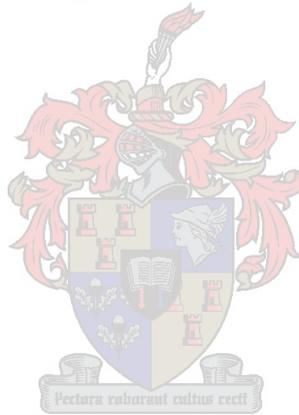


# Automatic Syllabification of Untranscribed Speech

Pieter W Nel

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in the Department of Electrical and Electronic Engineering at the University of Stellenbosch.



Advisor: Prof. J.A. du Preez

April 2005

# Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature:

*November 2004*

# Abstract

The syllable has been proposed as a unit of automatic speech recognition due to its strong links with human speech production and perception. Recently, it has been proved that incorporating information from syllable-length time-scales into automatic speech recognition improves results in large vocabulary recognition tasks. It was also shown to aid in various language recognition tasks and in foreign accent identification. Therefore, the ability to automatically segment speech into syllables is an important research tool. Where most previous studies employed knowledge-based methods, this study presents a purely statistical method for the automatic syllabification of speech.

We introduce the concept of hierarchical hidden Markov model structures and show how these can be used to implement a purely acoustical syllable segmenter based, on general sonority theory, combined with some of the phonotactic constraints found in the English language.

The accurate reporting of syllabification results is a problem in the existing literature. We present a well-defined dynamic time warping (DTW) distance measure used for reporting syllabification results.

We achieve a token error rate of 20.3% with a 42ms average boundary error on a relatively large set of data. This compares well with previous knowledge-based and statistically-based methods.

# Opsomming

Die syllabe is voorheen voorgestel as 'n basiese eenheid vir automatiese spraakherkenning weens die sterk verwantskap wat dit het met spraak produksie en persepsie. Onlangs is dit bewys dat die gebruik van informasie van syllabe-lengte tydskaal die resultate verbeter in groot woordeskat herkennings take. Dit is ook bewys dat die gebruik van syllabes automatiese taalherkenning en vreemdetaal aksent herkenning vergemaklik. Dit is daarom belangrik om vir navorsingsdoeleindes syllabes automaties te kan segmenteer. Vorige studies het kennisgebaseerde metodes gebruik om hierdie segmentasie te bewerkstellig. Hierdie studie gebruik 'n suiwer statistiese metode vir die automatiese syllabifikasie van spraak.

Ons gebruik die konsep van hierargiese verskuilde Markov model strukture en wys hoe dit gebruik kan word om 'n suiwer akoestiese syllabe segmenteerder te implementeer. Die model word gebou deur dit te baseer op die teorie van sonoriteit asook die fonotaktiese beperkinge teenwoordig in die Engelse taal.

Die akkurate voorstelling van syllabifikasie resultate is problematies in die bestaande literatuur. Ons definieer volledig 'n DTW (Dynamic Time Warping) afstandsfunksie waarmee ons ons syllabifikasie resultate weergee.

Ons behaal 'n TER (Token Error Rate) van 20.3% met 'n 42ms gemiddelde grensfout op 'n relatiewe groot stel data. Dit vergelyk goed met vorige kennis-gebaseerde en statisties-gebaseerde metodes.

# Acknowledgements

I would like to thank the following people:

- My parents and family for providing me with an education and continuous support.
- Prof. J.A. du Preez, my supervisor, for his patient guidance over eight years of undergraduate and graduate studies.
- Gerhard Esterhuizen and Zelda Weitz for their support, good pasta, great wine and excellent conversation.
- The DSP Lab and all its people for providing a stimulating environment.
- Everyone who contributed to the PatrecII system over the years. Without it, this thesis would not have been possible.
- Koos Hugo and SwerwerII for False Bay sailing.

Dedicated to my father, Flip Nel (11 November 1937 - 10 August 1998)

# Contents

<b>1</b>	<b>Introduction, Motivation and Context</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Background . . . . .	2
1.3	Literature Synopsis . . . . .	3
1.4	Objectives . . . . .	3
1.5	Contributions . . . . .	4
1.6	Overview . . . . .	4
<b>2</b>	<b>Literature Study</b>	<b>5</b>
2.1	Introduction . . . . .	5
2.2	Fujimura . . . . .	5
2.3	Mermelstein . . . . .	7
2.4	Kahn . . . . .	9
2.5	Prinsloo . . . . .	10
2.6	Wu . . . . .	12
2.7	Howitt . . . . .	13
2.8	Discussion . . . . .	14
<b>3</b>	<b>The Role of Syllable-based Information</b>	<b>16</b>
3.1	Introduction . . . . .	16
3.2	Phonological concepts and definition of the syllable . . . . .	16
3.3	The structure of monosyllabic words . . . . .	21
3.3.1	The onset . . . . .	21
3.3.2	The coda . . . . .	22
3.3.3	The peak . . . . .	22
3.3.4	The rhyme . . . . .	22
3.4	Properties of Acoustic Speech in the Syllable . . . . .	22
3.4.1	Vowels and Diphthongs . . . . .	23
3.4.2	Glides, Liquids and Nasals . . . . .	23
3.4.3	Stops (Plosives) . . . . .	24
3.4.4	Fricatives . . . . .	24
3.5	Chomsky and Halle's binary features . . . . .	25
3.6	Phonotactic constraints for the syllable . . . . .	26
3.6.1	Onset phonotactics . . . . .	27
3.6.2	Rhyme phonotactics . . . . .	28
3.6.3	Peak phonotactics . . . . .	29
3.7	Syllable definition used in this study . . . . .	29
3.8	Discussion . . . . .	31

<b>4</b>	<b>Language Model</b>	<b>32</b>
4.1	Introduction . . . . .	32
4.2	Regular Expressions, Regular Languages and Finite Automata . . . . .	32
4.3	A Syllable Grammar . . . . .	34
4.4	HMM as model for an FSA . . . . .	36
4.5	Hierarchical HMM . . . . .	37
4.6	Discussion . . . . .	38
<b>5</b>	<b>Database</b>	<b>40</b>
5.1	Introduction . . . . .	40
5.2	SUNSpeech . . . . .	41
5.3	Cleaning up the database . . . . .	41
5.4	Errors found . . . . .	45
5.5	Data composition . . . . .	46
5.6	Discussion . . . . .	47
<b>6</b>	<b>Acoustic syllable segmenter</b>	<b>49</b>
6.1	Introduction . . . . .	49
6.2	Signal processing . . . . .	50
6.3	Phones . . . . .	50
6.4	DTW distance measure . . . . .	50
6.5	HMM . . . . .	52
6.6	Syllabification Results . . . . .	54
6.7	Discussion . . . . .	55
<b>7</b>	<b>Conclusion and Directions for Future Research</b>	<b>57</b>
7.1	Discussion . . . . .	57
7.2	Future Work . . . . .	58
7.3	Conclusion . . . . .	58
<b>A</b>	<b>Glossary</b>	<b>61</b>
<b>B</b>	<b>Sunspeech Corpus</b>	<b>65</b>
B.1	The Allowed Phonemes with their ‘ASCII code’ for the DSP Speech Database	65
B.2	Addendum to original documentation . . . . .	65
B.2.1	Vowels . . . . .	66
B.2.2	Diphthongs . . . . .	67
B.2.3	Fricatives . . . . .	67
B.2.4	Glides . . . . .	67
B.2.5	Liquids . . . . .	68
B.2.6	Nasals . . . . .	68
B.2.7	Stops . . . . .	68
B.2.8	Affricates . . . . .	69
B.2.9	Other . . . . .	69
B.3	Chomsky Features for Phonemes in Sunspeech . . . . .	70
B.3.1	Vowels . . . . .	70
B.3.2	Diphthongs . . . . .	71
B.3.3	Fricatives . . . . .	71
B.3.4	Glides . . . . .	71

B.3.5	Liquids . . . . .	72
B.3.6	Nasals . . . . .	72
B.3.7	Stops . . . . .	72
B.3.8	Affricates . . . . .	73
B.3.9	Other . . . . .	73
B.4	Description of features used . . . . .	74
B.5	Utterances in Sunspeech . . . . .	75

# List of Figures

2.1	Mermelstein's loudness function and convex hull for a speech segment (from [11]).	8
2.2	Syllable representation of the word <i>Jennifer</i> (from [17]).	9
2.3	Simplified model of the acoustical syllabifier, showing acoustical feature extraction with automatic adaptation and the HMM as syllable and phoneme class parser (from [4]).	11
3.1	A block diagram of the human speech production system (from [19, p. 103]).	17
3.2	Relative sonority of the words a) <i>clamp</i> and b) <i>Andrew</i> .	18
3.3	Relative sonority of the words <i>hidden aims</i> and <i>hid names</i> .	19
3.4	Relative sonority of the word <i>phonology</i> .	19
3.5	Relative sonority of the word /pljəʊlmp/.	20
3.6	Relative sonority of the word <i>sticks</i> .	20
3.7	The syllable structure of monosyllabic words.	22
3.8	Average formant locations for vowels in American English (from [19]).	23
3.9	Sonority scale (feature-based version) (from [3]).	26
3.10	A complete template for English language syllables (from [3]).	27
3.11	Syllable structure for the word, <i>new</i> , as suggested by [3].	28
3.12	Syllable structure for the word, <i>next</i> .	29
3.13	The structure of monosyllabic words (repeated here).	29
4.1	The relationship between finite automata, regular expressions, and regular languages (from [20]).	33
4.2	State diagram of a finite automaton for the syllable grammar presented in section 4.3 (from Prinsloo [4]).	36
4.3	The syllable HMM model used in this study. The HMM states are defined in table 3.3.	37
4.4	An example of a two-level hierarchical HMM where the upper level is a simple $n$ -state parallel HMM model, and the lower level consists of a collection of phoneme HMMs which constitutes the parallel states.	39
6.1	The overlap between two syllables, $i$ and $j$ , where one will be the original transcription and the other generated by our segmenter model from the acoustical data.	51
6.2	The syllable segmenter model used to generate syllable-level transcriptions from the acoustical data.	53
6.3	The trained syllable model. Transition probabilities as trained are shown and rounded to two decimals.	53
6.4	The parallel HMM model used to represent the phoneme classes.	54

# List of Tables

2.1	Syllable constituents for English (from [1]). The vertical distribution of elements is not our concern here. The elements in the same horizontal position, i.e. with the same vowel affinity value, do not co-occur in the same initial or final consonant cluster. The structures and phonotactic constraints of vowels (V) and glides (W, J, and H) are not discussed in Fujimura's paper. . . . .	6
2.2	Prinsloo's syllable classes . . . . .	10
2.3	Prinsloo's phoneme class segmentation results (from [4] [6]) . . . . .	11
2.4	Word-error rates for each individual system and combined for evaluation test set (from Wu [2]) . . . . .	13
2.5	Mermelstein's algorithm as applied to TIMIT database by Howitt (from [12] [15]). The "broadband" condition is Mermelstein's original frequency range (500Hz - 4KHz), and the " $F_1$ " range is 0 - 650Hz. . . . .	14
3.1	The sonority scale with rising sonority from left to right and classified by phoneme class (from [3, p. 133]). . . . .	18
3.2	Chomsky and Halle's binary phonological features applied to a selection of consonants and vowels. . . . .	25
3.3	The syllable classes used in our syllable model. . . . .	30
3.4	Binary features for syllable classes (slightly adapted from Prinsloo's work and repeated here) [4]. . . . .	31
4.1	Chomsky hierarchy and the corresponding machine that accepts the language (from [21]). . . . .	33
5.1	The binary features not used in the discrete features experiment. . . . .	42
5.2	The binary features used in the discrete features experiment. . . . .	42
5.3	An example of a transcription file and the corresponding binary feature file for utterance HJM10051. . . . .	44
5.4	The syllable structures found in the SUNSpeech database. . . . .	46
5.5	The frequency of the eight most frequent syllable structures in the Switchboard corpus (from [13]). . . . .	46
5.6	The frequency of words with $N$ syllables in the Switchboard vocabulary and corpus (from [13]). . . . .	47
5.7	A summary of the sizes and composition of the data used by prominent syllabification studies. . . . .	47
6.1	The syllabification results achieved by our model on the SUNSpeech test set. . .	54
B.1	Sentences in the English subset of Sunspeech . . . . .	76

# Chapter 1

## Introduction, Motivation and Context

*Could mortal lip divine  
The undeveloped freight  
Of a delivered syllable  
'Twould crumble with the weight.*  
- Emily Dickinson, "A Syllable"

### 1.1 Motivation

The syllable was proposed as a unit of automatic speech recognition as early as 1975 due to its strong links with human speech production and perception [1]. It has been suggested that many prosodic properties such as pitch, accent and stress are most naturally expressed in terms of syllables. Some researchers hypothesize the syllable to be the primary unit of segmentation in speech and the basic unit of lexical access in the human mind. Recently, it has been proved that incorporating information from syllable-length time scales into automatic speech recognition improves results in large vocabulary recognition tasks [2].

Segmenting a speech signal into reliable pre-defined segments, or recognition units, is a difficult problem in continuous speech recognition. Typically, a set of acoustic cues and rules is defined and employed to segment units such as phones, syllables or words. The syllable has, however, particularly interesting phonological properties which can be employed to segment speech [3] [4] [5]. Previous studies have used acoustical features such as peak fundamental frequency ( $F_0$ ) values, energy integrals, duration and rising  $F_0$  contours. On a more language-specific level, the phonotactic constraints of phonemes

within the syllable have also been used in previous studies to segment syllables.

Most studies on syllabification over the last three decades have focused on knowledge-based methods. This study will investigate acoustical methods for syllable segmentation, with a focus on being language independent.

## 1.2 Background

During the early 1990's, Prinsloo conducted a study at the University of Stellenbosch on automatic syllabification in Afrikaans [4] [6]. At the same time, the development of an extensive software toolkit for use in pattern recognition, PatrecII [7], was started in the department under the leadership of Prof. Johan du Preez. This toolkit matured to such an extent in the late 1990's that it suddenly made complex speech experiments feasible. Du Preez developed the theory of higher-order hidden Markov models as part of his PhD in the same period and applied this to the domain of automatic language recognition [8].

This study was launched as a vehicle to revisit Prinsloo's earlier work and to use these new tools at our disposal on a complex problem. Work on topic-spotting was under way in the same lab at the time, and investigating the syllable, as a bigger unit of speech for possible use in these problems, was a natural step.

In 2000, the African Speech Technology project was launched with the goal of promoting African languages for use in speech recognition. As a result, speech databases were created for various indigenous African languages. During the initial part of this work, the syllable model we developed showed promise as a tool to cross-check the hand transcription and labelling of such databases. Other studies have also shown that the syllable is useful in automatic foreign accent identification which might have had possible use in the AST project [9].

It was anticipated that automatic syllable segmentation and the notion of a syllable as a basic recognition unit, could also be used as a tool for linguists in empirically analyzing syllable rules and comparing these empirical results with actual linguistic theory. Especially in the domain of higher order hidden Markov models, it was anticipated that the higher order links might show up context dependant rules and phonotactics.

## 1.3 Literature Synopsis

We summarize briefly the most prominent works on syllabification over the past three decades. Fujimura proposed the syllable as a basic unit of speech recognition as far back as 1975 [1]. In the subsequent year, Daniel Kahn published his influential study on English syllabification in his PhD "*Syllable-based Generalizations in English Phonology*" [10]. This is still one of the most widely used and complete definitions of the English syllable.

Most of the early work on automatic segmentation of speech into syllables used knowledge-based methods. Various algorithms have been proposed to automatically segment speech into syllables, of which Mermelstein's convex hull method was one of the first in 1975 [11]. He achieved a token error rate (TER) of 9.5%, albeit on a limited data set of eleven sentences produced by only two male speakers. When the same algorithm was used to segment TIMIT into syllables, overall performance dropped to a TER of 26.6% [12]. When modified slightly as reported by Howitt in [12], it improved to a TER of 14.6%.

Recently, the focus has been on statistically-based methods. These have their own inherent problems in that statistical methods are unable to handle conditions that are not present in their training data. Most recently Wu [2] [13] reported a 21% error rate on a subset of OGI Numbers95 using RASTA PLPs as input to a multilayer perceptron.

Prinsloo introduced the concept of using binary phoneme class features as input to a HMM-based syllable parser [4]. He achieved good results, but again on a very limited dataset. We build and extend on his research in our own work.

## 1.4 Objectives

Our goal with this study was to define a model for the syllable that could be used to automatically segment syllables from an untranscribed acoustic speech signal. The model was to be descriptive enough to account for the large majority of syllables found in the English language, but it was to be general enough for it to be easily adapted and retrained for other languages.

This model was then to be used to segment a speech database from the acoustic signal, and to compare the inserted syllable boundaries with the original hand-transcribed ones. This segmenter could be used as a tool to build databases of syllables. These could be used in various aspects of speech research.

A detailed token error rate (TER) measure was to be defined to compare results as this has been a common downfall of previous studies.

Therefore, our objectives were threefold:

- develop an algorithm to automatically segment speech into syllables
- use this algorithm to segment an untranscribed English acoustical speech signal
- clearly report the results in a well-defined success measure that can be used by others in future studies

## 1.5 Contributions

We succeeded in developing a generic model for the English syllable. This model can easily be adapted for other languages by incorporating the phonotactic constraints of such languages. We used the concept of hierarchical hidden Markov models to construct a syllable model. When used for syllabifying a speech database, we achieved a token error rate of 20.3% which compares well with previous studies.

We defined a detailed success measure for our dynamic time warping (DTW) algorithm which can be used in future studies where syllable recognition is to be measured and results consistently reported and compared.

During the course of this study, results were published in [5] and [14].

## 1.6 Overview

Chapter 2 provides an overview of the pre-eminent articles in the field of syllable recognition and syllable linguistics over the past thirty years. Amongst linguists there are a number of different viewpoints on the definition of a syllable. Chapter 3 will discuss syllable linguistics and will define and refine the English syllable model that was used in this study. In Chapter 4, we present key computational linguistic concepts used in the modelling of syllables by means of regular grammars. We also discuss our use of hierarchical hidden Markov models. In Chapter 5, the SUNSpeech corpus used in our experiments is discussed as well as a preliminary syllable model which we used to do a sanity check on the existing transcriptions. Chapter 6 explains the acoustic model and syllable segmenter used in our syllable recognition experiments as well as the results achieved. Finally we conclude in Chapter 7 and give some direction for future research.

# Chapter 2

## Literature Study

In this chapter, we discuss briefly all the pre-eminent works on syllables and syllabification that were published in the last thirty years. They are discussed in rough chronological order of publication.

