Lexical Stress in Speech Recognition

ROGIER VAN DALEN

25th June 2005

Delft University of Technology
Faculty of Electrical Engineering, Mathematics, and Computer Science
Mediamatics: Man–Machine Interaction

Graduation committee: /ˌɡrædʒuˈeɪʃn̩ kəˈmɪti/
Dr. drs. L. J. M. Rothkrantz
Ir. P. Wiggers
Dr. E. D. Botma (Leiden University, Department of English)
Dr. eng. R. Bidarra
Dr. A. H. J. Oomes

Rogier van Dalen
Master's thesis, 25th June 2005
Lexical Stress in Speech Recognition

Delft University of Technology
Faculty of Electrical Engineering, Mathematics, and Computer Science
Mediamatics: Man–Machine Interaction

Keywords: Lexical stress; in Dutch; speech recognition

This work has been typeset with LaTeX. The text is composed in Computer Modern and the corresponding TIPA fonts, with the display set in Venetian 301.

Contents

Abstract

Every native speaker can hear the difference between (English) *súbject* and *subjéct* or between (Dutch) *voorkómen* and *vóorkomen*. Human listeners use lexical stress for segmentation and disambiguation. However, lexical stress is not normally modelled in automatic continuous speech recognisers. In this work it is modelled how lexical stress can be used in a speech recogniser. Though earlier efforts have not modelled stress for consonants, they appear to contain stress information as well. Furthermore, different spectral features are needed for different phonemes.

A baseline speech recogniser for Dutch and one that uses lexical stress information are trained. The stress-enabled recogniser's word error rate is lower by 2.6 %.

Preface

The work you have before you is the result of my Master's thesis project at Delft University of Technology, at the Man–Machine Interaction group. I would like to thank everyone involved in providing me with the facilities, CPU cycles, and time to discuss my system and thesis. In the latter department, I expressly thank Leon Rothkrantz and Pascal Wiggers. Leon Rothkrantz has been my supervisor, who would not only give me loads of suggestions, but also always look at the longer-term plans, warning me whenever I would come up with a blatantly and unrealistically optimistic planning. After discussing important things, or in between, he invariably takes the time for a joke, an anecdote, or two, or more. Pascal Wiggers provided me with the sources for his working systems (see section 5.1 on page 57) and explanations on the Hidden Markov Toolkit (see section 3.1 on page 31) where needed, which allowed me to pick up steam most immediately. Furthermore, throughout my work on this thesis we had interesting discussions about all kinds of aspects of my system, this report, the future of speech recognition and the world in general.

I would also like to thank Bert Botma of Leiden University for taking a seat on my graduation committee and commenting on the phonology-related part of this report, giving numerous invaluable remarks. Similarly invaluable comments have come from Mike Spaans and Birgit van Dalen.

Any mistakes, over-generalisations, or other inaccuracies in this report are mine, and are probably due to my not taking these people's advice on something.

A paper about the first part of my thesis work that was submitted to the Text, Speech and Dialogue conference is reproduced in appendix B on page 87.

Orthography

Literature about automatic speech recognition sometimes uses acronyms, apparently because authors wish to save trees or key strokes. The latter argument is obsolete; the former is better served by using another typeface or point size, or a smaller leading. I will therefore solely use understandability as a criterion for abbreviating terms. ASR for Automatic Speech Recognition does not pass this criterion. On the other hand, I trust the opaqueness to the general reader and the familiarity to the specialist reader will warrant the use of HMM (for Hidden Markov Model), MFCC (for Mel-Frequency Cepstral Coefficients), and LPC (for Linear Predictive Coding).

The dialects used in phonemic examples are ABN (Algemeen Beschaafd Nederlands, the prestige dialect of the Netherlands) and RP (Received Pronunciation, the prestige dialect of Great Britain), unless noted otherwise. The conventions used for the transcriptions are shown in appendix A on page 85.

Chapter 1

Introduction

> *Substantives should be wrote with a Capital Letter.*
>
> Daniel Fenning, *The universal spelling-book*

The first edition of Daniel Fenning's highly successful spelling book appeared in 1756. In the beginning of the eighteenth century English printed prose had come to use capital letters for virtually all nouns ("substantive Nouns" or "Substantives", as Fenning calls them), just like in present-day German. Writers would introduce an initial capital for any noun "if it bear any considerable Stress of the Author's Sense upon it". Their printers needed to deal with increasing, and sometimes haphazard, use of capitals in manuscripts; the result was that all nouns were capitalised. When after 1750 "the 'extra' initial capital had become merely a marker of word-class" (Osselton 1984), and thus fully predictable, it was dropped again (Osselton 1985). How come Fenning recommends initial capitals for nouns even in his 13th edition (Fenning 1770)?

Table 1.1 reproduces a table from this book that contains pairs of words "of the same Sound but of different Signification" and different stress patterns, as

Table 1.1 *Table XX from Fenning (1770, p. 60). The initial capitals on nouns are meant to set them apart from the verbs and render visible the differences in the stress patterns. The acutes are* after *rather than* on *capitals. 'Rebel' without an acute and 'presén' without a 't' are supposedly mistakes.*

	A'bsent	*To*	absént	*A*	Mínute		minúte
An	A'ttribute	*To*	attribúte	*An*	O'bject	*To*	objéct
	Aúgust		Augúst	*A*	Présent	*To*	presén
A	Cóllect	*To*	colléct	*A*	Prójject	*To*	projéct
A	Cómpact	*To*	compáct	*A*	Rebel	*To*	rebél
A	Cómpound	*To*	compoúnd	*A*	Récord	*To*	recórd
The	Cónfines	*He*	confínes		Réfuse	*To*	refúse
A	Cónduct	*To*	condúct	*A*	Súbject	*To*	subjéct
A	Désert	*To*	desért	*A*	Tórment	*To*	tormént
A	Férment	*To*	fermént	*An*	U'nit	*To*	uníte
	Fréquent	*To*	frequént				

shown by the acutes ´. That they form different parts of speech, which is audible when they are pronounced, is shown in writing: nouns start with a capital. In other words, schoolchildren are told to capitalise their nouns to differentiate between stress patterns. This makes for an easier understanding of the text.

Even though the current-day orthography of languages like Dutch and English does not show *lexical stress* (word stress, as shown in a dictionary), the stress pattern is important for understanding the syntax and semantics of speech. Some examples are:

(1) English *récord* (n.) – *recórd* (v.)
English *pórtrait* – *portráy*
Dutch *óvervallen* 'robberies' (n.) – *overvállen* 'to rob' (v.)

(2) English *shórthand* – *short hánd*
Dutch *áanbod* 'offer' – *aan bód* 'first in line'

(3) English *trústy* – *trustée*
English *thírty* – *thirtéen*
English *digréss* – *tígress*
Dutch *vóorkomen* 'happen' – *voorkómen* 'prevent'
Dutch *avontúur* 'adventure' – *ávonduur* 'evening hour'
Dutch *Sérvisch* 'Serbian' – *servíes* 'tea-set'

Fenning's examples and those in (1) show how word classes for semantically related words may be signalled by the stress pattern. In (2) the spelling (word division) changes to form a different meaning depending on the stress pattern, though the pronunciation is otherwise the same. (3) has orthographically and phonemically similar words with different stress patterns and meanings, often because one is morphologically simplex and on complex. The pairs that are so similar phonemically that the stress pattern may be vital for telling them apart.

1.1 Human speech perception

Different languages have different ways of pronouncing and using stress. Germanic languages, such as English and Dutch, are *stress-timed* languages: the time between stressed syllables is roughly the same. In languages like French all syllables are evenly distributed over time; these languages are called *syllable-timed* (Ewen and van der Hulst 2001, p. 206). According to Jespersen (1952), Germanic languages stress the important parts (the root of lexical words), while there are languages "whose traditional stress rests or may rest on other syllables than the root".[1] I will consider stress in Germanic languages only, and focus on Dutch and English. How regular is stress assignment in those languages? Collins and Mees (1999, p. 230) say that in certain languages

> the stress falls overwhelmingly on a syllable in a particular position in the word (positional stress). For instance, in Czech and Slovak the stress is regularly on the first syllable. In many languages, it is on the penultimate (last but one) syllable, e.g. Italian, Welsh and

[1] The *root* is the element that carries the main component of meaning in a word. Affixes or inflectional endings can be attached to it.

> Polish. Some languages have stress on the final syllable, e.g. Farsi. In certain languages, e.g. French and many Indian languages, e.g. Hindi, Gujarati, native speakers do not appear to consider stress of significance. For instance, in French, although the tendency is for the word in isolation to have stress on the final syllable, this is often shifted to other syllables in connected speech.
>
> In English and Dutch, stress behaves in none of these ways. Stress is not easily and regularly predictable (an effect which may be termed dynamic). On the other hand, it is of importance to the word shape, and is not (as a rule) shifted from one syllable to another in connected speech. Consequently, we may say that for English and Dutch, and many other languages (e.g. Frisian, German, Russian, Danish, Spanish), stress is usually fixed for each word, but may occur on any syllable. Furthermore, in these languages, stress is of paramount importance to the native speaker in determining the meaning of the word.

Germanic used to have stress on the first syllable of the root. Though English and Dutch now have a stress assignment rule that works backwards from the last syllable (see section 2.1.1.1 on page 10), they have a general tendency to stress the first syllable of words, inherited from Germanic. Harley (2001, p. 221), in a book on psycholinguistics, describes how speakers use this for speech perception.

> [S]trategies that we develop to segment speech depend on our exposure to a particular language. Strong syllables bear stress and are never shortened to unstressed neutral vowel sounds; weak syllables do not bear stress and are often shortened to unstressed neutral vowel sounds. In English, strong syllables are likely to be the initial syllables or main content-bearing words, while weak syllables are either not word-initial, or start a function word (Cutler and Butterfield 1992; Cutler and Norris 1988). A strategy that makes use of this type of information is called the *metrical segmentation strategy.* It is possible to construct experimental materials that violate these expectations, and these reliably induce mishearings in listeners. For example, Cutler and Butterfield described how one participant, given the unpredictable words “conduct ascends uphill”[2] presented very faintly, reported hearing “The doctor sends the bill”, and another “A duck descends some pill”. The listeners have erroneously inserted word boundaries before the strong syllables and deleted the boundaries before the weak syllables. This type of segmentation procedure, whereby listeners segment speech by identifying stressed syllables, is called *stress-based segmentation.*

It appears that English-hearing infants have already learned to associate stressed syllables with word onsets at the age of seven months (Thiessen and Saffran 2003). Native speakers of Dutch appear to use the stress pattern to discriminate between words more than native speakers of English do, even when listening to English words. For example, when hearing the beginning of a word *octo-*, Dutch listeners will decipher whether it is *octó-* or *ócto-* and reconstruct *october* or *octopus*, respectively (Cooper *et al.* 2002).

[2] Note that two of the three words have a Romance (i.e. non-Germanic) origin. [RvD]

Though word segmentation and discrimination are major problems for current-day automatic speech recognisers, they do not recognise nor use stress. If the importance of lexical stress has been known since the eighteenth century, and psycholinguists have shown that listeners first identify stressed syllables to segment speech, why do automatic speech recognisers not even contain the notion of a 'stressed syllable'? This is probably because it is perceived as difficult to recognise stress, as it does not have one clear acoustic correlate (Sluijter 1995). Also, the literature on speech recognition often uses the English language, while Dutch apparently uses stress for discrimination as well. Furthermore, many speech recognition systems try and convert speech into text, and as we have seen, stress patterns have not been encoded in writing since over 200 years ago. That should not be an excuse, however: (4a) and (4b) differ both in stress pattern and spelling.[3]

(4) a. *Wanneer komt dat aan bod?*
[aːmˈbɔt][4]
When comes that first in line?
'When will that be discussed?'

b. *Wanneer komt dat aanbod?*
[ˈaːmbɔt]
When comes that offer?
'When will that offer come?'

Current speech recognition systems typically assume that speech consists of just a concatenation of *phonemes*[5]. They do not use the properties of the syllable as a whole. An automatic speech recogniser tries and find the sequence of words (and thus the sequence of phonemes) that best matches what it hears. Thus, a speech recogniser may mistake (5a) for (5b).

(5) a. *In Africa, one can meet a lion face-to-face!*
[əˈlaɪən]

b. *In Africa, one can meet alien face-to-face!*
[ˈeɪliən]

The fact that the difference in stress placement between *a líon* and *álien* is overlooked is unfortunate. It means that information that *is* encoded in speech is not used for recognition — and we are talking information that can be found in every dictionary, not something exotic or overly difficult to formulate. This information may not be strictly necessary to recognise what is said (as in the *a lion – alien* case) but humans encode the message in various channels to ensure the message gets across. This redundancy appears to be very much intertwined with prosodic structure (Aylett 2000), of which lexical stress is a part. Not using information from speech in a speech recogniser while the information is there

[3] *Aan* is a *clitic* that attaches to *bod* to form a *prosodic word* (Booij 1999), so the two sentences can be compared. See section 5.7 on page 62 for more information.

[4] *Phonetic transcriptions* are delimited by square brackets []. A stress mark ˈ is put in front of a stressed syllable. See Collins and Mees (1999) for the conventions of phonetic notations of English and Dutch.

[5] *Phonemes* are the smallest units of sound speakers are aware of. *Hor* and *hoor* contain a phonemic difference (/hɔr/ vs. /hoːr/). *Lel* /lɛl/ has the same phoneme, /l/, twice, but with different realisations, which are called *allophones*.

and humans do use it is hardly recommendable considering the deplorable state speech recognition is in, especially when compared to the dazzling foresights science fiction writers and technological experts alike (e.g. Baker 1975) have provided us with.

1.2 Definitions

Linguistic terms in the area of prosody often have different meanings. This actually means that terms used interchangeably by some are used with totally different meanings by others, and their meanings may be swapped by even other authors (the terms *prosody* and *intonation* come to mind). I shall use a demarcation that is more or less common in the field of research I am interested in: the following definitions are from Ladefoged (1975) or based on a linguistic tradition (Sluijter 1995; Ladd 1996).

Prosody Speech attributes that are not bound to phonemes.

Pitch The auditory property of a sound that enables a listener to place it on a scale going from low to high, without considering the acoustic properties, such as the frequency of the sound.

Fundamental frequency (or F_0) The acoustic correlate of pitch.

Intonation The pattern of pitch changes that occur during a phrase, which may be a complete sentence.

Accent A kind of prominence that accents semantically important words in an utterance, the main acoustic correlate of which is usually described (e.g. Bolinger 1986) as a pitch peak. An example in writing: "I didn't drown his *hamster*; I drowned his *cactus*!".

Stress The use of extra respiratory energy during a syllable. In utterances of more than two syllables multiple levels of stress may be distinguished.

Lexical stress Word stress. This makes the difference between *súbject* and *subjéct*; it is what this work is about.