### 2.1 Introduction

The syllable has been revisited several times over the last three decades as a unit of speech recognition. Since so much success has been achieved with phonemes as a basic unit in large vocabulary continuous speech recognition (LVCSR) systems which are finally usable commercially, attention is going back to the syllable as a possible path through which to improve performance. It has been proved that the syllable can increase recognition, especially in noisy environments where context plays an important function in human speech recognition. We briefly summarise the most important works in the past thirty years in terms of proposing the syllable as a unit of recognition, as well as the most used syllabification algorithms.

### 2.2 Fujimura

In 1975, Osamu Fujimura published his paper "*Syllable as a Unit of Speech Recognition*", one of the first proposing the syllable as a basic unit of speech recognition [1]. He argued the necessity of a successful phonetic unit as a productive constituent element of phonetic forms for a given language so that both recognition and synthesis can be based on it. He proposed the syllable as a unit that can account for context sensitive effects but which is still computationally tractable to be of use for large vocabulary tasks.

The most common form of recognition used at the time was stored templates and dynamic programming. As such, they were greatly limited by the number of different variations implicit in morpheme templates, and they had to find a way to reduce this number by using a phonological unit smaller than the morpheme. Hence, Fujimura's suggestions are largely based around optimising this technique. The variation due to allophonic effects in phonemes make it a less than ideal unit to use in template matching speech algorithms. He therefore suggests the syllable as a more practical unit.

Fujimura discussed the syllable structure of English to some length and proposed a sub-classification of syllabic components based on vowel affinity of consonantal segments. The higher the vowel affinity for a consonantal segment, the closer it must be to the syllable nucleus, which itself is typically a vowel. This is shown in Table 2.1.

He expanded on the phonotactic constraints present in the English language by explaining that the maximum length for initial consonant clusters in English is three phonemes, invariably starting with /s/ and followed by a voiceless stop and then a liquid or a semivowel (glide), e.g. "strike," "skew".

High	←	Vowel Affinity						→	Low
		Sonorants			Obstruents				
			l	m	b	ɔ̥	d		
				n	g				
	V	W	r				ʒ		z
		J					ð		
		H				f	ʃ		x
					sp				
					sk				
					p	ts	t	st	

Table 2.1: Syllable constituents for English (from [1]). The vertical distribution of elements is not our concern here. The elements in the same horizontal position, i.e. with the same vowel affinity value, do not co-occur in the same initial or final consonant cluster. The structures and phonotactic constraints of vowels (V) and glides (W, J, and H) are not discussed in Fujimura's paper.

Various exceptions to these general constraints were discussed. One example is the fact that English has a peculiarity in respect of the vowel affinity ordering principle in the use of /s/, and Fujimura proposed that the phonemic sequences /sp/, /st/, /sk/ be treated as single consonantal units for this purpose. This treatment explained those contrasts like "task," "tax," or "text," which would otherwise have to be exceptions to the ordering principle. In the case of initial clusters, Fujimura suggested that we are able to assume that each of the classes like sonorants versus obstruents as such, is given one position in the scale, and only one element in the class can be selected for the same class position.

This principle was used in our own model of the syllable which we explain in Chapter 3.

The placing of syllable boundaries between two contiguous syllables of multisyllabic forms is a difficult problem in languages like English. Fujimura suggests a way of solving the problem by observance of the stronger adherence of the intervocalic consonant to the stressed syllable. We will see in section 2.4 how Daniel Kahn, in the following year, expanded on the problem of ambisyllabic syllable parts and produced a system for treating them.

Fujimura suggested that when syllables are analysed in terms of syllabic features it is most natural to classify syllables in terms of classes of features, i.e., nucleus, initial, and final, and further in terms of subclasses within each class. This system is used in most subsequent work on syllabification [3] [4].

## 2.3 Mermelstein

Mermelstein in his 1975 paper "*Automatic Segmentation of Speech into Syllabic Units*" describes his "convex hull" algorithm - one of the first knowledge-based segmentation algorithms for syllables [11]. The algorithm allows the significance of a loudness minimum to be a potential syllabic boundary from the difference between the convex hull of the loudness function and the loudness function itself. Mermelstein calculated this loudness or intensity measure over a "broadband" range of 500Hz to 4KHz.

Mermelstein's algorithm uses a unique recursive technique to find syllable peaks and boundaries, which reflects the effect of context by comparing the dips and peaks to their immediate surroundings. It has been used in several successive projects on knowledge-based speech recognition systems in the years following the publication, most recently in Howitt's work which we discuss in section 2.7 [12] [15].

Figure 2.1 illustrates the implementation of the convex-hull algorithm. Since this is such a well-used algorithm and one against which we compare our results, we repeat Mermelstein's own explanation of this algorithm in the following paragraph.

"An original speech segment over the interval  $(a-c)$  is found to possess a loudness function  $l(t)$  with maximum at point  $b$ . The convex-hull computed for the segment  $(a-b-c)$  is  $h_1(t)$ . Over the interval  $(a-c)$ , the maximum hull-loudness difference is  $d_1$  at  $c'$ . If  $d_1$  exceeds the threshold, segment  $(a-b-c)$  is cut up into segment  $(a-c')$  followed by segment  $(c'-b-c)$ . The hull for segment  $(a-c')$ , defined around the new maximum point  $b'$ , follows the loudness curve. This results in a zero hull-loudness difference over that interval and hence that

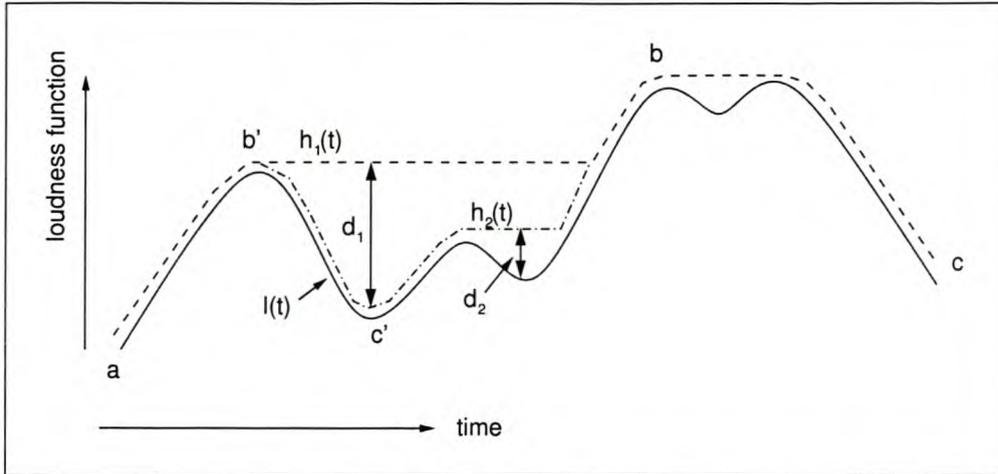


Figure 2.1: Mermelstein's loudness function and convex hull for a speech segment (from [11]).

portion is not segmented further. The hull for segment  $(c'-b-c)$ , denoted by  $h_2(t)$ , is shown by the short dashed line where it differs from  $h_1(t)$  over the segment interval. The new maximum hull-loudness difference is found to be  $d_2$ . If  $d_2$  does not exceed the threshold then the segment  $(c'-c)$  is not divided further." [11].

The threshold parameter mentioned is 2dB. The algorithm includes durational and absolute level constraints as well. Syllable segments are required to be at least 80ms long. Syllabic peaks are required to be no more than 25dB below the overall intensity peak. Mermelstein's algorithm does not proceed from left to right in time. It assumes that the entire utterance is stored before processing commences, but requires only that a complete segment delimited by silent intervals be captured before segmentation starts.

Mermelstein suggested that inclusion of alternative fluent-form syllabifications for multi-syllabic words and the use of phonological rules for predicting syllabic contractions can further improve agreement between predicted and experimental syllable counts. We will see in section 2.4 that Kahn uses this principle extensively to syllabify words according to their rate of speech.

Mermelstein reported a 9.5% TER, 6.9% deleted, 2.6% inserted. He used two male speakers to generate slow read speech, 11 sentences each (half of which included monosyllabic words), for a total of 22 utterances and 418 syllable tokens. While this was adequate to demonstrate the algorithm's utility, this research required a more comprehensive database. Howitt's work, discussed in section 2.7, repeated this algorithm on a much bigger dataset.

## 2.4 Kahn

Kahn introduced the notion of the hierarchical theory of the syllable in his influential PhD thesis at MIT, *"Syllable-based Generalizations in English Phonology"* (1976) [10]. Kahn proposed in this study to extend the notion of phonological representation assumed in such works as Chomsky and Halle's *"The Sound Pattern of English"* (1968) [16] by introducing a new tier of representation involving strings of the symbol 'S', representing the node "syllable". As seen in Figure 2.2, by counting the S's in the upper tier in this representation we determine the number of syllables.

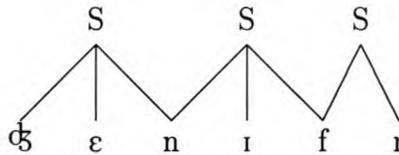


Figure 2.2: Syllable representation of the word *Jennifer* (from [17]).

We also see that the syllables in question are the sequences /dʒɛn/, /nɪf/ and /fr/. A feature of Kahn's representation is the fact that certain segments may be ambisyllabic in that they belong to two elements of the upper tier. In this example, /n/ and /f/ are ambisyllabic and are shared between two syllable segments.

Kahn showed a number of benefits in representing the syllable as a hierarchical unit as opposed to a linear phonological representation. Several phonological productive processes operate on each other on a lower hierarchical level which the linear phonological representations are incapable of adequately describing. Kahn also proposed different rules of syllabification based on the rate of speech used in producing the sample.

Kahn's hierarchical representation is deemed by some researchers to be insufficiently rich in that it does not distinguish syllable peaks from marginal elements [17]. Other studies proposed an extra set of constituents smaller than the syllable, which took vowel and consonant segments as their members. These are termed the onset, nucleus (or peak) and the coda. Clements and Keyser in their book *"CV Phonology"* [17] argue that the notion of syllable and peak is enough of a theoretical framework in order to fully describe phonological concepts, and that the extra tier of onset and coda unnecessarily add to the complexity without much value added in terms of descriptive rules, etc. They propose a 3-tiered approach on only the CV level (consonant, vowel).

Kahn's set of hierarchical rules for syllabification of English language syllables is still the most recognised definition and has been used in various subsequent studies as the benchmark against which syllabification results are measured [2] [12]. His theory has also been implemented by NIST in a software package **TSYLB2**, which is now widely used to syllabify linear phoneme transcription inputs and insert syllable boundaries for various rates of speech [18].

## 2.5 Prinsloo

Prinsloo and Coetzer, then with the same department as the author, published their article "*Automatic syllabification and phoneme class labelling with a phonologically based hidden Markov model and adaptive acoustical features*" in 1990. It used phonological features as input to an HMM model which parsed syllables and indicated syllable boundaries in continuous speech. They used a single layer neural network to first determine phoneme classes for the features [*sonorant*] and [*syllabic*], as described by Chomsky and Halle [16]. This was used as input to an HMM where the various states represented syllable classes and the inter-state connections modelled some of the phonotactic constraints of the Afrikaans language.

The syllable classes were determined from the speaker independent binary phonological features [*syllabic*] and [*sonorant*] as shown in Table 2.2. The class of [*+syllabic*] segments includes the ordinary vowels as well as the so-called syllabic sonorants, such as the last segment of the English word, *button*, in its most common pronunciation, /'bʌtʰn/.

Description	Features
vowels and diphthongs	[ <i>+syllabic</i> ] [ <i>+sonorant</i> ]
voiced consonants	[ <i>-syllabic</i> ] [ <i>+sonorant</i> ]
unvoiced consonants	[ <i>-syllabic</i> ] [ <i>-sonorant</i> ]
Silence	

Table 2.2: Prinsloo's syllable classes

They then used the basic rules of sonority to define a regular grammar for the classes in Table 2.2, allowing a finite automaton to parse these classes and syllables in acoustic speech. An HMM was then used to implement this FSA.

Figure 2.3 describes the process used by Prinsloo.

The notion of using an HMM to parse a set of syllable classes is also used and extended in our own study to model and segment syllables. We however used the direct acoustic signal

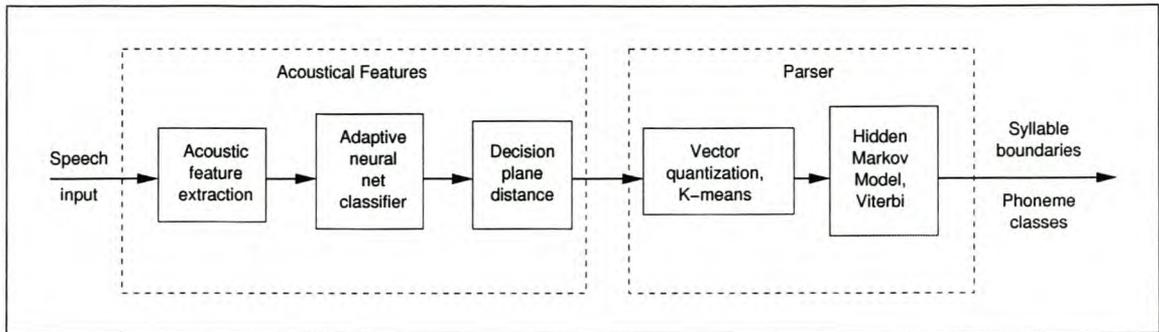


Figure 2.3: Simplified model of the acoustical syllabifier, showing acoustical feature extraction with automatic adaptation and the HMM as syllable and phoneme class parser (from [4])

to first recognise individual phonemes, which are then categorized in the broad syllable classes described by Table 2.2 by way of hierarchical sets. From that level onwards, our model essentially operates in the same way as Prinsloo's.

Prinsloo performed his experiment on a speech database with recordings of 10 male Afrikaans speakers, each with two excerpts of 7 seconds (2.5 minutes of speech comprising 370 syllables).

He detected nuclei with 97.3% accuracy. However, he does not describe a measure for calculating accuracy when boundary errors are taken into account.

Type	Percentage
Correct	96.5%
Deletions	2.1%
Substitutions	1.3%
Insertions	0.1%

Table 2.3: Prinsloo's phoneme class segmentation results (from [4] [6])

Prinsloo's work, like Mermelstein, suffers from the fact that it was performed on a very small dataset. In Mermelstein's case, the data was sufficient to prove the principle. However, Prinsloo reports exceptional results which are not really significant when the size of his dataset is taken into account. Keeping in mind that this study was undertaken more than 15 years after Mermelstein's, more than enough advances were made in computing power in order for Prinsloo to test his results on a bigger set of data. The effort to test some of these principles on a bigger dataset is thus also one of the primary reasons for our own study. We built on Prinsloo's use of an HMM to model a regular grammar, which parses syllables based on broad phoneme classes. Our own set of data is sufficient in size to be more comparable to recent studies in the field of syllabification. Two such studies are discussed in the next two sections.

## 2.6 Wu

In her 1998 Berkeley PhD thesis "*Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*", Su-Lin Wu investigates the effect of using syllable information in automatic speech recognition (ASR). She concedes that the implementation of syllable-based recognisers has proven to be challenging; she states that the syllable is an attractive unit for recognition for these reasons:

1. Syllable representations and durations may exhibit greater stability to phoneme-based representations and durations.
2. Syllables appear to offer a natural interface between speech acoustics and lexical access.
3. Syllables constitute a convenient framework for incorporating suprasegmental prosodic information into recognition.

A subset of the OGI Numbers corpus, which consisted of a 32-word vocabulary restricted to numbers, was used for her experiments. It had a training set of 3600 utterances (example "eighteen thirty one") and a development test set and an evaluation test set, each containing about one hour of speech (1200 utterances). Wu implemented a baseline recogniser using traditional phone-based recognition, a syllable-based recogniser using modulation spectrogram features over an extended context window, and a combined recogniser using elements of both of the above in a weighted manner.

The baseline recogniser used RASTA-PLPs as input to a multilayer perceptron (MPL) phonetic classifier with 400 hidden units. It took features from 105ms of speech (9 frames of 25ms with a 10ms step) and classified them into 32 phone categories. A multiple-pronunciation lexicon with simple minimum duration modelling was developed for the baseline recogniser. Embedded Viterbi alignment was applied iteratively to optimize the lexicon pronunciations, minimum phone durations, and training labels.

The syllable-based recogniser used modulation spectrogram features for the front-end speech representation. These were computed with a set of 15 channels using a FIR filter bank. Again an MLP was used with a single hidden layer of 400 units, though the syllable-based system used an extended context window of 185ms (17 frames) and classified the features into 124 "semi-syllable" categories. The syllable-based lexicon was derived from the baseline system's lexicon.

The combined recogniser combined results of the baseline and syllable-based recognisers at the whole-utterance level. Each recogniser was used to generate a word lattice for

each input utterance. For each utterance, an  $N$ -best list is generated from each lattice, the two  $N$ -best lists are concatenated, and duplicate hypotheses are eliminated. For each hypotheses in these merged lists, two acoustic scores were calculated via forced alignment using the baseline and syllable-based recognisers and a language model score was calculated from the backoff bigram grammar. The final score for each utterance is the weighted sum of the two acoustic scores and the language model score. Wu applied an empirically determined weighting factor.

The results of the three different recognisers are shown in Table 2.4.

System	Clean	Reverb.
RASTA, phone units, 105ms context window Baseline	6.8%	27.8%
ModSpec, syllable units, 185ms context window	9.8%	30.9%
Combined	5.5%	19.6%

Table 2.4: Word-error rates for each individual system and combined for evaluation test set (from Wu [2])

Wu proved to some extent that using syllable level information improves recognition rates in large vocabulary continuous speech recognition when combined with traditional phone-based recognisers. Her study finally backs up Fujimura’s suggestion of incorporating the syllable as a basic unit of speech recognition.

## 2.7 Howitt

Andrew Howitt in his MIT PhD dissertation “*Automatic Syllable Detection for Vowel Landmarks*” [12], investigated landmark-based speech processing as a component of lexical access from features (LAFF). His work implemented a Vowel Landmark Detector using a syllabic segmentation algorithm. He used Mermelstein’s segmentation algorithm as described earlier, but on a much bigger dataset. Howitt implemented this on the TIMIT database, and since his study was performed more than two decades after Mermelstein published this algorithm, computational advances allowed Howitt to do it on a much bigger dataset. This is, therefore, one of the few studies in recent times that provides us with a usable syllabification benchmark on a relatively large dataset against which we can compare the performance of our own algorithm.

As the first part of his experiment, and very useful for us, Howitt repeated Mermelstein’s experiment on a subset of the TIMIT database. TIMIT does not have syllable level transcriptions, and Howitt used the **TSYLB2** program from NIST to add syllable boundary transcriptions using the phonetic transcriptions as input [18]. The training set was 619

utterances and 7585 syllable tokens, and the test set, 373 utterances and 4404 syllable tokens.

	Train				Test			
	Detect	Insert	Delete	TER	Detect	Insert	Delete	TER
broadband	76.2%	2.40%	23.8%	26.2%	77.7%	2.61%	22.3%	24.9%
F1 range	88.4%	1.82%	11.6%	13.4%	87.1%	1.73%	12.9%	14.6%

Table 2.5: Mermelstein’s algorithm as applied to TIMIT database by Howitt (from [12] [15]). The ”broadband” condition is Mermelstein’s original frequency range (500Hz - 4KHz), and the ” $F_1$ ” range is 0 - 650Hz.

Howitt proved that substantial performance over Mermelstein’s original algorithm can be gained, by modifying the frequency range for peak detection to focus on the first formant. He stated that the definition of vowel landmarks specify that they should be located around peaks in energy in the region of  $F_1$ . If so, the performance of Mermelstein’s algorithm should improve when the intensity is measured in a band around  $F_1$ , nominally about 300 to 900Hz. He investigated the effect of this band by independently varying the upper and lower band edges and their rolloff values. He found that the optimal frequency band (0 to 650Hz) does indeed delineate the region where  $F_1$  is likely to be found. By using an intensity measure on this band, he increased his TER rate on TIMIT to 14.6% from Mermelstein’s original 26.2%.

He complains that none of the previous studies on syllabification provides a description on their accuracy measure. His own study bypasses this problem since it only focusses on detecting syllabic nuclei, and therefore does not venture into the more ambiguous area of syllable boundaries. In our own study, we therefore tried to define a detailed measure by which we report our accuracy on detecting syllable boundaries.

## 2.8 Discussion

We provided a brief overview of the most prominent works on syllables in speech recognition in this chapter. We saw that syllabification algorithms are still largely knowledge-based and that this knowledge built into the algorithms is very much dependant on the language to which it is applied. Though some good results were reported by Prinsloo, for instance, they do not carry much weight due to the limited sizes of the datasets on which the experiments were performed. Howitt revisited Mermelstein’s original algorithm and showed that, although it performs relatively well, performance drops on a bigger set of data. Work on acoustic syllabification models is almost non-existent, although Wu proved that the use of information of syllable-length time-scales does improve recognition in ASR

systems. Therefore, we will attempt to focus on a model that uses more of the acoustic information in the speech signal to syllabify, as well as aiming to be able to easily adapt to different languages by simple retraining on an appropriate dataset. At the same time we are attempting an accuracy of at least 20% TER in order to compare with Howitt. We will also define an accurate measure that fully describes our results and that can be used in future studies.

## Chapter 3

# The Role of Syllable-based Information

This chapter discusses the syllable as a possible basic unit of speech recognition, for which there is some empirical psychoacoustic support in the case of humans, and some engineering justification in the case of machines striving to imitate human abilities. It also gives a brief overview of the phonological and phonotactic properties of the syllable in English. Since there are a number of different linguistic theories on what constitutes a syllable, we define the model to which we conform.

### 3.1 Introduction

It has been suggested that many prosodic properties such as pitch, accent and stress are most naturally expressed in terms of syllables. Some researchers hypothesize the syllable to be the primary unit of segmentation in speech and the basic unit of lexical access in the human mind. There is however much ambiguity in the definition of a syllable. We briefly review the most commonly agreed-upon concepts and definitions relating to the syllable. We use examples from Giegerich's "*English Phonology*" [3] in our discussion.

### 3.2 Phonological concepts and definition of the syllable

Before we can start recognising syllables, we need to investigate the phonological concepts that characterise them. Intuitively, syllables seem to be fairly straightforward objects. Speakers will normally have little difficulty in deciding how many syllables a given word

of their language contains. It is still not easy, however, to define a syllable in phonetic and phonological terms. If we look at the initiation process in the production of speech, it is noted that a pulmonic air stream is required for the production of speech. This air stream does not flow at a constant rate; rather, it occurs as a series of pulses which is to some extent caused by bursts of activity on the part of the chest muscles, giving rise to variation in the flow rate of air. In addition to this, there are retardations caused to the flow of air, again by either the chest-muscles or the articulatory process, or both. Figure 3.1 describes this system in block diagram form.

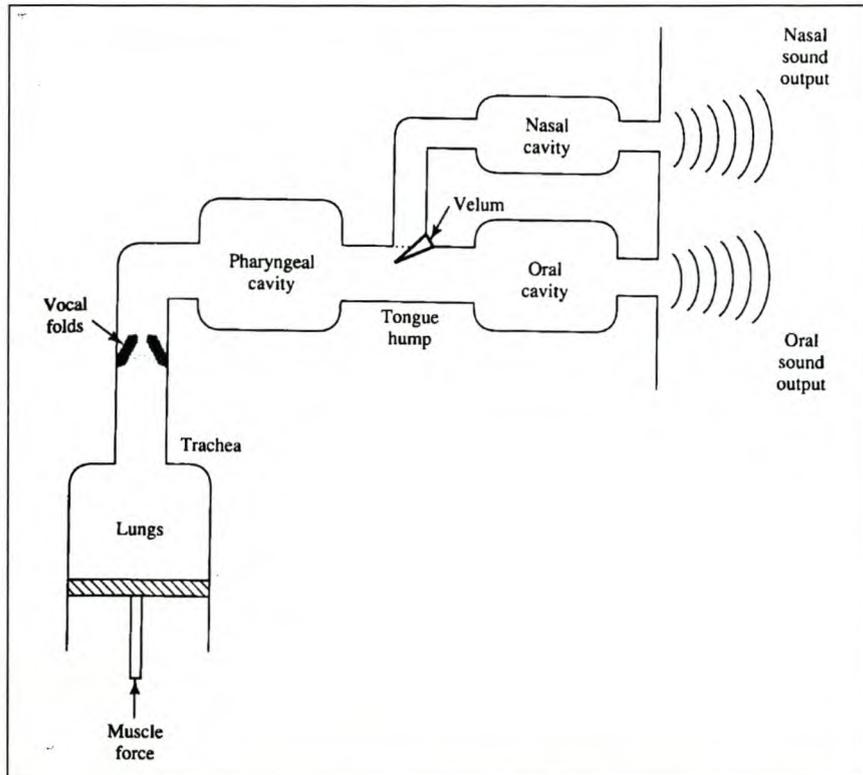


Figure 3.1: A block diagram of the human speech production system (from [19, p. 103]).

When we consider the pronunciation of a series of letters of the alphabet, say P-T-K [pitike], we note how the closures of [t] and [k] interrupt the air flow and how the string consists of three successive bursts of air, coinciding with three syllables. In *E-F* [iɛf], on the other hand, there are no such articulatory closures to be held responsible for the syllable division. Still, the transition from [i] to [ɛf] is marked by a syllable boundary. The retardation in effect here is thus from the first kind described - caused at the source of the air stream by the chest muscles.

According to such a “pulse theory” of the syllable, each syllable coincides with a peak in the flow rate of pulmonic air. The problem with this theory is that it explains very little. It is not measurable if one only has a sound recording. In speech the kinetic energy of the air-stream pulses is translated to acoustic energy, one manifestation of which is sonority. The pulses of the air stream correspond to peaks in sonority.

The sonority of a sound is its relative loudness compared to other sounds, everything else (pitch etc.) being equal. Speech sounds can be ranked in terms of their relative sonority: voiceless oral stops are the least sonorous, while the low vowels have the highest degree of sonority. All other speech sounds fall between these two extremes. Table 3.1 shows a representation of the sonority scale of speech sounds. Fujimura described the same scale as "vowel affinity", which we discussed in section 2.2.

Oral Stops		Fricatives		Nasals	Liquids	Semivowels	Vowels	
Voiceless	Voiced	Voiceless	Voiced				High	Low
p	b	f	v	m			i	a
t	d	θ	ð	n		j		
k	g	s	z	ŋ	l r	w	u	ɑ

Table 3.1: The sonority scale with rising sonority from left to right and classified by phoneme class (from [3, p. 133]).

With the help of this sonority scale, we are able to predict the correct number of syllables corresponding to sonority peaks in the majority of English words (and most other languages). The problem of accurately detecting syllable boundaries still remains.

If we consider a monosyllabic word such as *clamp* in Figure 3.2 a), it is clearly seen that it has only one sonority peak. In comparison, a bisyllabic word such as *Andrew* in Figure 3.2 b), clearly has two sonority peaks reflecting the fact that it has two syllables.

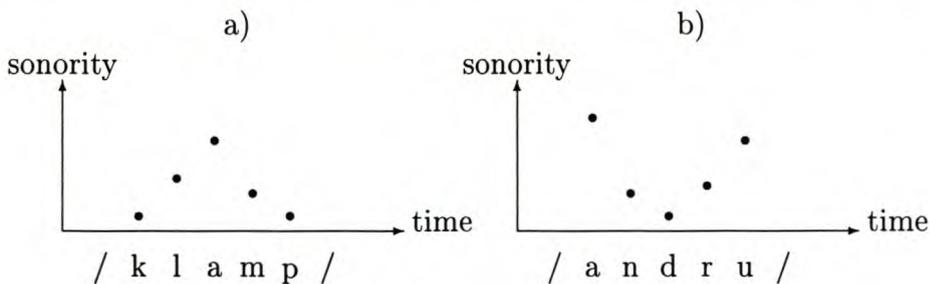


Figure 3.2: Relative sonority of the words a) *clamp* and b) *Andrew*.

However, the sonority theory leaves unexplained a few cases, some of them universal and some of them specific to the English language. For instance, if we consider the two phrases *hidden aims* and *hid names*, both are represented phonetically in an identical manner by /hɪdnemz/. However, although their phonetic representation is identical, *hidden aims* has three syllables and *hid names* has two. The reason for this is quite obvious, but not provided by sonority theory: *hidden* has two syllables and *aims* one, and *hid* and

*names*, one each. If we look at the relative sonority associated with both phrases, they are identical and can be represented as seen in Figure 3.3.

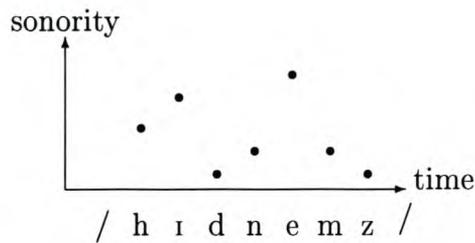


Figure 3.3: Relative sonority of the words *hidden aims* and *hid names*.

It is then clear that sonority does not account for the manner in which syllabification rules take into account the boundaries between words. Stated differently, words are syllabified individually and then put into phrases and sentences.

Another effect for which sonority does not account is the syllable boundaries within words. For example, if we look at longer words such as *phonology*, it is clear from Figure 3.4 that it has four syllables. It does not show, however, the position of the syllable boundaries.

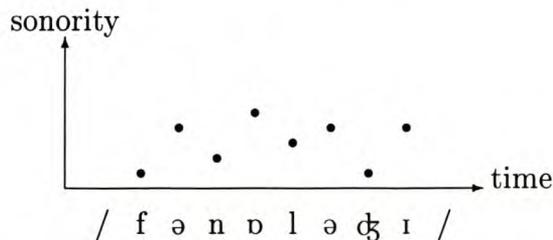


Figure 3.4: Relative sonority of the word *phonology*.

Most speakers would agree on the following syllabification *pho.no.lo.gy*. Sonority theory identifies the troughs between syllable peaks, but does not predict what appears to be quite a simple regularity - the consonant constituting the trough belongs to the following, rather than the previous, syllable. Typically, speakers do not syllabify *phonology* as *phon.ol.og.y*. This is what is referred to as the maximum onset principle.

Next it should be investigated how many phonemes a syllable can contain and what phonemes can occur next to each other in a syllable. If we look at a nonsense example /pljəʊlmp/, it is intuitively clear that this syllable is impossible in English. It is not because of sonority, as it has only one sonority peak as seen in Figure 3.5.

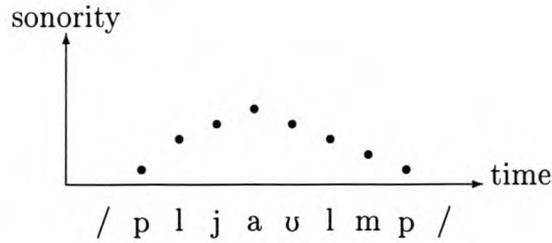


Figure 3.5: Relative sonority of the word /pljauɫmp/.

The reason is clearly that in English no syllable can have that many phonemes. A word like *clamp* just about exhausts the potential of a single syllable. A new problem is also encountered here: if /klamp/ is permissible in English, then why is /knamp/ not? Here again, the sonority theory is not enough to define a syllable. We have to additionally define the number of permissible phonemes and also the way in which the phonemes can cluster. /kn/ is forbidden in English even though it constitutes an upward sonority slope. These phonotactic constraints are discussed in more detail in section 3.6.

Finally, if the monosyllabic word, *sticks*, is considered, we must ask ourselves why it constitutes a single syllable in the English language, if it clearly contains three syllable peaks as seen in Figure 3.6. In the previous examples, none of them actually contradicted the sonority principle. Rather, they amended the rather loose sonority theory of the syllable. In the case of *sticks*, however, we have a clear contradiction.

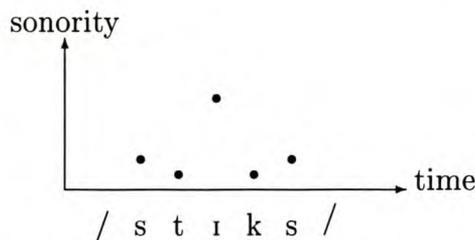


Figure 3.6: Relative sonority of the word *sticks*.

Definitions striving for more technical accuracy are problematic. Every definition seems to have exceptions and caveats or is unsatisfying for practical implementation. For example, consider the following two popular definitions:

- 1) *A syllable is a vowel between optional consonant clusters.*

This, the most popularly understood rule, has many exceptions, since a syllable does

not necessarily contain a vowel. A syllable can instead have a "syllabic consonant" that functions as the nucleus of the syllable; for example, the /l/ in "noodle" or the /s/ in the onomatopoeia, "psst".

2) *Syllables correspond to peaks of sonority.*

Sonority is roughly analogous to the energy contour. Peaks of sonority are, therefore, analogous to regions of greater sound energy and are thought to correspond to the nuclei of syllables. This definition allows consonants to take the place of syllable nuclei, but the sonority-based specification is vague in some cases and can lead to confusions. For example, the unmistakably monosyllabic word, *spa*, is considered by some to have two peaks of sonority. Mechanically segmenting speech into syllables is also difficult. The "maximum onset principle" defines the onset of syllables (the initial consonant clusters) to be as long as possible within the context of the word. For example, the word, *estate*, would be pronounced as *e-state*, according to this rule. The /s/, however, often sounds as if it is shared between syllables. Speakers can pronounce the word as "es-tate," if the first syllable is stressed, an exception to the maximum onset principle. This phenomenon is described in detail by Kahn in his work on ambisyllabicity [10]. He states that it is simply not necessary or possible for polysyllabic words in English to have a well-defined syllable boundary in such cases.

### 3.3 The structure of monosyllabic words

We now show the common perceptual model of the syllable structure. We will expand on this model in later sections, but we need to briefly define the terminology related to the substructure of syllables in order to discuss acoustic and phonological features in subsequent sections. Figure 3.7 shows the parse tree for a syllable.

#### 3.3.1 The onset

Syllables need not necessarily have onsets: *eye*, *eat* and *ink* begin with the syllabic element (the peak). The change in loudness or intensity at the onset of a syllable is generally more abrupt than at its end; thus there is less uncertainty about the onset time of a syllable than about its termination [11].

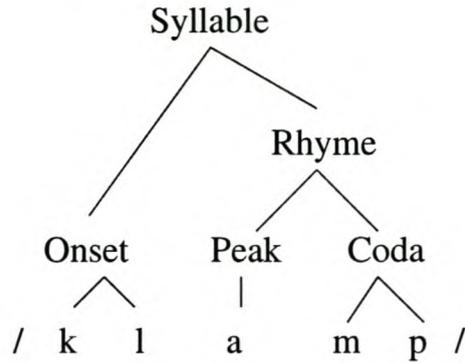


Figure 3.7: The syllable structure of monosyllabic words.

### 3.3.2 The coda

The coda of a syllable is the consonant or group of consonants that follow the peak. It is in several respects similar to the onset, except that it is a mirror image of the onset of course. Like the onset, a coda is optional.

### 3.3.3 The peak

The peak (or nucleus) of the syllable contains the “syllabic” element: the segment that is more sonorous than both its neighbours.

### 3.3.4 The rhyme

The rhyme of a syllable is a unit that contains both the peak and the coda of the syllable. The reason for having a unit, such as the rhyme, is that the peak and coda function together in several settings, and it is indeed this unit that is responsible for the rhyme in poetry.

## 3.4 Properties of Acoustic Speech in the Syllable

In this section, we will discuss the acoustic behaviour of various phoneme classes in the context of the syllable. We will describe a more scientific way to classify these phonemes in section 3.5, where we introduce Chomsky and Halle’s binary features.

### 3.4.1 Vowels and Diphthongs

Vowels can vary widely in duration (typically from 40-400 msec [19, p. 120]). The variation in cross-sectional area along the vocal tract determines the formants of the vowel. Vowels can be distinguished by the location of formant frequencies (usually the first three formants are sufficient). The formant frequencies for a male speaker occur near 500, 1500, 2500, 3500 Hz and so forth.  $F_1$  and  $F_2$  are closely tied to the shape of the vocal-tract articulators. The frequency location of the third formant,  $F_3$ , is significant to only a few specific sounds. The fourth and higher formants remain relatively constant in frequency, regardless of changes in articulation. The formant,  $F_1$ , usually has the highest energy while the second formant,  $F_2$ , varies more than others. Figure 3.8 shows the average formant locations for vowels in American English.