1.3 Scope

There is much more to prosody than just lexical stress. However, the word level is consistently defined in phonological theory and automatic speech recognition practice alike. Speech recognisers have a list of words called the *lexicon* from which the words to be recognised are taken. Linguistic theory assumes a similar lexicon from which words are extracted with their stress pattern specified to form phrases when speaking. This work will not take into account higher-level phenomena than lexical stress, such as phrase stress and accent. This is for two reasons.

The first is the lack of knowledge about the sentence structure during recognition. For recognising stress on a phrase level, it must be known what parts of speech are concatenated and how these interact (see section 2.1.1.2 on page 12).

This requires knowledge about the syntax to be available during recognition, i.e. when the words are not yet known. Knowledge about word stress, on the other hand, may be built into the lexicon, a tool that is available in every typical speech recognition system.

The second is the lack of a theory about the phonological level. Kompe (1997, p. 109), trying to use intonation in a dialogue system, uses the "form" of the pitch contour to directly find its "function" ("a *rising* contour indicates a *question*"). However, this may be an oversimplification. Ladd (1996, p. 19) thinks that efforts to find things like the physical or auditory cues to question intonation while skipping the phonological level are misguided. It would be like studying the "physical cues to properties like plurality or verb aspect or negation — it seems obvious that it would be pointless to do so".

Many aspects of human speech come together in prosody. Phonologically, phrasal stress rules take lexical stress and prosodic phrase structure and produce difference in relative prominence (see section 2.1.1.2 on page 12). Semantics interact with this: words can be accented and thus be made more prominent as well. Syntactic structure is important for both, but the mapping from syntactic to prosodic structure is not straightforward (see section 5.7 on page 62 for an example). A comprehensive theory taking in all these aspects is necessary for a full account of lexical stress.

This overarching theory is yet to be found. Taglicht (1998) shows that for a satisfactory account of the interaction of syntax and phonology it is necessary to introduce "syntactic constraints on intonational phrasing". Hyde (2002) proposes a new theory to account for the interaction of stress and metrical structure. From a more psychological viewpoint, Port (2003) finds a "perceptual 'beat' that occurs near the onset of vowels" and suggests "a phonological grammar should probably be built on top of this sloshy, dynamical timing system". It seems that a proper model of intonation needs to derive from a phonological description of intonation, derived from a morphological description, which in its turn needs to derive from a syntactic description. A metrical system probably interacts with it. None of these are already well enough defined and comprehensive enough to be used in a computerised recognition system.

Though we know in advance that the contents of the lexicon cannot fully account for stress, let alone of the whole area of prosody, we do know that lexical stress has a set of acoustic correlates (Sluijter 1995). Recognising lexical stress should be a major first step in recognising phrasal stress. Though much variance from factors that cannot yet be modelled is expected, future linguistic research will make the intonation part easier to separate from the stress part. To cope with the variance now we shall use probability functions, as is usual in speech processing.

1.4 Research questions

If lexical stress information should be an important subject of research in speech recognition, where to start? First, we must find out how stress is realised acoustically. It is not straightforward, however, to implement in a computer system a comprehensive set of rules for all subtleties of stress. It should however be possible to model at least the basics of lexical stress.

Question 1 *What are the acoustic correlates of lexical stress?*

Even with a good qualitive description of the acoustic correlates of stress—and even with a quantitative description to match it—it may not be obvious how to convert lexical stress information in the speech signal into the best form for automatic recognition. The representation should consist of few values and contain as little noise as possible.

Question 2 *How can the acoustically coded stress information be extracted from recorded speech and be represented as feature values?*

Computers can take into account only a finite number of possibilities in finite time. This has led to a number of simplifying assumptions representing a trade-off between recognition accuracy and feasibility. Such assumptions are found in the probability functions used by *Hidden Markov Models* (see section 2.2.3 on page 26). These model the phonemes in most of today's speech recognition systems and are used in this work as well.

Question 3 *How can lexical stress data best be processed using the technology at hand—Hidden Markov Models?*

As noted, current-day speech recognition systems do not take syllables into account. This may be a problem since lexical stress is specified per syllable.

Question 4 *How can lexical stress information from the segments in one syllable be integrated?*

Even if lexical stress is recognised correctly in an overwhelming majority of cases, it will not necessarily improve overall recognition. For example, minimal pair recognition (think *súbject – subjéct*) may improve while confusion between phonemes and words increases.

Question 5 *What influence does automatic lexical stress recognition influence have on what is recognised? To what extent can it theoretically help recognition? Does it help recognition in practice?*

It has already been hinted that current speech recognition technologies may limit the implementation of lexical stress recognition. Hidden Markov Models and search algorithms (see section 2.2.4 on page 27), if reasonable for everyday speech recognition, may pose such constraints on a speech recogniser that make the recogniser incapable of capturing the acoustic properties of stress to their full extent. Thus, it might be rewarding to look into new technologies that better suit lexical stress recognition and can be used in the future.

Question 6 *What speech processing technologies may in the future be used to integrate lexical stress information into speech recognisers?*

1.5 Operational

Answering the research questions in section 1.4 requires looking into the phonology and acoustics of stress, and the relevant properties of speech recognisers. I will take the following steps:

- Identify the properties of lexical stress, both phonological and acoustic, from the literature. Find the features of lexical stress that can help automatic speech recognition from the literature. Test whether these are also found in continuous speech corpora; find out how they can best be represented in a speech recognition system. This will require modelling how lexical stress can be mapped onto speech recognition building blocks.
- Identify a subset of lexical stress-related speech recognition techniques to model lexical stress within the given constraints. Implement these in a speech recogniser and test it to whether recogition improves.
- Identify what can not yet be modelled and how this may be done in the future.

Chapter 2 on the facing page goes into lexical stress and speech recognition theory. Chapter 3 on page 31 discusses the tools that will be used for the implementation. Chapter 4 on page 35 proposes speech recogniser models that enable stress modelling. Chapter 5 on page 57 discusses the system that will be built to test the hypotheses. Chapter 6 on page 69 goes into the results from the experiments. Chapter 7 on page 79 concludes this thesis and discusses future work.

Chapter 2

Theory

> *The one is the* to poiein*, or the principle of synthesis, and has for its objects those forms which are common to universal nature and existence itself; the other is the* to logizein*, or principle of analysis, and its action regards the relations of things, simply as relations; considering thoughts, not in their integral unity, but as the algebraical representations which conduct to certain general results.* Percy Bysshe Shelley, *A Defence of Poetry*

Computers do not have intuitions about language. As advanced as they sometimes seem to be, they cannot work with vague descriptions of phenomena they are supposed to spot. Uncertainty must be coded using statistical probabilities. This shows in the building blocks of speech recognition as presented in section 2.2 on page 23, which also discusses the connection between speech recognition and linguistic sciences and some outstanding problems in relating the two. Information on stress in linguistics can be found in section 2.1, which discusses the actual properties of lexical stress and should be uncontroversial, quite informal, and understandable to the non-linguist.

2.1 What is stress?

Before embarking on our journey towards the best way to make an automatic speech recogniser use lexical stress, we have to make sure we understand what lexical stress is.

The definition of stress from section 1.2 on page 5 indicates that stress is relative. Lexical stress is related to syllable prominence within a word. For example, the verb *to permít* and the noun *a pérmit* form a *minimal pair*[1], with the verb having stress on the second syllable and the noun having stress on the first syllable.[2] It is not perfectly clear what the phonetics and acoustics of lexical stress are. One major problem is where to draw the thin line between stress and intonation. Conceptually, lexical stress is no more than a phonological indication of prominence. As is indicated by the word "lexical", the difference between *to*

[1] I.e. they only differ on one point, in this case: the placement of lexical stress.

[2] For the intuitive meaning of 'syllable' for native speakers of Germanic languages (e.g. English, Dutch, German).

permit and *a permit* is clear from a dictionary: the phonemic transcription of the former is /pəˈmɪt/ and of the latter /ˈpɜːmɪt/.[3] What influence does that have on a speech recogniser's input? How can these vague notions be put in more specific terms?

2.1.1 Phonologically

Traditionally, phonetic literature does not distinguish between lexical stress and other prosodic traits clearly. There are good reasons for this. When you want to make a word stand out in a sentence ("It wasn't his *teacher*, it was his *father*!"), you use intonation by putting an *accent* on the word: you pronounce the word louder and with a higher pitch. More specifically, though, the accent is on the stressed syllable.[4] You do not say *fathér*; you say *fáther*. There is no specific reason for the position of the stress mark[5], other than how the word is specified in your internal 'dictionary', like it is in a conventional dictionary. If you call this internal dictionary the *lexicon*, as linguists do, you can call the kind of stress indicated in the dictionary *lexical stress*.

Bolinger (1986) looks at sentences in which words stand out and concludes "[t]he stressed syllable is the one that carries the potential for accent." In other words: a syllable may have lexical stress in the lexicon, but this abstract type of stress is only pronounced if the word has the accent, i.e. if the word is made to stand out in the sentence. This explains why pitch sometimes is such a reliable cue for detecting lexical stress (the syllable has a pitch accent when the word is accented) and is virtually irrelevant for words in unaccented positions.

Later theories (e.g. Ladd 1996) suppose there are two different categories of prominence features, separating *stress* and *accent*. Accents are explicitly linked to the structure of the intonation contour. Lexical stress, though interacting with accent structure, does have acoustic correlates of its own, which arise from differences in pronunciation such as an increase in respiratory energy, and is defined rhythmically. This makes sense of Bolinger's observations and at the same time introduces the notion that lexical stress has a phonetic reality. Before looking at the phonetic and acoustic impact of lexical stress, though, let us skim over some of its phonological properties.

2.1.1.1 Syllables

Though the word *syllable* came up a number of times, no definition was given yet. There are considerable theoretical difficulties in defining it.[6] However, for discussing stress a notion of what a syllable is must be established. The absence of a definition so far should not have mattered much, because native speakers of English and Dutch appear to have a quite consistent feeling for what a syllable is. However, computers do not share these feelings. So what is a syllable?[7] Apparently it is a vowel surrounded by consonants. As this vowel is the centre

[3] According to Procter (1995). Note that /ɜː/ becomes /ə/ when it is not stressed.

[4] Accents can also be on a lexically unstressed syllable, for example in *I didn't pour milk into my cof**fin**, but into my cof**fee***, but that needs not concern us here.

[5] No synchronic reason, at least; we are not now, however, concerned with historical phonology.

[6] Some (e.g. Harris 1994) do not believe there is such a thing as a syllable.

[7] This explanation is based on Ewen and van der Hulst (2001).

of the syllable, we call it the *nucleus*. It does not *have* to be surrounded by consonants, however. Let us look at the *syllabification* of various words.

(1) albatross [æl]$_\sigma$[bə]$_\sigma$[trɒs]$_\sigma$
America [ə]$_\sigma$[mɛ]$_\sigma$[rɪ]$_\sigma$[kə]$_\sigma$
slender [slɛn]$_\sigma$[də]$_\sigma$

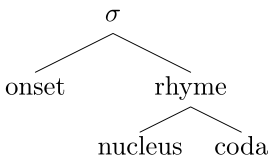
(1) seems to suggest that indeed every vowel forms a nucleus, and the consonants are pasted onto it. However, the word *confusion* consists of three syllables, as any native speaker can tell, but is pronounced with two vowels only.

(2) confusion [kən]$_\sigma$[fjuː]$_\sigma$[ʒn]$_\sigma$
bottles [bɒʔ]$_\sigma$[ɫz]$_\sigma$[8]

The last syllable of [kənfjuːʒn] has no vowel. The nucleus here is the consonant [n], which is called *syllabic*. A *syllabicity mark* below the consonant can be used to show this: [n̩]. But why does [n] form the nucleus in [ʒn̩], while [ə], and not [n], does in [kən]? It appears that it is not the kind of segment, but the *relative sonority* (something like 'loudness') that makes a syllable a syllable: the vowel or syllabic consonant is surrounded by less sonorous segments, forming a *sonority peak*.

What parts does a syllable consist of, other than the nucleus? Something may come before the nucleus, and something after it. Because 'first bit' and 'last bit' sound decidedly unacademic, these parts are commonly referred to as *onset* and *coda*.

(3)



The structure in (3), with the *rhyme* consisting of the nucleus and the coda, is commonly used. Why is the *rhyme* a constituent? First of all, it is the part of a syllable that makes it rhyme with other words (hence the name), but also the position of English lexical stress is sensitive to rhyme structure. Let us take a look at stress assignment for nouns.

(4) a. arena [ə]$_\sigma$[ˈriː]$_\sigma$[nə]$_\sigma$
angina [æn]$_\sigma$[ˈdʒaɪ]$_\sigma$[nə]$_\sigma$
b. America [ə]$_\sigma$[ˈmɛ]$_\sigma$[rɪ]$_\sigma$[kə]$_\sigma$
cholesterol [kə]$_\sigma$[ˈlɛ]$_\sigma$[stə]$_\sigma$[rɒl]$_\sigma$
c. agenda [ə]$_\sigma$[ˈgɛn]$_\sigma$[də]$_\sigma$

Stress assignment in English works from the last syllable (Liberman and Prince 1977).[9] The rule for nouns is:

- if the penultimate syllable's rhyme consists of more than one element, it will be stressed;

[8] In dialect speech (for example, Cockney). [ʔ] is used rather than RP [t] for expository purposes: it makes the syllable division clearer, but does not affect the argument.

[9] The first syllable of many content words is stressed because they are often short.

- otherwise, the antepenultimate syllable will be stressed (whether it is long or not).

Thus, we have *aréna* (rhyme /iː/[10], *angína* (rhyme /aɪ/) and *agénda* (rhyme /ɛn/), but not *Ameríca* (rhyme /ɪ/) or *cholestérol* (rhyme /ɛ/). This demonstrates that the number of elements in the rhyme does indeed determine stress assignment: the rhyme must therefore be a constituent.

This shows that there is a difference between the onset and the coda. This is not trivial: do not consonants after one vowel usually precede another? There is also a difference in the realisation of consonants in the onset and in the coda. For example, Dutch *roer* 'ruther' /rur/ may be realised as [ʀüəɹ]. This realisation, which is used by many speakers of ABN, the Dutch prestige dialect, shows the tendency of /r/ to become an approximant in the coda.

An example where the stress of the syllable, the segment's position in the syllable, and its neighbouring segments influence the realisation of the segment is English /t/.