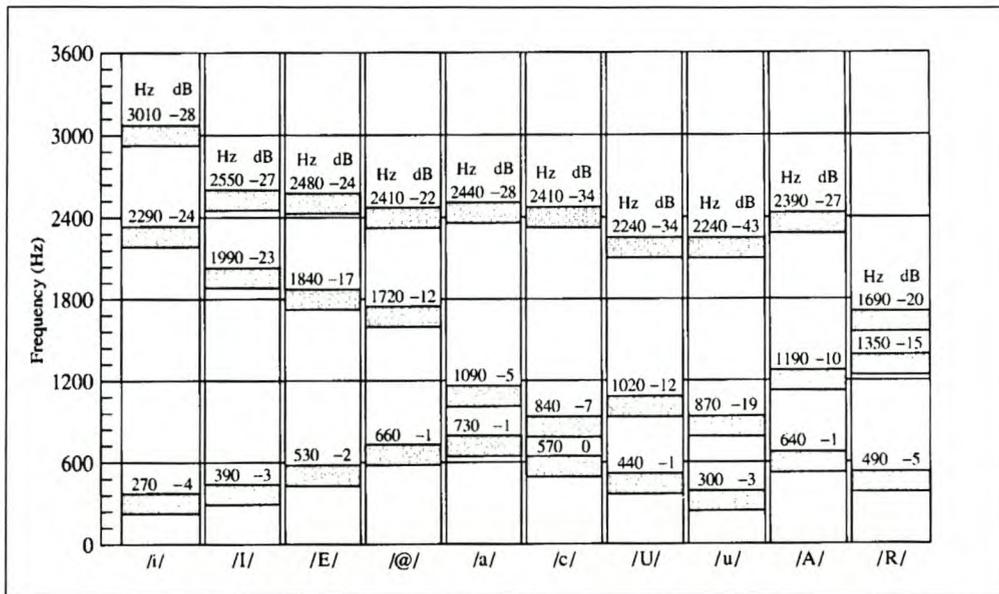


Figure 3.8: Average formant locations for vowels in American English (from [19]).

It is interesting to note that in nuclei the low vowels are usually longer due to the fact that the speaker has to move the slow jaw.

Nuclei of stressed syllables are longer, have more energy, and the  $F_0$  contour is characterized by sharper slopes.

### 3.4.2 Glides, Liquids and Nasals

Phonologically speaking, these consonants are easy to identify as they are always next to nuclei. Voiced consonants that could exist in onset and coda /l,r,m,n/ are usually longer

with more energy in syllable onset. The mean duration of voiced consonants tends to be shortened in onset clusters. This decrease in length is most obvious if the preceding phone is an unvoiced stop, less so for a fricative, and least for a voiced stop. Voiced consonants, with the exception of nasals, tend to lengthen the coda.

Voiced consonants in the onset are longer in stressed syllables while phrase final lengthening in read speech also stretches these consonants to a large extent.

### 3.4.3 Stops (Plosives)

In English, the stop consonants /b,d,g,p,t,k/ are transient, noncontinuant sounds that are produced by building up pressure behind a total constriction somewhere along the vocal tract, and suddenly releasing the pressure. This sudden explosion and aspiration characterizes the stop consonants.

Unvoiced stops are most easily identified by the place of articulation in onset, and formant transitions in the coda. Stop durations are not much influenced in final phrases, but are usually more aspirated in stressed syllables.

### 3.4.4 Fricatives

Fricatives are produced by exciting the vocal tract with a steady air stream, without the vocal cords vibrating. At some point, a constriction then causes this air stream to become turbulent. They are aperiodic with more energy in the higher frequency region. Statistical studies show these consonants to be usually longer in the coda [4].

In onset clusters, unvoiced fricatives tend to be shortened most if followed by an unvoiced stop, a little less for a nasal, and least for semivowels. Coda clusters usually shorten these fricatives in the hierarchy of ensuing phones, unvoiced stops and fricatives.

Voiced fricatives are produced in the same way as their unvoiced counterparts, only with the vocal cords vibrating. Although the voicing bar in these consonants is not prominent, voiced fricatives are shorter than unvoiced fricatives with the difference averaging 40ms in English.

If voiced fricatives are followed by a stop or other fricative in a coda cluster, the phone tends to lengthen. In the onset, however, these consonants stay relatively stable in length. Prosody does not affect fricatives as much as glides, liquids and nasals, while fricatives also lengthen in reading final phrases.

### 3.5 Chomsky and Halle's binary features

According to phonological theory, a phoneme cannot be broken up into smaller units. However, in order to further describe phonemes in a more scientific way phonological features are used. The sum of these individual phonological features or properties fully describes each phoneme. Chomsky and Halle developed these binary features in their landmark, 1968 book, "*The Sound Pattern of English*" [16]. The set of features is shown in Table 3.2 with some examples of how they apply to a selection of phonemes. Refer to Appendix B.4 for a full description of these features as applied to phonemes, appearing in the SUNSpeech corpus, which we used in this study.

feature	consonants					vowels			
	ŋ	θ	ʃ	v	ð	ε	a	ɑ	ɔ
[consonantal]	+	+	+	+	+	-	-	-	-
[sonorant]	+	-	-	-	-	+	+	+	+
[continuant]	-	+	+	+	+	+	+	+	+
[anterior]	-	+	-	+	+				
[coronal]	-	+	+	-	+				
[strident]	-	+	+	+	-				
[round]	-	-	-	-	-	-	-	-	+
[high]	+	-	+	-	-	-	-	-	-
[low]	-	-	-	-	-	-	+	+	+
[back]	+	-	-	-	-	-	-	+	+
[tense]	-	+	+	-	-	-	-	+	+
[voice]	+	-	-	+	+	+	+	+	+
[nasal]	+	-	-	-	-				
[lateral]	-	-	-	-	-				

Table 3.2: Chomsky and Halle's binary phonological features applied to a selection of consonants and vowels.

In a given set of phonemes and features, each phoneme will differ from every other phoneme in terms of at least one of the plus/minus specifications of the features. As can be seen from the table above, these features can, therefore, be used to define various phoneme classes.

When these features are applied to the sonority hierarchy, a diagram as shown in Figure 3.9 is formed. Of particular interest to us is, of course, the feature  $[\pm\text{sonorant}]$  as this indicates the level of sonority of the phoneme.

Using this sonority scale, we can now investigate the phonotactic constraints that govern the syllable in English.

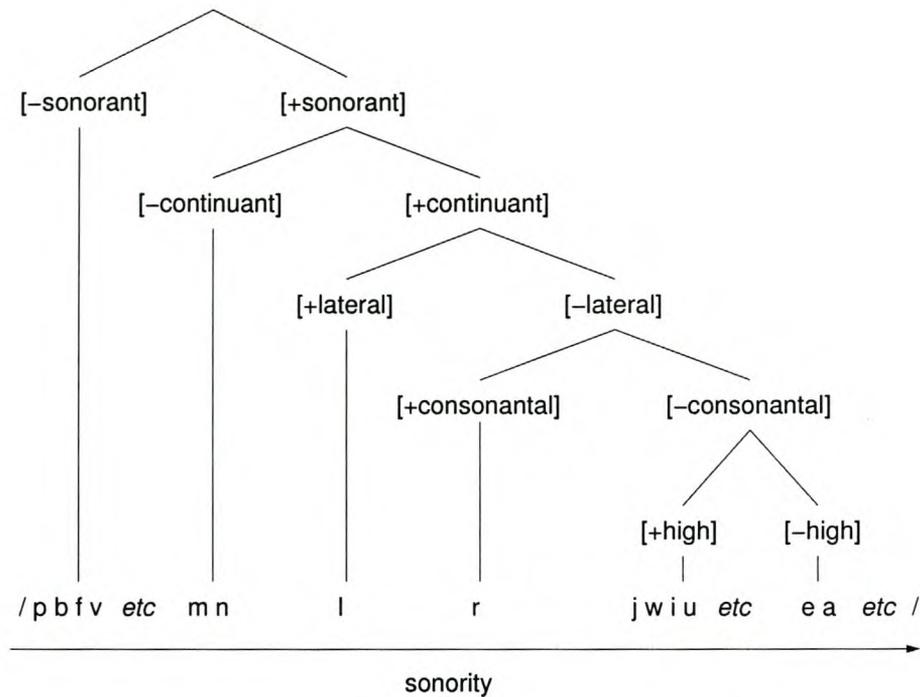


Figure 3.9: Sonority scale (feature-based version) (from [3]).

### 3.6 Phonotactic constraints for the syllable

A powerful property of the syllable, which is relatively under-utilized in automatic speech recognition, is the phonotactic constraints or positional constraints of phonemes within the syllable. These constraints allow only 49 and 60 consonant combinations in English onset and coda positions respectively [17].

When we expand on Figure 3.7 by using the phonological features, we are able to construct a syllable template for English as shown in Figure 3.10.

Figure 3.10 now fully defines the syllable in English with the following conditions:

- 1  $X_1$  plus one  $X_{>1}$  are obligatory
- 2  $X_2$  is associated with the peak if [-consonantal], otherwise with the coda
- 3 Further features of  $X_{1-3}$  decrease in sonority from left to right, in accordance with the sonority scale.

These conditions are further amended for unstressed syllables:

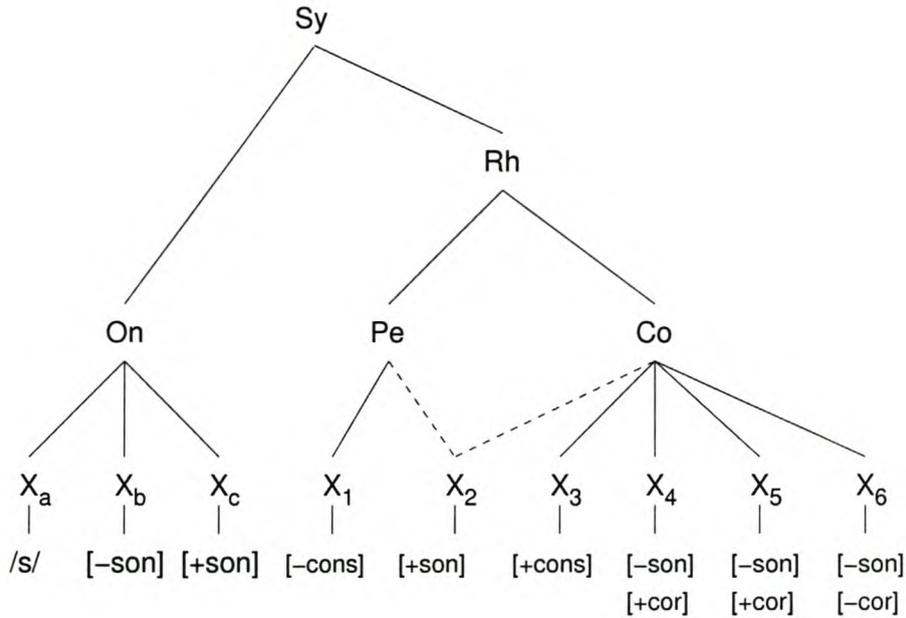


Figure 3.10: A complete template for English language syllables (from [3]).

- 4  $X_1$  may be occupied by any [+sonorant]
- 5 Only  $X_1$  is obligatory

We discuss onset and rhyme phonotactics in more detail below.

### 3.6.1 Onset phonotactics

As seen in Figure 3.10, in English, the onset normally contains a maximum of two consonants, except when it starts with an /s/ as in /str/ and /spr/. This is because a leading /s/ also violates the principle that onset sequences increase in sonority. However, it is only /s/ that can violate the sonority generalisation of the onset. Whenever an onset contains three X-positions, the first *must* be /s/. Onset appendices can only occur before /p/ /t/ and /k/. The first consonant in position  $X_b$  must be [-sonorant], with the following one in position  $X_c$  which must be [+sonorant]. In English, the phonemes /ʒ/ and /ŋ/ are never found in single X-position onsets.

In onsets containing two X-positions, the phonemes /v ð z ʒ/ (all [-sonorant, +continuant, +voice]) do not occur at all.

One other interesting and common case is the phoneme, /j/, in onset clusters. The word *new* is pronounced /nju:/ and its syllable structure is shown in Figure 3.11.

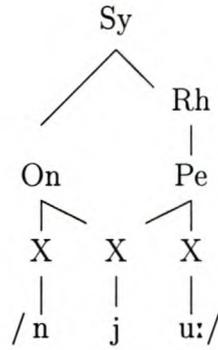


Figure 3.11: Syllable structure for the word, *new*, as suggested by [3].

/nj/ as an onset cluster clearly does not conform to the sonority theory since it is preceded by an already [+sonorant] phoneme. Linguistic theory suggests that the only way to treat this cluster is to consider /j/ to be part of both the onset and the peak. This only happens when the preceding phoneme is [+coronal][+sonorant]. However, this is not practical in a speech recognition system. It is more accurate to treat the /ju:/ as a single diphthong. (This then is also how the word, *few*, is treated as a real example in the database we use in our study). The /j/ in this case is quite different perceptually from the /j/ as it occurs in a word like *you* (/ju:/). If words such as *new* and *few* are transcribed as linguistic theory suggests, it will confuse the training of our models.

We see that a statement of the possible syllable onsets of a language (and, as we shall see, a statement of possible rhymes), consists of a positive template amended by a set of negative filters [3].

### 3.6.2 Rhyme phonotactics

#### Appendices

The words, *next*, (/nekst/) and *texts* (/tɛksts/), are examples where the sonority principle is again violated. The class of consonants that can follow the core syllable can be defined in terms of distinctive features: such consonants must have the feature composition [-sonorant] [+coronal]. To accommodate these cases, we allow for the core rhyme plus further X-positions, which must contain coronal obstruents, and which are referred to as the "appendix". Figure 3.12 illustrates the syllable parse tree for the word, *next*.

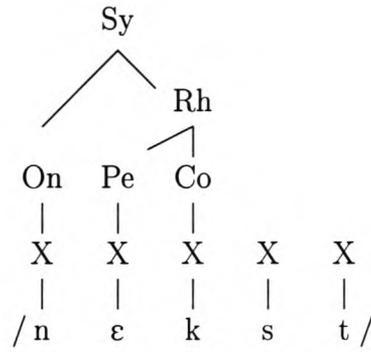


Figure 3.12: Syllable structure for the word, *next*.

### 3.6.3 Peak phonotactics

In stressed monosyllables, the only segments that can occur in the peak are vowels. When the syllable is unstressed, however, the constraints are relaxed as expressed in Figure 3.10, and we may find elements that are merely [+sonorant] as in the second syllable of the word, *button* (/bʌtn/). For unstressed syllables, a single phoneme as a syllable is allowed, as is the case with the first syllable in the bisyllabic word, *about* (/əbat/), where the schwa forms a syllable by itself.

## 3.7 Syllable definition used in this study

Now that we have reviewed the common principles and issues of syllable structure, we can define the model we will use in this study. We base our model of the syllable on the commonly accepted perceptual model shown in Figure 3.13 [3].

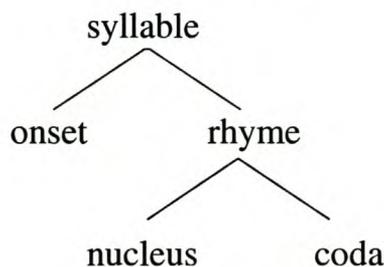


Figure 3.13: The structure of monosyllabic words (repeated here).

As we have seen, the structure in Figure 3.13, when applied to the English language, can be represented as:

$$C_0^3VC_0^3 \quad (3.1)$$

where  $C_0^n$  signifies 0 to 3 consonants and  $V$  signifies a vowel. Employing the phonotactic constraints that apply specifically to English language syllables allow us to further specify it as

$$/s/C_uC_vVC_vC_{u+v+s}C_{u+s} \quad (3.2)$$

where the members of each group are shown in Table 3.3.

Phonotactic constraints for the English syllable specify that when the onset is three consonants long, the first consonant can only be an /s/. According to sonority theory, there must be a rising sonority curve in the onset leading up to the nucleus. A further constraint is that, by referring to the list of binary features in Table 3.4, the second consonant must be [-sonorant] and the third [+sonorant] [3]. Therefore, the /s/ in the onset is followed by the unvoiced consonants,  $C_u$ , and then the voiced consonants,  $C_v$ .

A syllable must always have at least a nucleus,  $V$ , which we define as all vowels, diphthongs and the schwa. Syllabic consonants are treated as /ə/+C.

Group	Phonemes
$V$	/a/ /e/ /i/ /o/ /u/ /aɪ/ /ɛ:/ /ɛ/ /ɔ:/ /ɔɪ/ /ø/ /ə/ /ə:/ /æ/ /œ/ /œ:/ /eɪ/ /ui/ /iu:/ /əv/ /œy/ /aʊ/ /ɑ/
$C_u$	/b/ /d/ /f/ /g/ /k/ /p/ /t/ /x/ /θ/ /tʰ/ /ɰ/ /ʃ/ /tʃʰ/ /ʒ/ /r/
$C_v$	/h/ /j/ /l/ /m/ /n/ /r/ /v/ /w/ /z/ /ð/ /ʒ/ /ŋ/ /R/
$S$	/s/
$C_{u+v+s}$	$C_u \cup C_v \cup S$
$C_{u+s}$	$C_u \cup S$

Table 3.3: The syllable classes used in our syllable model.

In the coda we must conform to decreasing sonority. However, the sonority generalisation fails to account for one specific class of possible English codas: those with clusters like /sp/ and /sk/ as present in words, like *lisp* and *disk*. We, therefore, include /s/ in the second to last coda position,  $C_{u+v+s}$ .

<b>Group</b>	<b>Description</b>	<b>Features</b>
<i>V</i>	vowels and diphthongs	[+ <i>syllabic</i> ] [+ <i>sonorant</i> ]
<i>C<sub>v</sub></i>	voiced consonants	[- <i>syllabic</i> ] [+ <i>sonorant</i> ]
<i>C<sub>u</sub></i>	unvoiced consonants	[- <i>syllabic</i> ] [- <i>sonorant</i> ]
<i>S</i>	/S/	[- <i>syllabic</i> ] [- <i>sonorant</i> ]

Table 3.4: Binary features for syllable classes (slightly adapted from Prinsloo’s work and repeated here) [4].

### 3.8 Discussion

Our syllable definition can be applied in defining a regular grammar [16] for the classes in Tables 3.3 and 3.4 as was described by Prinsloo in [4]. This regular grammar has an exact non-deterministic Finite Automaton equivalent which can be implemented as an HMM.

These phonotactic constraints can be transformed to a regular phrase structure grammar, which could be employed to parse consonants and syllables in the acoustic speech signal (while taking account of the acoustic differences between syllable onset and coda consonants [4]). However, this direct parsing approach would require acoustical features related to manner and place of articulation, which is difficult to obtain. We use the HMM to represent syllable classes as HMM states and deduce interstate connections from the phonotactic constraints.