(5)	a.	tile	[t^{h}aɪl]$_{\sigma}$
		try	[tr̥aɪ]$_{\sigma}$
	b.	stile	[staɪl]$_{\sigma}$
		strive	[straɪv]$_{\sigma}$
	c.	light	[laɪt]$_{\sigma}$
(6)	a.	retire	[rɪ]$_{\sigma}$[ˈt^{h}aɪə]$_{\sigma}$
		retry	[rɪ]$_{\sigma}$[ˈtr̥aɪ]$_{\sigma}$
	b.	distend	[dɪ]$_{\sigma}$[ˈstɛnd]$_{\sigma}$
		distress	[dɪ]$_{\sigma}$[ˈstrɛs]$_{\sigma}$
(7)	a.	mutter	[ˈmʌ]$_{\sigma}$[tə]$_{\sigma}$
		mattress	[ˈmæ]$_{\sigma}$[trəs]$_{\sigma}$
	b.	muster	[ˈmʌ]$_{\sigma}$[stə]$_{\sigma}$
		nostril	[ˈnɒ]$_{\sigma}$[strəl]$_{\sigma}$

In (5) the only syllable of the word is stressed. When /t/ is in the coda, as in (5c), it is not aspirated. It *is* aspirated in (5a), where it is in the onset; and if it is followed by a liquid, /r/ in this case, the aspiration spreads to the liquid, which becomes voiceless. However, before /s/ (5b) /t/ is unaspirated. From (6) it appears that the aspiration depends on stress rather than the position of the syllable: the same pattern is exhibited in the second syllable when it is stressed. In (7), the syllable containing /t/ is unstressed; thus, it is never aspirated.

This shows that stress of the syllable, the position of the segment (in the onset or in the coda) and the neighbouring segments may have an influence on a segment's realisation. English fortis stops (/p, t, k/) are aspirated when they are in the onset of a stressed syllable and not preceded by /s/.

2.1.1.2 *Rhythmicality*

The rhythmicality of stress and the correlation with pitch accent can be expressed in a *bracketed metrical grid* (Ewen and van der Hulst 2001).[11] A grid of

[10] /iː/ means /ii/.

[11] Ewen and van der Hulst (2001)'s definitions of the terms "accent" and "stress" are quite different from mine (see section 1.2 on page 5); the latter will be used for consistency.

this type shows stress on various levels. Let us use the phrase *a lengthy oratorio* /ə ˈlɛŋθɪ ˌɒrəˈtɔːrɪəʊ/[12] to show this.

```
(8)                          H              tone
                             |
        (·                   ×          )   phrasal stress
        (×     )  (·         ×          )   word stress
        (×    ·)  (×    ·)  (×    ·    ·)   foot stress
    σ    σ    σ    σ    σ    σ    σ    σ
    ə    lɛŋ  θɪ   ɒ    rə   tɔː  rɪ   əʊ
```

Every line in (8) represents a level in the rhythmic hierarchy. So far, we have only dealt with word level stress. A syllable with a cross over it is the *head* of a domain, which is delimited by brackets. At the top level of (8) an abstract accent is depicted on a 'tone' tier. This is where stress interfaces with sentence accent.

Before deciding which part of (8) is going to be relevant for the speech recogniser, let us take a look at more complicated examples. Some of the more intriguing aspects of stress assignment become visible when we look at compound formation. Inserting a new level for compounding, we get the following for *White House politics*:

```
(9)                                          (·           ×             )   phrasal stress
     (×      ·)                              (×      ·)   (×            )   compound stress
     (×)    (×)   + (×          ·)   →       (×)    (×)   (×          ·)   word stress
     (×)    (×)     (×     ·)  (×)           (×)    (×)   (×     ·)  (×)   foot stress
      σ      σ       σ     σ    σ             σ      σ     σ     σ    σ
     waɪt   haʊs  +  pɒ    lɪ   tɪks  →      waɪt   haʊs  pɒ    lɪ   tɪks
```

Note that the first syllable of *politics*, which is not a compound, vacuously receives a cross from the compound stress rule, now that we have introduced a new level. Other than that, this all looks pretty straightforward. However, what happens when we make a compound from two elements such that the stress pattern at one level has two adjacent crosses? Let us take *New York pizza*. First try it out aloud. You should find that even though you say *New Yórk*, in *New York pizza*, *Néw* has more prominence than *York* (but still less than *pízza*).

```
(10)  (     ·    ×    )      (·         ×    )      compound stress
      (·    ×)  (×    )      (×    ·)  (×    )      phrasal stress
      (×)  (×)  (×    )  →   (×)  (×)  (×    )      word stress
      (×)  (×)  (×   ·)      (×)  (×)  (×   ·)      foot stress
       σ    σ    σ   σ        σ    σ    σ   σ
      njuː jɔːk piː tsə      njuː jɔːk piː tsə
```

On the left side of (10) a *rhythmic clash* occurs because the two rhythmic 'beats' corresponding with the lexical stress, /njuː ˈjɔːk/ and /ˈpiːtsə/, are adjacent, yielding /njuː ˌjɔːk ˈpiːtsə/. The clash is resolved by moving the secondary stress to the left: the result is a /ˌnjuː jɔːk ˈpiːtsə/.

How difficult is it for a speech recogniser to cope with this effect? The speech recogniser is going to parse Dutch, and not English; however, Dutch has

[12] The symbol ˌ is used to indicate *secondary stress.*

pairs similar to the English ones, like *Nieuwescháns* and *Níeuweschans-Oóst*. The problem is that *Nieuweschans* is transcribed in the lexicon we use for the recogniser (yes, it is actually there) as /niːwəˈsxɑns/, which is wrong whenever a stress shift happens. Using the obvious remedy, adding to the recogniser lexicon all possible combinations that yield an exception, would lead to a combinatorial explosion of the lexicon size. This information is very difficult to build into a speech recogniser. This can be a problem if it hears a stress pattern different from what is specified in the lexicon: it may not recognise the word at all.

2.1.1.3 *Feasibility*

In the theoretical corner we have now found two major problems:

- It is not clear whether stress is *always* realised phonetically, and whether it can always be detected.
- It is not clear whether lexical stress often enough leads to actual stress: as we have seen, different mechanisms work together to yield a stress pattern in a way that is rather more complex than looking up transcriptions in a dictionary and blindly copying them.

As for the first point, more elaborate phonological theories than Bolinger's assume there *is* a phonetic reality to stress. Whether the computer will be able to detect stress must be determined. These very phonological theories, however, also significantly complicate the prediction of stress patterns. For now, we will disregard those theories; however, see section 4.2 on page 41 for ideas on how they may be incorporated into a more sophisticated speech recognition system.

2.1.2 *Acoustically*

If stress is formulated as prominence from a phonological point of view, how can it be seen acoustically? This is not an idle question: if a speech recogniser is to detect stressed and unstressed segments, we have to tell it how to detect those. Stress seems to have to do with effort and taking the time to pronounce a sound properly. Intuitively, it would seem that stressed syllables are in general pronounced with more force and care than unstressed ones, which are subject to all kinds of reduction, especially in 'careless' speech.

De Jong *et al.* (1993) claim that "[s]tressed syllables have more distinctive articulations, whereas unstressed syllables have 'undershoot' due to greater coarticulatory overlap with their neighboring segments' gestures." This basically means that the influence of adjacent sounds on the unstressed syllable is larger than the influence on stressed syllables. It is as though stressed syllables are so strong they can 'fight off' the influence of neighbouring segments. Ewen and van der Hulst (2001) speak of duration, amplitude and pitch as phonetic exponents of stress, at least in Dutch and English. They mean, one may presume, that stressed segments have a longer duration, higher amplitude and most likely, higher pitch (or something like a pitch peak).

Let us do a small experiment showing what properties have actually been found to be useful for automatic recognition of lexical stress from sound recordings. We will look at a minimal pair from the Dutch language: *kanón* /kaːˈnɔn/

'gun' and *cánon* /ˈkaːnɔn/ 'canon'. We use these words because they are also used by Sluijter (1995), and because the words differ only in stress, and not in phonemes.[13] We will look at a number of different representations of the sound as pronounced by a native speaker of the language (i.e. me). I pronounced them in isolation, so that the effects will be quite clearly visible, similarly to Sluijter's laboratory experiment. Magnified though the effects may therefore be, they are found in more real-world scenarios as well (van Kuijk and Boves 1999). The pictures have been made with the Praat program (Boersma 2001).

2.1.2.1 *Wave form*

The first way we will represent the sound is by a graph showing the air pressure over time. The most interesting thing we can see here is the amplitude (the loudness) of the sounds. In figure 2.1(a) on the next page the final syllable /ˈnɔn/ is much louder than the final syllable in figure 2.1(b) /nɔn/.

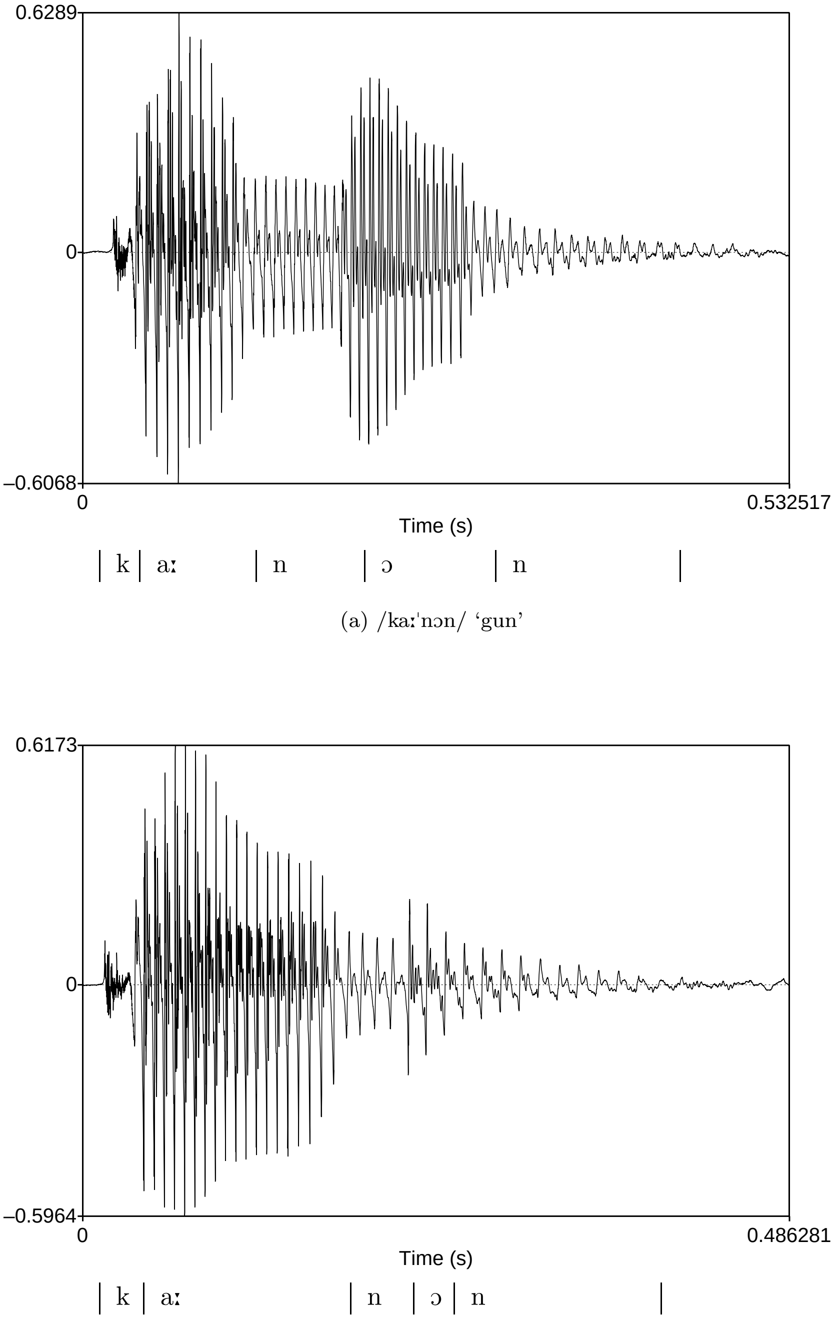
2.1.2.2 *Pitch*

An obvious candidate for showing which syllable is stressed is a picture showing the *fundamental frequency.* This is the string of 'notes' the utterance is on. Stressed parts of words, phrases and sentences are commonly signalled by higher pitch in Dutch. If you look at the two pitch contours in figure 2.2 on page 17, it is immediately clear which represents *kanón* and which represents *cánon*, given that the segmental content (i.e. the sounds) is /kaːnɔn/. This makes it seem that it is easy to figure out the stress pattern by just looking at the pitch contour.

However, it is not *that* easy. In figure 2.3, where I pronounced the two words with question intonation, it is clear that word stress is not in all cases as directly linked to intonation contour as it seemed just now. And this is just one example, copied more or less from Ladd (1996). Ladd uses the English minimal pair *pérmit – permít* though, and the intonation contours are even more similar. Another problem is that though I pronounced the words in isolation (in *citation form*) for this experiment, in fluent speech there often are effects from the intonation of the sentence as a whole breaking the seemingly nice pattern of a higher fundamental frequency accompanying word stress. As Collins and Mees (1999) say, the pitch change "may be either to a higher or lower pitch, or may involve a sustained pitch on a low or high tone".

There is not in general a relation between word stress and a high fundamental frequency: sentence intonation interferes with the citation form pattern. Ladd (1996, p. 55) draws a distinction between (phonetic) *alignment* and (phonological) *association.* Even if in an abstract sense pitch peaks are associated with stressed syllables, in a phonetic and acoustic sense "the peak may be early in the syllable or late, and indeed it may be outside the temporal limits of the syllable altogether. For example, it is particularly common in accented syllables at the beginning of an utterance to see the high F_0 peak aligned in time with the following unstressed syllable." This would make using pitch for interpreting the position of lexical stress difficult without modelling and overlaying a prosodic or intonational tier to be matched later with the segmental content; this would be outside the scope of this thesis (see section 1.3 on page 5).