This implementation of our syllable model is discussed in the next chapter.

# Chapter 4

## Language Model

This chapter will build on the definition of the syllable in the previous chapter and will produce a language model which we can use to parse syllables.

### 4.1 Introduction

In the previous chapter we described our definition of the syllable. This definition should now be implemented as a model that can be used to parse and syllabify English syllables. As seen in the last section of Chapter 3, we simplified this model into a computationally tractable set of rules that will cover the vast majority of cases. In this chapter, we will first discuss grammar theory and how to specify a grammar to describe our syllable model. We will then see how this grammar can be implemented with a Finite State Automaton (FSA). A hidden Markov model can be used to fully describe an FSA. Using HMMs in a hierarchical fashion, we are able to build in levels of complexity to allow for the underlying acoustic model. These hierarchical HMMs will be discussed in the last section.

### 4.2 Regular Expressions, Regular Languages and Finite Automata

A regular expression is an algebraic notation for characterizing a set of strings. It can be used to specify search strings as is used in the Unix operating system's **grep** utility and various word processing packages. Regular expressions can also be used to define a language in a formal way.

Any regular expression can be implemented as a finite-state automaton (except for regular

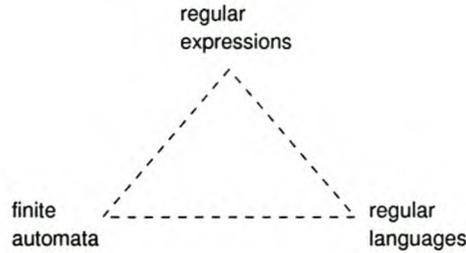


Figure 4.1: The relationship between finite automata, regular expressions, and regular languages (from [20]).

expressions that use the memory feature). Symmetrically any finite-state automaton can be described with a regular expression. A regular expression is one way of characterizing a particular kind of formal language, called a regular language. Both regular expressions and finite automata can be used to describe regular languages. The class of languages that is definable by regular expressions is exactly the same as the class of languages that is characterized by finite-state automata (whether deterministic or non-deterministic). This relationship between regular expressions, regular languages and finite automata is depicted in Figure 4.1.

A model, which can both generate and recognize all and only the strings of a formal language acts as a definition of the formal language. There are four major languages with their associated grammars in formal language theory and they are hierarchically structured as in Table 4.1.

Types	Constraints	Automata
Phrase structure grammar	$\alpha \rightarrow \beta$ . This is the most general grammar.	Turing machine
Context-sensitive grammar	A subset of the phrase structure grammar. $ \alpha  \leq  \beta $ , where $ \cdot $ indicates the length of the string.	Linear bounded automata
Context-free grammar (CFG)	A subset of the context sensitive grammar. The production rule is $A \rightarrow \beta$ , where $A$ is a non-terminal. This production rule is shown to be equivalent to Chomsky normal form: $A \rightarrow w$ and $A \rightarrow BC$ , where $w$ is a terminal and $B, C$ are non-terminals.	Push down automata
Regular grammar	A subset of the CFG. The production rule is expressed as: $A \rightarrow w$ and $A \rightarrow wB$ .	Finite-state automata

Table 4.1: Chomsky hierarchy and the corresponding machine that accepts the language (from [21]).

While it seems likely that we can't model all of English syntax with a finite-state grammar, it is possible to build an FSA that approximates English (for example by expanding only a certain number of NPs). There are algorithms for automatically generating finite-state grammars that approximate context-free grammars [20, p. 349].

Context-free grammars are the backbone of many models of syntax of natural languages (and, for that matter, of computer languages) [20, p. 324]. Context-free grammars are more powerful than finite-state automata, but it is nonetheless possible to approximate a context-free grammar with an FSA.

### 4.3 A Syllable Grammar

The basic rules of the sonority hierarchy can be applied in defining a regular grammar for the classes in Table 3.3. A finite automaton can then be used to parse these classes and syllables [4].

According to Chomsky's formal language theory, a grammar is defined as

$$G = (V_N, V_T, P, S) \quad (4.1)$$

where  $V_N$  and  $V_T$  are finite sets of non-terminals and terminals, respectively.  $P$  is a set of production or rewriting rules and  $S$  is the "syllable" as head of the language and start variable. The language to be analyzed is essentially a string of terminal symbols, (such as "Mary loves that person") (from [21]). It is produced by applying production rules sequentially to the start symbol. The production rule is in the form  $\alpha \rightarrow \beta$ , where  $\alpha$  and  $\beta$  are arbitrary strings of grammar symbols  $V_T$  and  $V_T$ , and the  $\alpha$  must not be empty.

The grammar for our model of the syllable can be described in terms of the phoneme classes and the rules of sonority described in the previous chapter. We use the same classes as Prinsloo in [4] who defined it as follows:

$$V_N = \begin{cases} S = Syllable \\ O = Onset \\ P = Peak \\ C = Coda \\ E = End \end{cases} \quad (4.2)$$

where  $V_N$  is the set of non-terminals and,

$$V_T = \begin{cases} u_o = \text{Unvoiced consonants (onset)} \\ v_o = \text{Voiced consonants (onset)} \\ n = \text{Nucleus} \\ v_c = \text{Voiced consonants (coda)} \\ u_c = \text{Unvoiced consonants (coda)} \end{cases} \quad (4.3)$$

$V_T$  is the set of terminal symbols. Prinsloo then formulated a set of production rules as shown below:

$$\begin{aligned} S &\rightarrow n & P &\rightarrow n \\ S &\rightarrow nC & P &\rightarrow nC \\ S &\rightarrow u_oP & C &\rightarrow u_c \\ S &\rightarrow v_oP & C &\rightarrow v_c \\ S &\rightarrow u_oO & C &\rightarrow v_cE \\ O &\rightarrow v_oP & E &\rightarrow u_c \end{aligned} \quad (4.4)$$

This is a right regular grammar which fully describes the syllable grammar in terms of phoneme classes. Since we have seen that a regular grammar can be represented by an FSA, we are able to define a syllable grammar as shown in the FSA in Figure 4.2.

The FSA is represented as a directed graph: a finite set of vertices (also called nodes), together with a set of directed links between pairs of vertices called arcs. The vertices are represented as circles and the arcs with arrows. Our FSA has five states, which are represented as nodes in the graph. State "S" is the start state, indicated by the incoming arrow. There are three terminal states, P, C and E, represented by double circles. The transitions are represented by arcs in the graph.

This non-deterministic finite state automaton can now be used to recognize syllables with phoneme classes as inputs. It is non-deterministic since it sometimes has to make a choice between multiple paths, given the same current state and next input.

In a practical speech model, however, the terminal symbols,  $V_T$ , are not directly observable from the acoustic signal. We, therefore, use a hidden Markov model, which can be viewed as an extension of the FSA [4]. The following section implements our model as an HMM able to parse syllables in the acoustic speech signal.

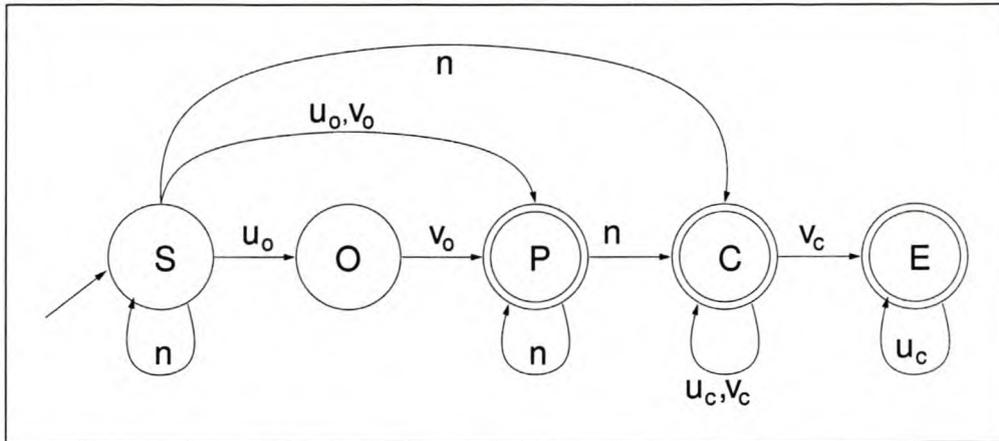


Figure 4.2: State diagram of a finite automaton for the syllable grammar presented in section 4.3 (from Prinsloo [4]).

## 4.4 HMM as model for an FSA

We expand the FSA when represented as an HMM in order to take advantage of specific phonotactic constraints for English. We have seen, for example, in Chapter 3 that when an onset is three characters long, the first must be an /s/.

In order to account for coda phonotactics, we have to adapt our model in order to allow for most commonly found exceptions to the sonority theory. Since our aim is to construct a generic model and "train in" these rules, we did not extend our study to define an exact grammar for these phonotactics, even though this would have been possible by using the information in Chapter 3 on syllable construct exceptions. Another motivation is that, as we will see in the next chapter, 85% of syllables in the database used are in the simple V, CV, VC and CVC forms, making detailed models of onset and coda behaviour unnecessary.

Rather, we created a generic model that will be able to parse the majority of cases. The fact that the model is generic will also enable us to easily modify it for use on other languages, this being one of the objectives of this study.

Figure 4.3 shows our HMM representation of the syllable model.

It has one position for the peak, which takes all vowels and diphthongs. As is clear from the transitions, the "V"-state is obligatory in order for the model to parse. Each syllable must have one, and only one, phoneme of this class. This implies that our model will exclude unstressed syllables where the peak might only be indicated by a syllabic consonant as in *button*. Again, as we will see in Chapter 5, the vast majority of words in

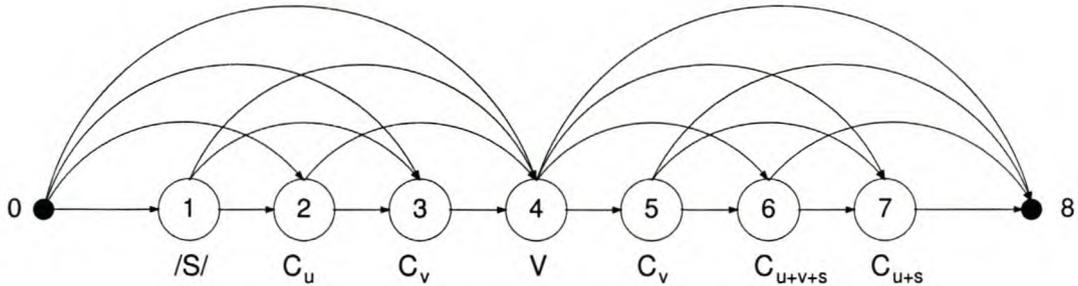


Figure 4.3: The syllable HMM model used in this study. The HMM states are defined in table 3.3.

the database is monosyllabic, which implies that there are very few cases where this will be a problem.

The onset states are straight forward and correspond to the states previously described in our syllable definition.

The coda states are more complex and some trade-offs had to be made. We wanted to keep our model simple, yet able to parse the majority of coda clusters. To accomplish this, we implemented the basic states in decreasing sonority. To allow for words such as *list* where the */s/* is not the final consonant, we added */s/* to State 6. State 6 is also a union of both voiced and unvoiced consonants. This allows for a combination of States 5 and 6 to parse two voiced consonants in the coda. States 6 and 7 will parse two unvoiced consonants in the coda, as in the case of the word, *sticks*.

## 4.5 Hierarchical HMM

The notion of hierarchical HMMs is a relatively simple one, which is not often attempted due to the complexity of implementing it in software. Existing experimental packages such as HTK [22] simply do not provide this feature. It is quite a challenge to implement a software system using this feature from scratch. Most researchers, therefore, opt for different approaches and easier-to-implement HMM structures for use in pattern recognition problems. Du Preez et al, however, did extensive development on the PatrecII software toolkit to allow complex structures such as these [7]. PatrecII is a collection of pattern recognition toolkits developed by the Digital Signal Processing Group of the University of Stellenbosch over a number of years. It is written in C++ and is highly adaptable, allowing one to easily connect together modules, which allows complex experiments. The hierarchical way of using HMMs, a simple concept but difficult to implement in software,

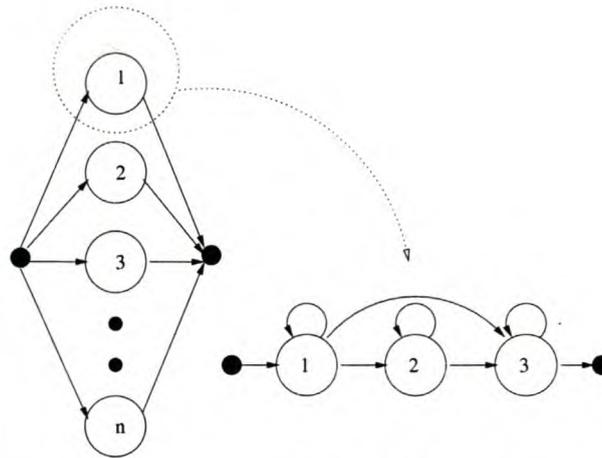


Figure 4.4: An example of a two-level hierarchical HMM where the upper level is a simple  $n$ -state parallel HMM model, and the lower level consists of a collection of phoneme HMMs which constitutes the parallel states.

was made possible entirely by the features of PatrecII.

Figure 4.4 is a graphical representation of a hierarchical structure. The model on the left is a simple parallel HMM model with  $n$  parallel states. The  $n$  parallel states together represent a collection of phonemes to form a class as described in Table 3.3. They are built as a parallel combination of their constituent phone models. The phone models in this example interact directly with the feature vectors obtained from the acoustical signal.

We construct a hierarchical model like this by first creating a template. In the example in Figure 4.4, a skeleton parallel HMM will be constructed in PatrecII. Separately, we would have trained  $n$  phoneme models on the data. Instead of populating the states of the parallel model with PDFs as a subsequent step, we "plug" into the parallel model states, the collection of phoneme HMMs that we have already trained. The input into state  $n$  of the parallel model now simply becomes the input to the phoneme model that is used in the place of state  $n$ . Similarly the output of the phoneme model in state  $n$  connects to the output of the original state.

Let us say we grouped all vowels together in a parallel HMM model. We can now train this high-level parallel HMM model by using the transcriptions for all vowel transcriptions. The phoneme models themselves are held constant and will not be trained. This will give us the transition probabilities to all the parallel states from the input node, without the phoneme models being affected on the lower hierarchy.

When we test the hierarchical model, we can essentially view it as one big model. The fact that it has different layers is transparent to us.

## 4.6 Discussion

We showed in this chapter how a regular grammar can describe our syllable model and that an HMM can fully model such a grammar. We introduced the relatively simple, although rarely attempted, concept of the hierarchical hidden Markov model.

In Chapter 6, we will use this model of the syllable to construct a segmenter to automatically tag syllable boundaries on a database, using only the acoustical features as input. First we have to discuss the specifics of the database used in the experiment in the next chapter.

# Chapter 5

## Database

We chose the SUNSpeech corpus for this experiment primarily because it was one of the few databases at our disposal that had a large amount of speech data with syllable-level transcriptions. This chapter will discuss the database, the composition of the data, and the actions that were necessary to clean up the inevitable mistakes in transcriptions. It will also compare the data with that used in previous studies.

### 5.1 Introduction

A suitable database for this study needed syllable-level transcriptions. The SUNSpeech database was developed by the Digital Signal Processing Department of the University of Stellenbosch in the early 1990s [23]. It contains both South African English and Afrikaans utterances and was hand-transcribed by two trained linguists on the phone, word, and syllable level. As Howitt mentions in his own study, the problem with phonetic transcriptions is that they are time-consuming to generate. A more serious problem is that phonetic transcriptions are often not unique or unambiguous. A phonetic transcription imposes categorical decisions on acoustic information that varies across a continuum. How these decisions should be made is far from clear. Also, phonetic transcriptions are vulnerable to errors, because there is no way to check for consistency [12]. We encountered these same problems in SUNSpeech, and they will be discussed in this chapter. In the process of trying to clean up the database, we did find that our preliminary syllable model was very useful when employed as a sanity check on the syllable-level transcriptions.

## 5.2 SUNSpeech

The SUNSpeech corpus is a set of continuous, naturally spoken utterances in South African English and Afrikaans. The recognition experiments were performed on the English subset of the corpus. The English subset consists of 40 different sentences spoken by 97 different speakers with a total of 1942 utterances. All sentences are not spoken by all speakers. Data was sampled at 16000Hz and recorded in a noise-free environment. It was divided into a training set and test set with 1316 utterances by 66 speakers and 626 utterances by the remaining 31 speakers, respectively. The list of sentences used in the English subset appears in Table B.1 in Appendix B.5.

The fact that there are only 40 different sentences being used is a limitation in the size of the set of unique syllables. Using linguists to hand-transcribe a database has its drawbacks in that the syllable definitions used are still open to some interpretation. We found this to be the case in a significant number of transcriptions on SUNSpeech, where on closer inspection, the syllable boundaries were inconsistently allocated and were not conforming to the overall philosophy of transcription employed for labelling the database. This is particularly the case with unstressed syllables, where traditionally the most ambiguity in interpretation will be encountered.

We will first describe how we used a preliminary version of our model as a sanity check on the database transcriptions. We will then discuss the most common mistakes that were encountered and how they were treated.

## 5.3 Cleaning up the database

In order to be able to use a speech corpus, it is often necessary to first do a sanity check on the correctness of the data that will be used to train recognition models. Since the database is hand-labelled, there is always a percentage of the data which will not be correctly labelled due to human error.

We took a cue from Prinsloo's work and used binary phoneme class features as input to the syllable model designed in the previous chapter. Prinsloo only used the phoneme classes [*sonorant*] and [*syllabic*] as inputs to his HMM. We decided to provide more features as input to our syllable HMM model and to ignore the actual acoustic input at this stage. By referring back to Figures 3.9 and 3.10, we see that the phoneme classes provide sufficient information in terms of general sonority theory, so that we can characterise syllable structure.

When we considered the full set of fifteen features defined in Appendix B.3, we decided to simplify the set by ignoring the following ones:

feature	reason
[lateral]	this is only applicable with /l/
[nasal]	only applicable with /n/, /m/ and /ŋ/
[round]	only differentiates specific phonemes and is not that important in terms of the broad syllable classes
[high]	
[low]	
[back]	
[tense]	this is only applicable to vowels and diphthongs

Table 5.1: The binary features not used in the discrete features experiment.