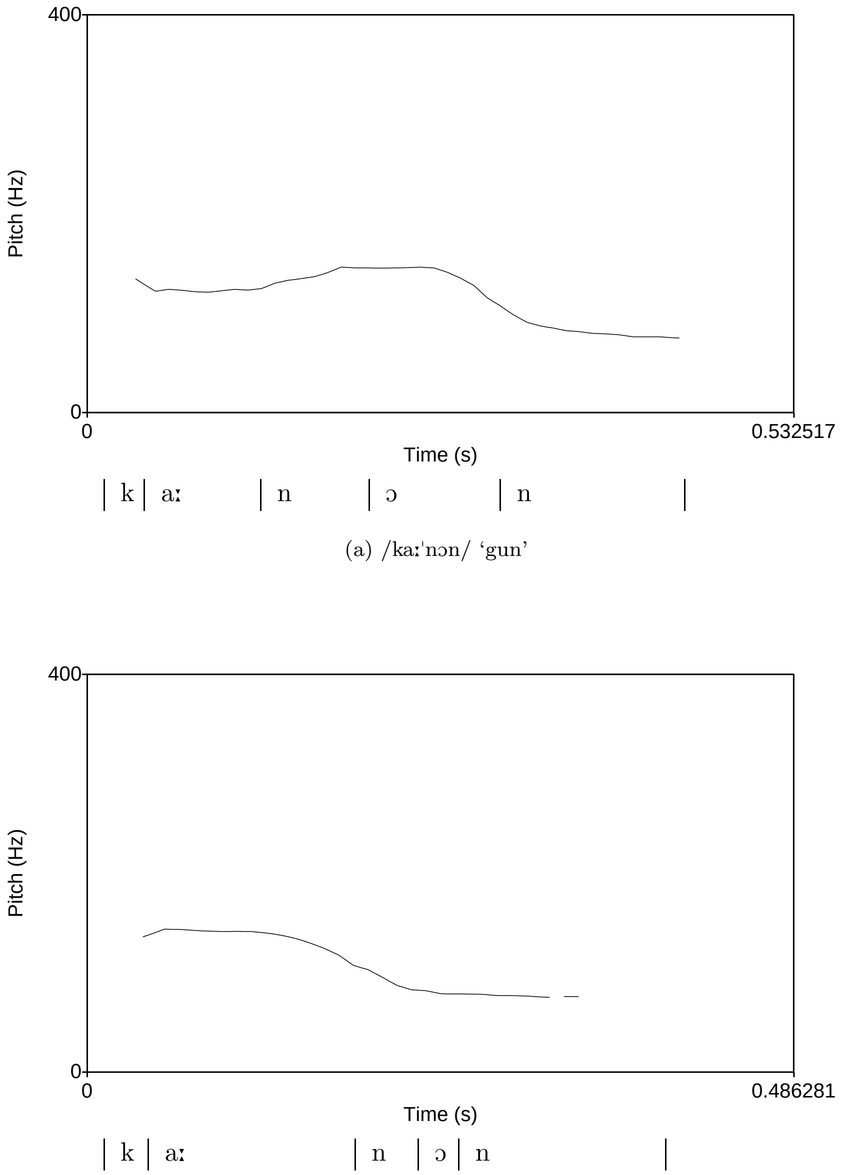
[13] According to Geeraerts (2000).



(a) /kaːˈnɔn/ ‘gun’

(b) /ˈkaːnɔn/ ‘canon’

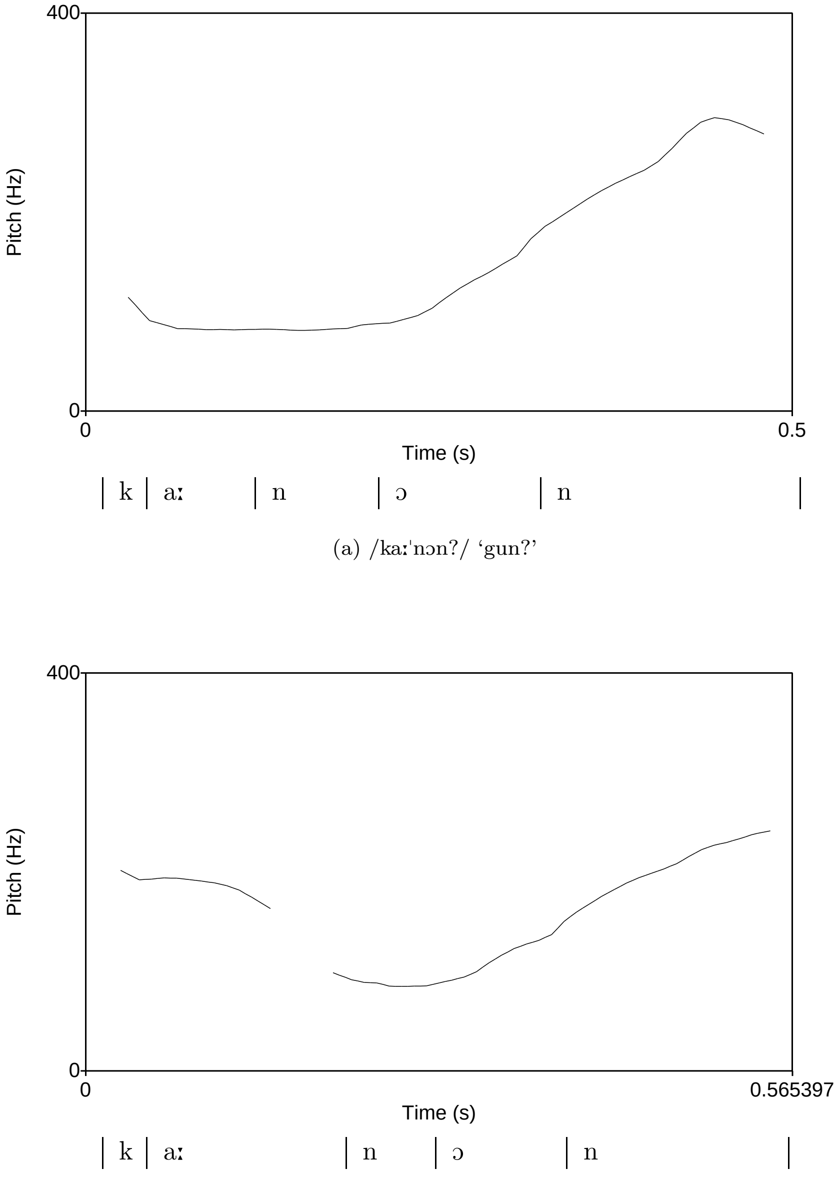
Figure 2.1 *Waveforms for the* kanón – cánon *minimal pair. Stressed syllables are notably louder and longer.*



(a) /kaːˈnɔn/ ‘gun’

(b) /ˈkaːnɔn/ ‘canon’

Figure 2.2 *Pitch contours for the* kanón – cánon *minimal pair. The pitch peaks are more or less on the stressed syllables.*



(a) /kaːˈnɔn?/ ‘gun?’

(b) /ˈkaːnɔn?/ ‘canon?’. (There is a gap in the contour that should not be there: the Praat program failed to pick up my voice at that position. This does not influence the argument, but if you wish, you can fill in the obvious missing line yourself.)

Figure 2.3 *Pitch contours for the* kanón? – cánon? *minimal pair. Due to the interaction with the intonation the pitch peaks are much less clear than those in figure 2.2.*

2.1.2.3 Duration

The diagrams have so far included a transcription of the sounds underneath them, showing which sound is pronounced at what exact point in time, and for how long it lasts. The latter is clearly of interest to us: stressed vowels are longer than unstressed vowels. This phenomenon appears to be quite robust and it has been consistently found in works like Sluijter (1995) and van Kuijk and Boves (1999).

2.1.2.4 Reduction

According to Carr (1999), "We have said that English is stress-timed, and that the number of unstressed syllables may be quite high. Given that it is possible to have many such intervening unstressed syllables, this put pressure on such intervening syllables to simplify in various ways." It is not easy to see from a diagram, even for experienced phoneticians, but unstressed vowels tend to be *reduced*. Shall we claim that this is due to speakers not taking care when they pronounce unstressed vowels, pronouncing things sloppily? Shall we claim that unstressed vowels tend to be reduced to certain vowels, which are not necessarily the 'easy' ones, but apparently the number of 'target vowels' is small?

Being in the lucky situation of having to consider only Dutch, and maybe the language this is written in, English, we will skip the phonological theories and generalities, and just look at a few examples in text. The reason we chose *kanon* – *canon* as an example was that both are pronounced /kaːnɔn/. If you speak Dutch, now pretend not to notice whoever is currently sitting next to you, and say out aloud, but quickly, the word *kanón* 'gun'. Does your pronunciation start with /kaː/? Mine does not, if I say it quickly. I say /kə/. This is reduction of what 'ought to be' /aː/ to /ə/, which is the vowel Dutch reduces basically every other vowel to. (Shall we claim this is because /ə/ is the easiest vowel to pronounce we have? We shall not, though for Dutch it is a defensible notion.) /ə/ is the symbol for *schwa*, which is like the vowel you get when you let your mouth hang open and make a sound. It occurs in many Dutch words, for example: *prat**en*** 'talk' /ˈpraːt**ə**/. Note that the word *praten* also has reduction of /n/ to /∅/.