This left the following eight features that were used as input for our syllable model:

feature	reason
[sonorant]	indicates phonemes (vowels, glides, nasals and liquids) made up by the sound waves associated with voicing
[voiced]	produced with glottal setting consistent with vocal-fold vibration
[syllabic]	all segments that can act as the nucleus
[consonantal]	produced with a radical obstruction in the vocal tract
[continuant]	air stream is not blocked in the oral cavity during production
[anterior]	contains information on the location of the obstruction, thus the type of consonant which may be important in determining on-set/coda
[coronal]	contains information on coda constraints as explained in section 3.6.2
[strident]	contains information about the noisiness of a sound segment

Table 5.2: The binary features used in the discrete features experiment.

We created a mapping file with all the information contained in Appendix B.3 and used a PERL script to create binary feature files from the SUNSpeech transcriptions. Using the transcription files for each utterance, we created a corresponding feature file containing a combination of eight binary features representing each phoneme in the utterance. A "1" indicates that the feature is active ( $[+feature]$ ) and a "0" indicates that the feature is not active for that phoneme ( $[-feature]$ ). A "-1" indicates that the feature is not applicable for that phoneme.

We used discrete time segments in this new feature file; therefore each phoneme in the transcription is deemed to last one "unit" of time in the features file. All temporal information was essentially ignored. An example of a transcription file and the corresponding binary features set is shown in Table 5.3. A full explanation of the phoneme set used on SUNSpeech is provided in Appendix B.

Our model now simply acted as a pure FSA that recognised strings of input without any temporal information playing a role.

We used a slightly simpler model of the syllable for this experiment than the one described in section 3.7. We merely wanted to confirm the validity of the labelling on the database, as well as investigate the use of binary features as input to our syllable HMM. Our model was, therefore, the more generic version which merely allows three onsets (the first still being only /s/), a vowel, and then three coda positions of decreasing sonority.

$$/s/C_u C_v V C_v C_u C_u \quad (5.1)$$

We added an extra sub-syllable label on the transcriptions as shown in Table 5.3. This defined each phoneme as being part of either the set of voiced consonants (VC), unvoiced consonants (UC), vowels and diphthongs (V), and /s/ (S).

These transcriptions and features were then used to train four different discrete PDFs corresponding to the classes V, UC, VC, and S and which would be used in the states of our syllable HMM.

We initialised each PDF with an a priori weighting using actual counts from the transcriptions on the number of times each phoneme occurs in the database. We used a fixed circular gaussian density function with a single variance which we pinned on a very small value (0.01 in our this specific case).

These PDFs were then used in the seven states of the syllable HMM. The HMM was trained on the entire database. We did not use separate training and test sets in this experiment, since the objective was to use the model as a tool to highlight possible errors in transcriptions, before we proceeded to proper training and testing.

We then used this HMM to run through the entire database and score each labelled syllable. We have seen that the definition of a syllable is open to a fair amount of interpretation and there are different schools of thought on syllable definitions. The various inconsistencies in syllable labels on the database now showed up. Some of the syllables in the corpus simply would not be parsed by our syllable model and showed up as severely negative scores. Many of the scores were exactly the same, i.e. there was a number of numerically identical scores showing repeat problems found with syllables mislabelled in a similar way. When we consider that there are only 40 different sentences being used, it was clear that some syllables were consistently labelled incorrectly for all speakers.

Time	Phon	SubSyll	Syll	Word	Discrete features
0.030000	42	Sil	-	-	0 0 0 0 0 0 0
0.154000	126	V	Syllable	i	1 1 1 0 1 -1 -1 -1
0.203000	102	UC	03nonSyllable	have	0 0 0 1 -1 -1 -1 -1
0.279000	100	UC	-	-	0 1 0 1 0 1 1 0
0.305000	143	V	Syllable	-	1 1 1 0 1 -1 -1 -1
0.440000	115	S	-	-	0 0 0 1 1 1 1 1
0.591000	126	V	Syllable	-	:
0.635000	218	UC	-	-	:
0.680000	143	V	-	-	
0.730000	100	UC	Syllable	decided	
0.829000	116	UC	02nonSyllable	to	
0.923000	112	UC	-	-	
0.957000	149	V	Syllable	-	
1.056000	109	VC	-	-	
1.115000	149	V	-	-	
1.217000	116	UC	Syllable	permit	
1.260000	143	V	Syllable	the	
1.352000	112	UC	-	-	
1.450000	97	V	-	-	
1.527000	98	UC	Syllable	-	
1.555000	108	VC	-	-	
1.615000	151	V	-	-	
1.697000	107	UC	Syllable	public	
1.734000	106	VC	-	-	
1.791000	131	V	-	-	
1.849000	116	UC	Syllable	yet	
1.888000	143	V	-	-	
1.937000	110	VC	-	-	
2.030000	97	V	22nonSyllable	-	
2.070000	172	VC	-	-	
2.139000	131	V	Syllable	another	
2.223000	103	UC	-	-	
2.256000	108	VC	-	-	
2.302000	143	V	-	-	
2.381000	109	VC	-	-	
2.407000	112	UC	-	-	
2.490000	115	S	Syllable	glimpse	
2.569000	110	VC	01nonSyllable	-	
2.647000	116	UC	02nonSyllable	into	
2.712000	109	VC	-	-	
2.810000	97	V	Syllable	my	
2.911000	97	V	-	-	
2.960000	205	VC	Syllable	-	
3.049000	107	UC	-	-	
3.131000	149	V	Syllable	uncle	
3.279000	133	V	-	-	
3.359000	115	S	Syllable	-	
3.433000	119	VC	-	-	
3.588000	150	V	-	-	
3.717000	181	UC	Syllable	oswalds	
3.786000	108	VC	-	-	
3.951000	126	V	-	-	1 1 1 0 1 -1 -1 -1
4.063000	102	UC	Syllable	life	0 0 0 1 -1 -1 -1 -1

Table 5.3: An example of a transcription file and the corresponding binary feature file for utterance HJM10051.

Starting with the worst scores first, we inspected these syllables directly in the database and we were able to identify whether it was mislabelled, or whether it was an error in our model. The mislabelled syllables were marked in the transcriptions and the model ignored them and their corresponding acoustical features from that point onwards. We used the tag, *Syllable*, to indicate a syllable boundary in the transcriptions, and *XXnonSyllable* for these errors. *XX* is a number corresponding to a specific group of similar errors.

This process was continued up to a certain threshold in error score, when we could no longer see for certain that the problem was due to errors in labelling. As a result, most of the errors in labelling were removed and the database could be considered representative enough of usable data for us to proceed. This was an extremely labour intensive task with some 2000 different utterances having to be inspected and corrected by hand. It did show that binary features could be used as input to an HMM acting as a syllable parser. This confirmed, amongst others, Prinsloo's work.

## 5.4 Errors found

We encountered several inconsistencies in the syllable labels. Single consonants were labelled as syllables, mostly where these consonants are missing the label for a preceding schwa. Many syllables were transcribed containing two vowels. Some examples include:

- single consonants /n/, /v/, /t/, /d/, /ʒ/, /f/, /r/, /k/ labelled as syllables. In Table 5.3 some examples of this are labelled as *01nonSyllable* (single /n/), *02nonSyllable* (single /t/), and *03nonSyllable* (single /f/).
- words, like *reputation*, where the last syllable is transcribed as /ʃn/ which does not contain the implicit schwa. This is common with unstressed syllables.
- single /tʃ<sup>h</sup>/ /ts<sup>h</sup>/ /dʒ/
- *about*, labelled as single syllable /əbat/; therefore containing two nuclei
- *for*, labelled as /fr/; therefore missing a nucleus
- *thousands*, where the last syllable is labelled as /ʒnʒ/ missing a nucleus
- *evident*, split into two syllables, where the first is /ɛvə/, again containing two nuclei
- *another*, labelled as /ənə.ðɛ/. In Figure 5.3 this is labelled as *22nonSyllable*.

These mislabelled syllables were marked by hand in the transcriptions of both our training and test sets and ignored in all subsequent experiments. We tagged roughly 4% of the syllables in the database in this manner. All future training and testing ignored these segments, as well as their corresponding sections in the acoustic feature files for each utterance.

## 5.5 Data composition

The five simple syllable structures shown in Table 5.4 account for 94% of all syllables in the SUNSpeech database. This is similar to that reported for Switchboard by Wu in [13], where, as shown in Table 5.5, eight relatively simple structures also account for 84% of the syllables found in that corpus. Note that the transcriptions in Switchboard show diphthongs as double vowels (VV).

Structure type	% of corpus
V	10.43
VC	13.77
CV	36.25
CVC	28.47
CVCC	5.1

Table 5.4: The syllable structures found in the SUNSpeech database.

Structure type	% of corpus
CVV	21.19%
CVC	19.75%
CVVC	9.99%
CV	9.51%
VC	9.14%
VV	6.98%
CVCC	3.99%
VCC	3.85%

Table 5.5: The frequency of the eight most frequent syllable structures in the Switchboard corpus (from [13]).

Our use of a more generic model implies that we will not be able to account for all the phonotactic constraints and exceptions to these constraints, especially in longer coda clusters. When we consider the spread of syllable structures found in SUNSpeech, we see that we will, however, still be able to handle and parse the majority of syllables found in the database. The assumption is that this will also be the case in normal conversational speech, where the range of possible syllables is unbounded.

<b>N</b>	<b>% of vocabulary</b>	<b>% of corpus</b>
1	22.39%	81.04%
2	39.76%	14.30%
3	24.26%	3.50%
4	9.91%	0.96%
5	3.21%	0.18%
6	0.40%	0.021%
7	0.057%	0.0013%
8	0.0052%	0.000037%

Table 5.6: The frequency of words with  $N$  syllables in the Switchboard vocabulary and corpus (from [13]).

When we look at Table 5.6, we see that in Switchboard, 95% of all the words in that corpus do not contain more than two syllables. Again, if we assume that this is fairly representative of conversational speech, we see that it is not necessary for our model to include all the complexities of trying to model ambisyllabicity and complex coda cluster rules as described in Chapter 3.

When we compare the size of our dataset in Table 5.7 with that of previous studies, we see that it compares very well in terms of size and variety. Before Howitt, most previous studies used only enough data in order to prove a concept or algorithm. Only Howitt used a database of significant size against which we can compare our own results [12].

<b>Author</b>	<b>Dataset</b>
Mermelstein	2 male speakers, 11 sentences, 22 utterances, read at their comfortable reading rate. 418 syllables.
Prinsloo	10 male speakers, each with two excerpts of 7 seconds (2.5 minutes of speech), 20 utterances, 370 syllables
Howitt	male and female speakers of TIMIT whose numbers end in 8 or 9. Training set: 619 utterances and 7585 syllables, Test set: 373 utterances and 4404 syllables
<b>This study</b>	training and test sets with 1316 utterances by 66 speakers and 626 utterances by the remaining 31 speakers respectively, 14143 syllables in test set.

Table 5.7: A summary of the sizes and composition of the data used by prominent syllabification studies.

## 5.6 Discussion

Since SUNSpeech is not a widely used database, it is still difficult to compare our results to other studies. This was also a criticism when the results from this study were presented

at [5]. The most commonly used database in speech research is TIMIT; however, it does not contain syllable-level transcriptions. One way around this, as Howitt did in his study, is to use the **TSYLB2** program from NIST to add syllable boundaries in the existing transcriptions. This automatic labelling is, however, different from that done by trained linguists, who should be less error-prone. The benefit of using **TSYLB2**, which implements Kahn's theory of the syllable, is that it provides a repeatable result. This makes it easier to compare syllabification results. Howitt did, however, report that he too had to spend a significant amount of time to go through a similar process of cleaning up the transcriptions generated by **TSYLB2**.

# Chapter 6

## Acoustic syllable segmenter

In order for us to automatically syllabify a continuous speech stream, we need a recogniser that is able to create transcriptions with syllable boundaries. This recogniser should be able to handle and tag any non-syllable groupings such as silence and background noise. This chapter will discuss our implementation of an automatic syllable segmenter.

### 6.1 Introduction

In Chapter 5, our experiment with binary phoneme class features showed that, as Prinsloo proved in his article, these features are sufficient to provide an input to an accurate HMM model of a syllable grammar [4]. At this point in our research, we were faced with two possible avenues along which to continue our study and realise our goal of automatic syllabification from an untranscribed acoustic speech signal.

The first option was to create eight different recognisers for the eight phoneme classes used in the binary features experiment in Chapter 5. These recognisers, possibly in the form of neural networks, would provide the information from the acoustic signal and would act as input to the HMM. However, that would have simply been a repeat of Prinsloo's previous research, albeit on a much bigger set of data and with more phoneme classes as inputs.

Our second option was to use the acoustical features directly as input into a hierarchical set of HMMs. Our first hierarchical level would be normal phoneme recognisers, which are grouped into the phoneme classes according to their levels of sonority. These classes again constitute the seven high-level states of our original syllable model of section 3.7. We chose the latter route primarily to exploit and prove the concept of using hierarchies in HMMs.

## 6.2 Signal processing

As a first step in building our hierarchical model, we had to process the audio recordings on SUNSpeech and produce acoustical features for use in training and testing our HMM models.

We performed pre-emphasis and energy normalisation in a 100Hz - 7500Hz window on the data. It was then parameterized using 18 dimensional mel-frequency cepstral coefficients (MFCCs) with 22 filter banks [21, p. 317]. A frame length of 20ms with a frame skip of 10ms was used. Temporal information is of particular significance in syllabification and, therefore, the delta, and delta of delta between successive frames were computed. The dimensions of the result were reduced using linear discriminant analysis (LDA) [21, p. 427].

## 6.3 Phones

We trained phone models for the 56 distinct phonemes found in SUNSpeech, using both the training and test set. For the phonemes, we specifically did not follow the practice of splitting training and test sets. This was to eliminate the phoneme-level information and recognition as a factor in determining the success of our syllable model.

We used a simple left to right HMM structure with one state skip. Since we wanted the best possible input to our syllable model, the phone models were trained using the entire set of training and test data in order to minimise effects due to phone model inaccuracies. We achieved an accuracy of 53% for all 56 phones tested on the training and test set. Since we built our syllable model using groupings of these phones, this level of accuracy was deemed sufficient for our specific set of experiments.

## 6.4 DTW distance measure

A common problem with automatic syllabification algorithms is that their accuracy measures are often described inadequately [12]. We, therefore, provide the detail of our matching procedure here.

To align an automatically determined syllable labelling with its ideal hand-labelled version, a dynamic time warping (DTW) procedure is used to do the mapping between these two sequences in terms of correct labels, substitutions, insertions and deletions [21]. Our

token error rate (TER) is then defined as shown in equation 6.1.

$$TER = 100\% \times \frac{Subs + Dels + Ins}{\text{No. of syllables in the correct utterance}} \quad (6.1)$$

Two components play a role here, namely a) the relative costs of these various types of labelling errors, and b) the specific local cost describing how dissimilar a particular label is compared to another.

We give a small but equal weighting to DTW paths corresponding to substitutions, insertions and deletions (the specific weight was 0.1). Since substitution errors result in a shorter DTW path length than the others, this weighting results in a slight preference for substitution errors compared to insertions and deletions.

Our label distance measure algorithm takes, as input, the acceptable time error in fixing the boundaries of the syllables. We call this  $\epsilon$  and used 20ms as our acceptable error margin. Referring to Figure 6.1, we then define the overlap between the original syllable transcription and our generated syllable boundaries.

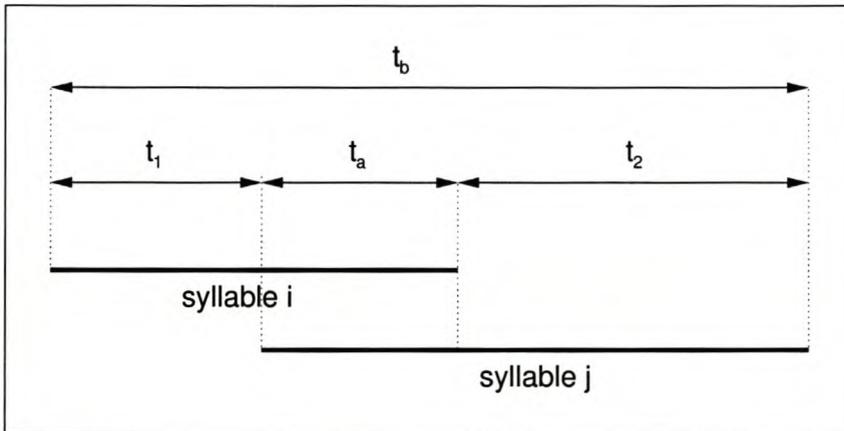


Figure 6.1: The overlap between two syllables,  $i$  and  $j$ , where one will be the original transcription and the other generated by our segmenter model from the acoustical data.

**Definitions:**

$$overlap = \frac{t_a}{t_b} \quad (6.2)$$

$$\epsilon = \text{acceptable boundary error} \quad (6.3)$$

**Step 1:**

$$D(i, j) = \begin{cases} 1 - overlap & \text{if } t_1 \text{ and } t_2 < \epsilon, \\ 3 - overlap & \text{if } t_1 \text{ or } t_2 < \epsilon, \\ 5 - overlap & \text{if } overlap > 0 \\ & \text{and } t_1 \text{ and } t_2 > \epsilon, \\ 10 & \text{if } overlap < 0. \end{cases} \quad (6.4)$$

**Step 2:**

$$D(i, j) = \begin{cases} D(i, j) + 5 & \text{if ids mismatch} \end{cases} \quad (6.5)$$

Our distance measure is based on the amount of overlap and whether the generated syllable falls within the accepted boundary error compared to the original transcription. In Step 1 we progressively penalize errors according to the degree of mismatch. The *overlap* is a positive number, when there is a degree of overlap between the syllable boundaries of the original transcription and that generated by our model, as shown in the example of Figure 6.1. However, it also happens that there is no overlap between syllables that are supposed to match. This then results in a negative number for  $t_a$  and, therefore, *overlap*.  $t_a$  is defined as the end time of syllable  $i$  minus the start time of syllable  $j$  and it will, therefore, become negative when there is no overlap between two syllables.

In Step 2, we add an extra penalty to the existing distance measure, should the IDs not match (i.e. a *garbage* tag is being aligned with a *Syllable* tag).

The actual values used in Steps 1 and 2 were determined mostly empirically by studying a large number of DTW transcription alignments, until we eventually achieved the desired behaviour.

## 6.5 HMM

Since the phenomena that we are modelling operates on a number of hierarchical levels, we chose to use a four-level hierarchical HMM (HHMM) with which to represent the speech. The top level represents a speech recording as a combination of syllables and garbage segments, as is shown in Figure 6.2. When used to analyse speech, this level of the model

generates tags with *Syllable* and *garbage* as labels.

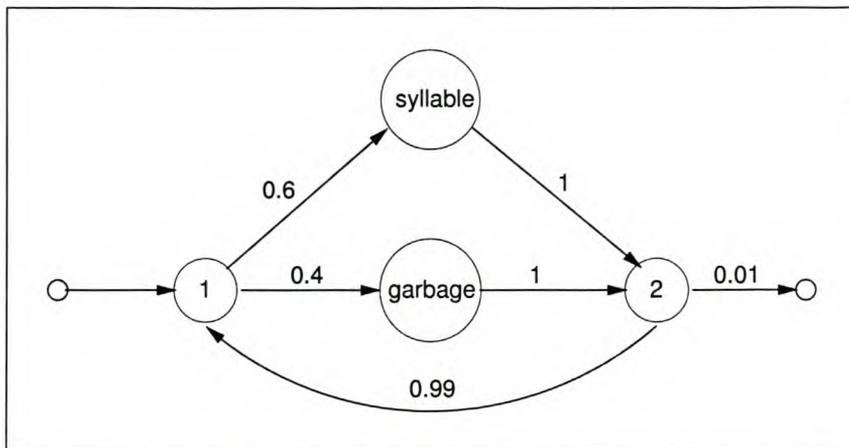


Figure 6.2: The syllable segmenter model used to generate syllable-level transcriptions from the acoustical data.

The input transition probabilities between the *Syllable* and *garbage* states in Figure 6.2 are based on a very conservative a priori count of the amount of correctly labelled syllables in the SUNSpeech set, compared to "silence" and *garbage* tags. (For this experiment, we replaced the *XXnonSyllable* labels in the transcriptions with *garbage*). The actual split between *garbage* and *Syllable* tags is closer to 85/15 %.

The syllable state from Figure 6.2 expands to the second level model as is shown in Figure 6.3. This implements an FSA of the syllable definition described in section 3.

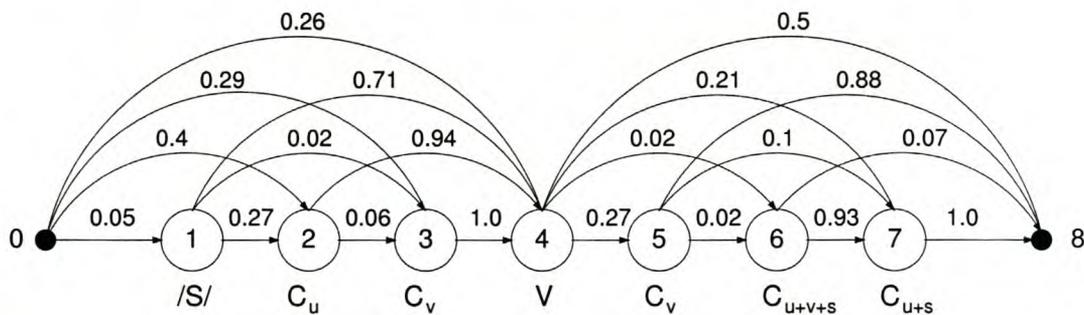


Figure 6.3: The trained syllable model. Transition probabilities as trained are shown and rounded to two decimals.

Similarly the garbage state of 6.2 expands to a 6-state ergodic HMM model on this second level. This garbage model is built using the */S/*,  $C_u$ ,  $C_v$ ,  $V$  classes together with a model for "silence" and one for the "unknown" tag, arranged in a fully connected configuration. (The "unknown" tag found in Sunspeech is a small collection of phonemes, which was not labelled by the transcribers.)

As shown in Figure 6.4, the third level in the hierarchy models each of the class groups described in Table 3.3. They are built as a parallel combination of their constituent phone models, which, in their turn, form the fourth and bottom-most level in the hierarchy. This level directly interacts with the MFCC feature vectors obtained from the acoustical signal.

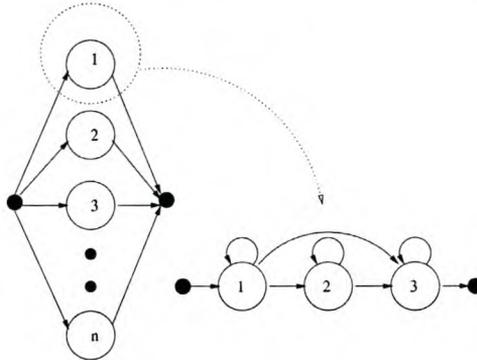


Figure 6.4: The parallel HMM model used to represent the phoneme classes.

These phoneme HMM models are trained separately and then integrated into the HHMM. After this integration, their parameters are kept frozen with further training impacting only on the higher levels of the HHMM. Specifically, the entire syllable HHMM model was trained using the time-aligned syllable markings available for the training set. The resulting transition probabilities in the trained syllable model are shown in Figure 6.3.

## 6.6 Syllabification Results

We tested our segmenter against the SUNSpeech test set. Table 6.1 summarizes the results achieved by our automatic syllabification system.

tokens	14143
deletions	12.7%
insertions	5.7%
substitutions	2%
correct	85.4%
accuracy	79.7%
<b>TER</b>	<b>20.3%</b>
avg boundary err	42ms
std dev	36ms
max err	406ms

Table 6.1: The syllabification results achieved by our model on the SUNSpeech test set.

Our token error rate of 20.3% compares well with results obtained by Howitt [12] and Wu [13] on TIMIT and OGI Numbers95 respectively.

Our model is, of course, very accurate in detecting syllable nuclei since every vowel implies a syllable peak. As expected, the syllable boundaries are less consistently detected. However, we achieved an average boundary error of 42ms. With an average English syllable length of 250ms [13], this can be considered fairly accurate. Most of the syllabification errors occurred in unstressed syllables and this supports the phonological expectations. Interestingly, our model did syllabify many of these unstressed syllables correctly. An example is the word, *reputation*, discussed previously, where, even though the last syllable is labelled /fn/ in the transcriptions, the model identified the implicit schwa phoneme. This allowed our model to parse and identify that syllable correctly. Since our schwa phoneme is trained on the entire database, it is further justification for our suggestion that there should have been a schwa phoneme inserted in the transcriptions in the first place.

Our maximum error of 406ms does imply that, in at least one case the model encountered a rather long string, which it could not parse at all. Since our model implements only the basic phonotactic constraints, it is inevitable that it will encounter some strings which it will not be able to parse. This trade-off in the design of our model was made in order to keep it simple enough, so that it could easily be trained and used for other languages, where the phonotactic constraints might be different.

When we consider the transition probabilities of the trained syllable model in Figure 6.3, we are able to make some interesting observations. We can see for instance that on the SUNSpeech database's test set, 50% of syllables do not have codas, 13% ( $0.26 \times 0.5$ ) of syllables consist of only single vowels or diphthongs  $V$ , 26% of syllables start with the  $V$  state and 5% of syllables start with /s/. This corresponds well with the data as summarised in Table 5.4

## 6.7 Discussion

We see that interesting observations can be made from studying the trained-in transition probabilities of our syllable model. By simply training a model like this on an entire database, we are able to analyse syllable structure empirically for that specific corpus. As such, it possibly can be used as a tool by linguists to inspect the syllable characteristics of a given set of data.

Our token error rate of 20.3% compares very well with previous studies, even those performed on significantly smaller sets of data. However, Howitt's experiments on TIMIT still remain the only really sensible results against which we can compare ourselves, due to the limited sizes of training data in other studies.

Our well-defined measure for accuracy solves a problem in syllabification research in that we define our results in a repeatable and exact way. This measure can be used and expanded upon in future studies, where it is necessary to report syllabification results.

## Chapter 7

# Conclusion and Directions for Future Research

### 7.1 Discussion

This study contributed to the field of syllabification research by presenting a purely acoustical syllable model that can be easily adapted and trained for different languages. As such, it could be used as a tool for linguists in sanity checks on databases' syllable transcriptions. It could also be employed to empirically investigate syllable structure of languages by training it on a database of sufficient size. For research tasks, such as text to speech and foreign accent identification, which rely on syllable-level information, it could be used to segment large amounts of speech data.

Our experiment, with binary features as raw input to a syllable HMM model, confirmed Prinsloo's work, and showed that this could be a simple tool to check syllable transcriptions. This presented us with the choice on whether to use binary feature classifiers on the acoustical data to provide input to the HMM, or to take a new direction and use a hierarchy of HMMs, build our syllable segmenter and for these HMMs to interact directly with the speech data. Our choice of the latter proved to be successful, and in the process we gained valuable experience in the use of the relatively novel concept of hierarchical hidden Markov models.

As an extension to the experiment described in Chapter 6, we used Du Preez's work on higher order hidden Markov models to see if this would improve our results. However, there was a negligible improvement in recognition accuracy. This is most likely due to data insufficiency and if we had to repeat these experiments today we would use tools such as Maximum a Posteriori estimation (MAP) to correct this.

## 7.2 Future Work

We used the SUNSpeech database because of its existing hand-labelled syllable-level transcriptions. Since this database is not very well known, it is difficult to compare results. The only other study that used a database of significant size was Howitt's, which used TIMIT. We intend repeating our experiment on the TIMIT database using Bill Fischer's **TSYLB2** program to generate syllable transcriptions. Fischer's program implements the syllable model defined for English by Kahn in [10]. By training our model on this data, we will essentially be able to create a statistical representation of Kahn's syllable model, as trained on TIMIT. A trained model like this can then easily be used as a diagnostic tool to indicate transcription errors on other similar English language databases.

We chose to apply our segmenter only to the English language. Future work might be focused on investigating the syllable structure of other languages. Much work has been done by others on languages such as Mandarin which has simple, mostly CV, syllable structure. Our model might allow a more syllable-centered approach to languages with more complex syllable structures.

## 7.3 Conclusion

We have applied the concept of hierarchical HMMs to model syllables. These statistical models are automatically inferred directly from acoustical speech data. It is, however, self-evident that the generalisation ability of these models is highly dependant on the specific training database being used. Evaluation showed the results to be fairly accurate and to compare well to knowledge-based approaches. The ability to observe the resultant regular grammars describing syllable structure also holds some benefit compared to neural-based approaches.

# Bibliography

- [1] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-23(1), pp. 82–87, February 1975.
- [2] S.-L. Wu, B. E. D. Kinsbury, N. Morgan, and S. Greenberg, "Incorporating information from syllable-length time scales into automatic speech recognition," in *ICASSP*, vol. 2, pp. 721–724, 1998.
- [3] H. J. Giegerich, *English Phonology - An introduction*. Cambridge University Press, 1992.
- [4] G. Prinsloo and M. Coetzer, "Automatic syllabification and phoneme class labelling with a phonologically based hidden markov model and adaptive acoustical features," *Computer Speech and Language*, vol. 4, pp. 247–262, 1990.
- [5] P. Nel and J. du Preez, "Automatic syllabification using hierarchical hidden markov models," in *ICASSP*, vol. 1, pp. 768–771, 2003.
- [6] G. Prinsloo, "Phoneme class recognition and automatic syllabification with a phonological based hidden markov model," Master's thesis, University of Stellenbosch, 1988.
- [7] J. du Preez et al, *PATRECII*. Digital Signal Processing Group, University of Stellenbosch, <http://dsp.ee.sun.ac.za>.
- [8] J. du Preez, *Efficient high-order hidden Markov modelling*. PhD thesis, University of Stellenbosch, March 1998.
- [9] K. Berkling, "Scope, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification," *Speech Communication*, no. 35, pp. 125–138, 2001.
- [10] D. Kahn, *Syllable-based Generalizations in English Phonology*. PhD thesis, Massachusetts Institute of Technology, September 1976.

- [11] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *Journal of the Acoustical Society of America*, vol. 58, pp. 880–883, October 1975.
- [12] A. W. Howitt, *Automatic Syllable Detection for Vowel Landmarks*. PhD thesis, Massachusetts Institute of Technology, July 2000.
- [13] S.-L. Wu, *Incorporating Information From Syllable-length Time Scales into Automatic Speech Recognition*. PhD thesis, University of California, Berkeley, 1998.
- [14] P. Nel and J. du Preez, "Automatic syllabification using hierarchical hidden markov models," in *IEEE Neural Networks for Signal Processing (NNSP)*, 2003.
- [15] A. W. Howitt, "Vowel landmark detection," *Eurospeech*, 1999.
- [16] N. Chomsky and M. Halle, *The Sound Pattern of English*. Harper and Row, Publishers, 1968.
- [17] G. N. Clements and S. J. Keyser, *CV Phonology*. The MIT Press, 1983.
- [18] B. Fischer, *TSYLB2*. Source code available from NIST at: [www.nist.gov/speech](http://www.nist.gov/speech), 1995.
- [19] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.
- [20] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2000.
- [21] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice Hall, 2001.
- [22] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The htk book," 1999.
- [23] Digital Signal Processing Group, University of Stellenbosch, <http://dsp.ee.sun.ac.za>, *SUNSpeech Corpus*.
- [24] J. Combrink and L. de Stadler, *Afrikaanse Fonologie*. Macmillan South Africa, 1987.

# Appendix A

## Glossary

The definitions below, unless otherwise indicated, are from the *Collins English Dictionary - Third Edition 1994*.

**allophone** *n.* any of several speech sounds that are regarded as contextual or environmental variants of the same phoneme. In English, the aspirated initial (p) in *pot* and the unaspirated (p) in *spot* are allophones of the phoneme /p/.

**anterior** *phonological feature* anterior sounds are produced with an obstruction that is located in front of the palato-alveolar region of the mouth; non-anterior sounds are produced without such an obstruction.[3, p. 116]

**aspire** *vb.* **a.** to articulate (a stop) with some force, so that breath escapes with audible friction as the stop is released. **b.** to pronounce (a word or syllable) with an initial *h*. *n.* **a.** a stop pronounced with an audible release of breath. **b.** the glottal fricative represented in English and several other languages as *h*.

**close** *adj.* denoting a vowel pronounced with the lips relatively close together.

**closed** *adj.* **a.** denoting a syllable that ends in a consonant. **b.** another word for **close** above.

**coda** the consonant or consonant group that follows the peak in a syllable.

**coronal** *phonological feature* coronal sounds are produced with the blade of the tongue raised above its neutral position; non-coronal sounds are produced with the blade of the tongue in the neutral position.[3, p. 116]

**demisyllable** defined as essentially half of a syllable that has been divided after the CV transition

**diphthong** *n.* a vowel sound, occupying a single syllable, during the articulation of which the tongue moves from one position to another, causing a continual change in vowel quality, as in the pronunciation of *a* in English *late*, during which the tongue moves from the position of (e) towards (ɪ). [C15: from Late Latin *diphthongus*, from Greek *diphthongos*, from DI- + *phthongos* sound]

**elision** *n.* the omission of a syllable or vowel at the beginning or end of a word, esp. when a word ending with a vowel is next to one beginning with a vowel.

**epenthesis** *n.* the insertion of a sound or letter into a word.

**flap** *n.* an (r) produced by allowing the tongue to give a single light tap against the alveolar ridge or uvula.

**foot** *n.* a group of two or more syllables in which one syllable has the major stress, forming the basic unit of poetic rhythm.

**fricative** *n.* a continuant consonant produced by partial occlusion of the air stream, such as (f) or (z).

**generative grammar** *n.* a description of a language in terms of explicit rules that ideally generate all and only the grammatical sentences of the language.

**glottis** *n., pl.* the vocal apparatus of the larynx, consisting of the two true vocal cords and the opening between them. [C16: from New Latin, from Greek *glōttis*, from *glōtta*, Attic form of Ionic *glōssa* tongue]

**glottal** *n.* articulated or pronounced at/or with the glottis.

**glottal stop** *n.* a plosive speech sound produced as the sudden onset of a vowel in several languages, such as German, by first tightly closing the glottis and then allowing the air pressure to build up in the trachea before opening the glottis, causing the air to escape with force.

**homorganic** *adj.* (of a consonant) articulated at the same point in the vocal tract as a consonant in a different class. Thus *ŋ* is the homorganic nasal of *k*.

**lax** *adj.* pronounced with little muscular effort and consequently having relatively imprecise accuracy of articulation and little temporal duration. In English, the vowel *i* in *bit* is lax. Compare **tense** [C14: from Latin *laxus* loose]

**morpheme** *n.* a speech element having a meaning or grammatical function that cannot be subdivided into further such elements.

**open** *adj.* **a.** denoting a vowel pronounced with the lips relatively wide apart. **b.** denoting a syllable that does not end in a consonant, as in *pa*. Compare **closed**.

**phonemics** *n.* that aspect of linguistics concerned with the classification, analysis, interrelation, and environmental changes of the phonemes of a language.

**phonology** *n.* the study of the sound system of a language or of languages in general.

**phoneme** *n.* one of a set of speech sounds in any given language that serve to distinguish one word from another. A phoneme may consist of several phonetically distinct articulations which are regarded as identical by native speakers since one articulation may be substituted for another without any change of meaning. Thus /p/ and /b/ are separate phonemes in English because they distinguish such words as *pet* and *bet*, whereas the light and dark /l/ sounds in *little* are not separate phonemes since they may be transposed without changing meaning. [C20: via French from Greek *phōnēma* sound, speech]

**phone** *n.* a single uncomplicated speech sound.

**phonetics** *n.* the science concerned with the study of speech processes, including the production, perception, and analysis of speech sounds from both an acoustic and a physiological point of view. This science, though capable of being applied to language studies, technically excludes linguistic considerations. "*The study of the full range of vocal sounds that human beings are capable of making is phonetics. The study of the sounds human beings employ when speaking a language is linguistic phonetics. Phonology is the study of the system underlying the selection and use of sounds in the languages of the world.*" (Kenstowicz and Kisseberth 1979).

There are three branches of phonetics each of which approaches the subject somewhat differently (from [19, p. 115]) :

1. *Articulatory phonetics* is concerned with the manner in which speech sounds are produced by the articulators of the vocal system.
2. *Acoustic phonetics* studies the sounds of speech through analysis of the acoustic waveform.
3. *Auditory phonetics* studies the perceptual response to speech sounds as reflected in listener trials.

This study represents a blend of articulatory and acoustic analysis.

[C19: from New Latin *phōnēticus*, from Greek *phōnētikos*, from *phōnein* to make sounds, speak]

**phonotactics** *n.* the study of the possible arrangement of the sounds of a language in the words of that language. Phonotactics refers to the way sounds combine with other sounds in a language. For example, the combination "nglib" cannot be a syllable, according to the phonotactic rules of English.

**prosody** the patterns of stress and intonation in a language.

**psycholinguistics** *n.* the psychology of language, including language acquisition by children, the mental processes underlying adult comprehension and production of speech, language disorders, etc.

**schwa** *n.* a central vowel represented as (ə). The sound occurs in unstressed syllables in English, as in *around*, *mother*, and *sofa*. There are 12 principle vowels in American English. Phoneticians often recognize the *schwa* as the thirteenth vowel, which is a sort of "degenerate vowel" to which many others gravitate when articulated hastily in the course of flowing speech. [19, p. 119]

**semantics** *n.* the branch of linguistics that deals with the study of meaning, changes in meaning, and the principles that govern the relationship between sentences or words and their meanings.

[C19: semantic, from Greek *sēmantikos* having significance, from *sēmainein* to signify, from *sēma* a sign]

**strident** *phonological feature* strident sounds are marked acoustically by greater noisiness than their non-strident counterparts [3, p. 118].

**suprasegmental** *adj.* denoting those features of a sound or sequence of sounds that accompany rather than form part of the consecutive segments of a word or sentence, as, for example, stress and pitch in English.

**syllable** *n.* a combination or set of one or more units of sound in a language that must consist of a sonorous element (a sonant or vowel) and may or may not contain less sonorous elements (consonants or semivowels) flanking it on either or both sides: for example, "paper" has two syllables. [C14: via Old French from Latin *syllaba*, from Greek *sullambanein* to collect together, from *sul-* SYN- + *lambanein* to take]

**sonority** The sonority of a sound is its relative loudness compared to other sounds, everything else (pitch, etc.) being equal. [3, p. 132]

**tense** *adj.* pronounced with considerable muscular effort and having relatively precise accuracy of articulation and considerable duration: *in English the vowel (i:) in "beam" is tense. Compare lax.*

# Appendix B

## Sunspeech Corpus

### B.1 The Allowed Phonemes with their ‘ASCII code’ for the DSP Speech Database

This document gives a list of allowed phonemes with their pseudo ASCII values for the DSP speech database. This list essentially includes only the underlying phonemes. This should keep the transcriptions as consistent as possible even though they are not entirely correct. An "A" in parenthesis refers to an Afrikaans word example.

### B.2 Addendum to original documentation

In this document, the author added the first column containing the more correct<sup>1</sup> transcriptions using the International Phonetic Alphabet<sup>2</sup>. It now also includes ALL 75 ASCII values found in the database<sup>3</sup>.

There also seems to be a discrepancy in the amount of files. There are 2673 .ana digitised sounds files but 2760 .trc transcription files.

Files *agv10001* and *sdp10001* do not have a syllable category of transcription in the t90 files.

---

<sup>1</sup>The author used a few well-known texts to find the binary features for the phonemes found in Sunspeech. These included the famous *The Sound Pattern of English* of Chomsky and Halle[16], the Afrikaans text, *Afrikaanse Fonetiek*, by Combrink and De Stadler[24] and *English Phonology* by Giegerich[3] whose data on binary features is based on Chomsky's.

<sup>2</sup>The L<sup>A</sup>T<sub>E</sub>X package TIPA, or T<sub>E</sub>XIPA, was used to create the phonetic symbols in this document. It can be found under **fonts/tipa** of the CTAN archives

<sup>3</sup>The previous version of this data was documented and distributed with Sunspeech and listed only 58 distinct phoneme codes

## B.2.1 Vowels

Transcription	Old Trans.	ASCII	Word example	Occ
a	a	97	dug	4969
e	e	101	fear	2278
i	i	105	meet	7374
o	o	111	poor	847
u	u	117	boot	1466
y	y	121	uur (A)	1422
ɛ:	ɛ:	130	bêre (A)	415
ɛ	ɛ	131	met	4817
ɔ:		132	paw	2378
ɔ:		133	bore	791
ɒ		135	commuters <i>only found once</i> <i>in m_b20022</i>	1
ø	ø	142	kleur (A)	1011
ə	ð	143	ago	17648
ə:	ð:	144	flower	718
æ	æ	145	bat	3055
œ	œ	149	nut (A)	3652
œ:	œ:	150	fur	1321
ɑ	a	247	bar	2732
?	-	86	<i>a mix of different things,</i> <i>mostly vowels</i>	51

## B.2.2 Diphthongs

Transcription	Old Trans.	ASCII	Word example	Occ
ʔ		68	<i>found four times in: Saudi Arabia few approach and boiled</i>	4
aɪ	a:i	126	bite	1812
o:i	o:i	128	mooi (A)	133
ɔɪ	oi	134	boy	632
ɛɪ	ɛi	140	bedjie (A)	81
eɪ	ði	151	fate	2238
ui	ui	153	moeite (A)	194
iu:	iu:	210	due	211
əʊ	œʊ	211	goat	907
œy	œy	217	lui (A)	449
au	ō:	245	brow	356

## B.2.3 Fricatives

Transcription	Old Trans.	ASCII	Word example	Occ
ʔ		70	<i>a collection of fricatives</i>	26
f	f	102	fat	3802
h	h	104	hat	1155
s	s	115	sit	9224
v	v	118	van	2983
x	x	120	gaan (A)	947
z	z	122	zip	2168
θ	θ	171	thin	687
ð	ð	172	then	1956
ʃ	ʃ	188	ship	1843
ʒ		195	vision	448

## B.2.4 Glides

Transcription	Old Trans.	ASCII	Word example	Occ
ʔ		71	<i>mostly w</i>	9
j	j	106	yet	958
w	w	119	win	2277

## B.2.5 Liquids

Transcription	Old Trans.	ASCII	Word example	Occ
r	r	114	rat	4835
l	l	108	lot	4821
R	R	82	brei (A)	1191
?		94	rou (A)	38
r	/	218	refers to a flap	873
?		76	<i>mostly r</i>	124

## B.2.6 Nasals

Transcription	Old Trans.	ASCII	Word example	Occ
m	m	109	mat	4335
n	n	110	net	11615
ŋ	ŋ	205	sing	2240
?		78	<i>some nasal</i>	13

## B.2.7 Stops

Transcription	Old Trans.	ASCII	Word example	Occ
?		83	t ts d	49
b	b	98	bat	2749
d	d	100	dog	4227
g	g	103	go	1093
k	k	107	kit	5110
p	p	112	pet	3180
t	t	116	tip	9050
mostly p		154	<i>glimpse mostly to be found in front of /b/ /d/ /s/</i>	36
mostly b		156	<i>object - p in gvd10001</i>	12
d or t		175		67
d or t		177	<i>almost always d, except in akb10057, avm10005 and whs10005</i>	46
k		201	impact	114
g		203		20

## B.2.8 Affricates

Transcription	Old Trans.	ASCII	Word example	Occ
t <sup>h</sup>	ts <sup>h</sup>	181	cats	621
ɖ	dz	184	cads	184
tʃ <sup>h</sup>	tʃ <sup>h</sup>	191	chin	1807
ɖʒ	d	193	jam	1007

## B.2.9 Other

Transcription	ASCII value	Word example	Occurences
Silence	42		5173
Silence	242	<i>This is a mistake in cvr10021, should have been 42</i>	1
?	61	<i>no symbol, see r_m10016</i>	28
?	63	Unknown transcription	2469

## B.3 Chomsky Features for Phonemes in Sunspeech

The set of binary features used to describe the phonemes is the following: [sonorant], [voiced], [syllabic], [consonantal], [continuant], [anterior], [coronal], [strident], [round], [high], [low], [back], [tense], [nasal] and [lateral]. Where a feature did not apply or was not known for a specific phoneme, it is indicated with a -1 entry.

### B.3.1 Vowels

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
a	97	1	1	1	0	1	-1	-1	-1	0	0	1	0	0	-1	-1
e	101	1	1	1	0	1	-1	-1	-1	0	0	0	0	1	-1	-1
i	105	1	1	1	0	1	-1	-1	-1	0	1	0	0	1	-1	-1
o	111	1	1	1	0	1	-1	-1	-1	1	0	0	1	1	-1	-1
u	117	1	1	1	0	1	-1	-1	-1	1	1	0	1	1	-1	-1
y	121	1	1	1	0	1	-1	-1	-1	1	1	-1	0	-1	-1	-1
ɛ:	130	1	1	1	0	1	-1	-1	-1	0	0	0	0	0	-1	-1
ɛ	131	1	1	1	0	1	-1	-1	-1	0	0	0	0	0	-1	-1
ɔ:	132	1	1	1	0	1	-1	-1	-1	1	0	1	1	1	-1	-1
ɔ:	133	1	1	1	0	1	-1	-1	-1	1	0	1	1	1	-1	-1
ɒ	135	1	1	1	0	1	-1	-1	-1	1	-1	-1	1	-1	-1	-1
ø	142	1	1	1	0	1	-1	-1	-1	1	1	-1	0	-1	-1	-1
ə	143	1	1	1	0	1	-1	-1	-1	0	0	0	0	-1	-1	-1
ə:	144	1	1	1	0	1	-1	-1	-1	0	0	0	0	-1	-1	-1
æ	145	1	1	1	0	1	-1	-1	-1	0	0	0	0	-1	-1	-1
œ	149	1	1	1	0	1	-1	-1	-1	1	0	0	0	-1	-1	-1
œ:	150	1	1	1	0	1	-1	-1	-1	1	0	0	0	-1	-1	-1
ɑ	247	1	1	1	0	1	-1	-1	-1	0	0	1	1	1	-1	-1
?	86	1	1	1	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

### B.3.2 Diphthongs

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
ʔ	68	1	1	1	0	1	-1	-1	-1	1	0	0	1	1	-1	-1
aɪ	126	1	1	1	0	1	-1	-1	-1	0	0	1	1	1	-1	-1
o:i	128	1	1	1	0	1	-1	-1	-1	1	0	0	1	1	-1	-1
ɔɪ	134	1	1	1	0	1	-1	-1	-1	1	0	0	1	1	-1	-1
ɛɪ	140	1	1	1	0	1	-1	-1	-1	0	0	0	0	0	-1	-1
eɪ	151	1	1	1	0	1	-1	-1	-1	0	0	0	0	-1	-1	-1
ui	153	1	1	1	0	1	-1	-1	-1	1	1	0	1	1	-1	-1
iu:	210	1	1	1	0	1	-1	-1	-1	0	1	0	0	1	-1	-1
əʊ	211	1	1	1	0	1	-1	-1	-1	1	0	0	0	-1	-1	-1
œy	217	1	1	1	0	1	-1	-1	-1	1	0	0	0	-1	-1	-1
aʊ	245	1	1	1	0	1	-1	-1	-1	1	0	0	1	1	-1	-1

### B.3.3 Fricatives

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
ʔ	70	0	0	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
f	102	0	0	0	1	1	1	0	1	0	0	0	0	1	0	0
h	104	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0
s	115	0	0	0	1	1	1	1	1	0	0	0	0	1	0	0
v	118	0	1	0	1	1	1	0	1	0	0	0	0	0	0	0
x	120	0	0	0	1	1	0	0	0	0	1	0	1	1	0	0
z	122	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0
θ	171	0	0	0	1	1	1	1	0	0	0	0	0	1	0	0
ð	172	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0
ʃ	188	0	0	0	1	1	0	1	1	0	1	0	0	1	0	0
ʒ	195	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0

### B.3.4 Glides

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
ʔ	71	1	1	0	0	1	0	0	0	1	1	0	1	0	0	0
j	106	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0
w	119	1	1	0	0	1	0	0	0	1	1	0	1	0	0	0

### B.3.5 Liquids

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
r	114	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0
l	108	1	1	0	1	1	1	1	0	0	0	0	0	0	0	1
R	82	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0
?	94	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0
r	218	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0
?	76	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0

### B.3.6 Nasals

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
m	109	1	1	0	1	0	1	0	0	0	0	0	0	0	1	0
n	110	1	1	0	1	0	1	1	0	0	0	0	0	0	1	0
ŋ	205	1	1	0	1	0	0	0	0	0	1	0	1	0	1	0
?	78	1	1	0	1	0	0	0	0	0	1	0	1	0	1	0

### B.3.7 Stops

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
?	83	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0
b	98	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
d	100	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0
g	103	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0
k	107	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0
p	112	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
t	116	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0
p	154	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
b	156	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
d/t	175	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0
d/t	177	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0
k	201	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0
g	203	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0

### B.3.8 Affricates

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
ts <sup>h</sup>	181	0	0	0	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
dz	184	0	0	0	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
tʃ <sup>h</sup>	191	0	0	0	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
dʒ	193	0	0	0	1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

### B.3.9 Other

Trans	ascii	Son	Voice	Syll	Cons	Cont	Ant	Cor	Stri	Rnd	High	Low	Back	Tense	Nasal	Lat
Sil	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sil	242	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
?	61	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
?	63	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

## B.4 Description of features used

**sonorant** A sonorant is a sound whose phonetic content is predominantly made up by the sound waves associated with voicing [3, p. 93]. Sonorous sound segments are: vowels, glides, nasals and liquids and non-sonorous segments are: stops, fricatives and affricates.

**voiced** A voiced sound is produced with a glottal setting consistent with vocal-fold vibration; a voiceless sound is produced with a glottal setting inconsistent with vocal-fold vibration [3, p. 122].

**syllabic** This feature is an indication of the role the segment plays in the structure of the syllable. All segments which can act as the nucleus of a syllable has the feature [+ *syllabic*]. This includes the vowels, although some consonants might also act as syllable nuclei. The feature is thus also necessary for differentiating syllabic nasals and liquids from their non-syllabic counterparts [24, p. 22].

**consonantal** Consonantal sounds are produced with a radical obstruction in the vocal tract [3, p. 94].

**continuant** A continuant is a sound during whose production the air stream is not blocked in the oral cavity [3, p. 93].

**anterior** Anterior sounds are produced with an obstruction that is located in front of the palato-alveolar region of the mouth; non-anterior sounds are produced without such an obstruction [3, p. 116].

**coronal** Coronal sounds are produced with the blade of the tongue raised above its neutral position; non-coronal sounds are produced with the blade of the tongue in the neutral position [3, p. 116].

**strident** Strident sounds are marked acoustically by greater noisiness than their non-strident counterparts [3, p. 118].

**round** Rounded sounds are produced with a narrowing of the lip orifice; non-rounded sounds are produced without such a narrowing [3, p. 107].

**high** High sounds are produced by raising the body of the tongue above the level that it occupies in the neutral position; non-high sounds are produced without such a raising of the tongue body [3, p. 105].

**low** Low sounds are produced by lowering the body of the tongue below the level that it occupies in the neutral position; non-low sounds are produced without such a lowering of the tongue body [3, p. 105].

**back** Back sounds are produced by retracting the body of the tongue from the neutral position; non-back sounds are produced without such a retraction from the neutral position [3, p. 104].

**tense** Tense sounds are produced with a deliberate, accurate, maximally distinct gesture that involves considerable muscular effort; non-tense sounds are produced rapidly and somewhat indistinctly [3, p. 98].<sup>4</sup>

**nasal** Nasal sounds are produced with a lowered velum, which allows the air stream to escape through the nose; non-nasal sounds are produced with a raised velum, so that the air stream can only escape through the mouth [3, p. 124].

**lateral** Lateral sounds are produced by lowering the mid-section of the tongue at one or both sides, thereby allowing the air to flow out of the mouth in the vicinity of the molar teeth; in non-lateral sounds no such side passage is open [3, p. 125].

## B.5 Utterances in Sunspeech

The following is the list of sentences used in the English subset of Sunspeech. Each utterance is labelled with three initials of the speaker followed by a number identifying the sentence used.

---

<sup>4</sup>According to [24, p. 25] the feature [tense] is not relevant for vowels and diphthongs

Label	Utterance
10021	troops were rushed to violence racked townships after running battles left many dead and injured
10022	commuters arrived late for work when a train ripped down overhead wires at a key point causing major disruptions
10023	a few fortunate south african women have thrown off the humdrum drudgery of housework with the aid of appliances
10024	ultrapasteurised milk simply means that it is heated to a higher temperature and for a shorter period than the ordinary product
10025	through observation experimentation contact with experts and extensive reading one can acquire an astounding knowledge
10026	a pourable marinade is a quick easy and utterly delicious way to give meals a succulent steakhouse touch
10027	the editor has taken account of the large number of suggestions for addition and improvement that have come to his attention
10028	the chief result is a dictionary that continues to justify its reputation as the most reliable authority on spoken and written english
10029	persons guilty of ungentlemanly conduct and vulgar behaviour are known as cads
10030	yellow flowers arrange in a vase provide a cheerful sight in a dreary room
10031	it is evident from spectrogram readings that the acoustic signal is rich in phonetic information
10032	yesterday the boys laughed at the vision of the cats chasing shivering rats
10033	taxation is applied not only to companies but to a greater extent to individuals
10034	the farmyard is an educational outing for children where they can experience many things about the farm life
10035	an agenda setting out all matters to be discussed at the annual general meeting will be sent to all members
10036	educational toys can help to develop a childs perceptions and to heighten various skills as well as providing thousands of hours of pleasure
10037	the man leisurely watched the documentary on the television about a treasure hunt
10038	the united nations food and agriculture organization has said it is time to stop destroying the worlds forests
10039	keeping a garden looking beautiful all year round need not be a back breaking task if the correct tools are used
10040	a lack of iron is known to be the most common deficiency in all young babies
10041	an autocratic style of management is unlikely to motivate staff as a haphazard approach
10042	effective open and honest communication makes all the difference between good performance and mediocrity
10043	in evaluating the market facts will emerge about the potential of the product or service
10044	even my sense of humour could evolve a better joke than that
10045	what on earth could be the object of telling us such a rigmarole of lies
10046	the conventional contractor cannot work in soweto because of gang violence
10047	the single storey buildings will ensure that the rural setting is retained
10048	the biggest threat to do matrix or impact printing is laser technology
10049	movie giant ryan oneal is not a pretty sight as he squeezes his bulk into shorts
10050	german companies and banks will be pouring money into south africa by the end of the year
10051	i have decided to permit the public yet another glimpse into my uncle oswalds life
10052	what a zimbabwean first notices when he comes to london is the immense age of everything around him
10053	the boys thoroughly enjoyed their outing to the beach with their toys
10054	there were tears of joy when the battle weary troops arrived home from saudi arabia
10055	the serval is one of the african wild cats a lovely spotted creature with long erect ears
10056	throughout their childhood young chimpanzees are constantly learning their own special language
10057	a good lawyer would not let a few useless fads upset his sense of purpose
10058	courgettes should be boiled until tender and crisp
10059	many fears have been expressed worldwide for the homeless kurds
10060	the pleasure of eating doughnuts is usually enjoyed by everyone

Table B.1: Sentences in the English subset of Sunspeech