Of course there are all kinds of stages in between the full and the reduced form of a vowel. A bit of reduction might yield something like /kɑˈnɔn/ (which starts with the same sounds as *kan* /kɑn/). You could say that because you take less time to put your *articulators* (tongue, lips, and jaw) in the proper position, the result tends more towards the neighbouring segments, and on average, towards the 'default' position.

Reduction has also led to historical changes. Two well-known dialects of English, American English and British English (the official varieties of which are called *General American*, or GA, and *Received Pronunciation*, or RP, respectively) differ in how much reduction is used even in dictionary forms. The word *secondary* may be pronounced by an American as /ˈsɛkənˌdɛri/, which is the more historical form, while a British speaker will use /ˈsɛkəndri/ or even /ˈsɛkn̩dri/ (Procter 1995). The third syllable has been reduced; its vowel which was /ɛ/ has been reduced to /∅/, nothing, and the second vowel, /ə/, optionally undergoes the same treatment. Deletion, of course, is the most drastic form of reduction.

But how will a computer recognise vowel reduction, if even a trained phonetician cannot tell from a diagram? Let it suffice for now that computers *can* recognise different vowels, and often are as good as humans at it (Chang *et al.* 2000), so introducing a number of new vowels should in principle not be much of a problem. As we have seen, reduction comes in various shapes and sizes; but recognising large or even overlapping ranges of sounds as one vowel is something a computer can do as well. This enables it to hypothesise that something sounding like [aː], [ɑ], [ə], or even [∅] is a realisation of unstressed /aː/.

2.1.2.5 *Spectrum*

The spectrum of a sound is the energy at the different frequencies. A spectrogram is a plot of this information as a function of time. What you see in figure 2.4 on the facing page is the plots for the by now well-known *kanón* – *cánon* pair. The frequencies are on the y-axis. The more energy is output, the darker the colour. The interesting thing here is that stressed vowels have more energy in the higher frequencies relative to the lower frequencies than unstressed vowels. This can be checked by comparing the upper halves of the plots at the position where /ɔ/ is pronounced. In figure 2.4(a) this area is darker than in figure 2.4(b). This effect is called 'spectral balance' or 'spectral tilt', both signifying that relatively much energy goes into the higher frequency band. 'Spectral tilt' refers to the tilt of a line in a plot for one point in time.

Note that this difference in pronunciation between the two words does not disappear under question intonation: in figure 2.5, which contains the spectrograms for *kanón?* – *cánon?* (with question intonation), the difference between stressed /ˈaː/ and unstressed /aː/ can easily be seen. Sluijter (1995) also finds that the spectral tilt consistently correlates with lexical stress.

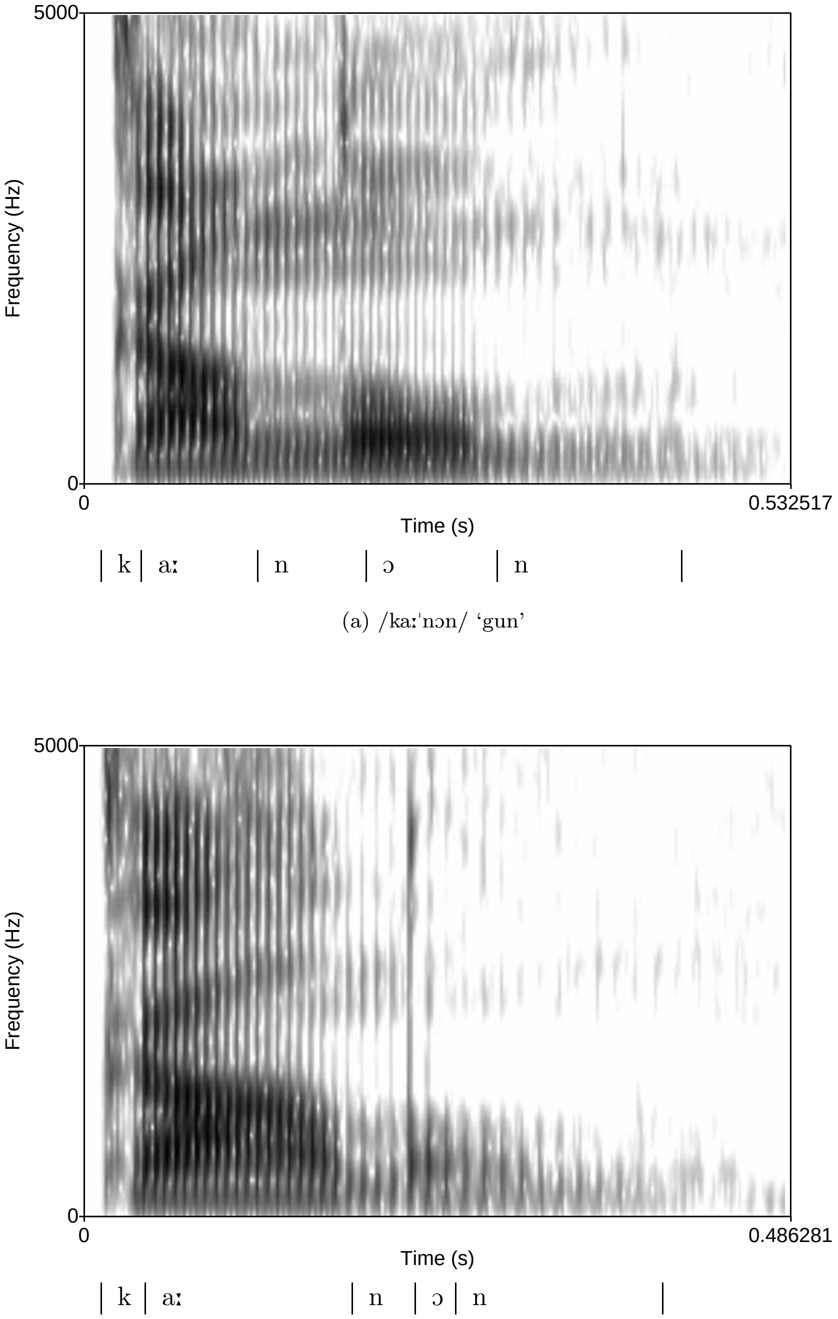
Note also that the occurrence of more high frequencies seems to coincide with increased overall energy. The actual correspondence lies in the amount of 'effort' that goes into pronouncing the vowel. Sluijter (1995) has this tying in nicely with "older literature" that claimed that stressed syllables were louder: "A stressed syllable might be perceived as louder, and therefore more prominent, than an unstressed one due to the increased intensity levels in the higher part of the spectrum."

2.1.2.6 *Consonants*

So far, consonants did not enter the picture. Consonants are generally disregarded in the literature about stress. This may be because vowels are the most noticeable[14] part of syllables, and they most strikingly carry acoustic information about stress. However, the stress property of all segments in a syllable should match. For example, in /kaːnɔn/ either both /k/ and /aː/ are stressed, or they are both unstressed. All segments in the second syllable /nɔn/ should be unstressed if the first syllable is stressed, and they should be stressed if the first syllable is not. This may be perceived as difficult to model in an automatic speech recognition system (see section 4.3.1 on page 43.)

However, there *is* information available in the literature about the influence of another type of 'sloppiness' on consonants. Van Son and Pols (1996) recorded

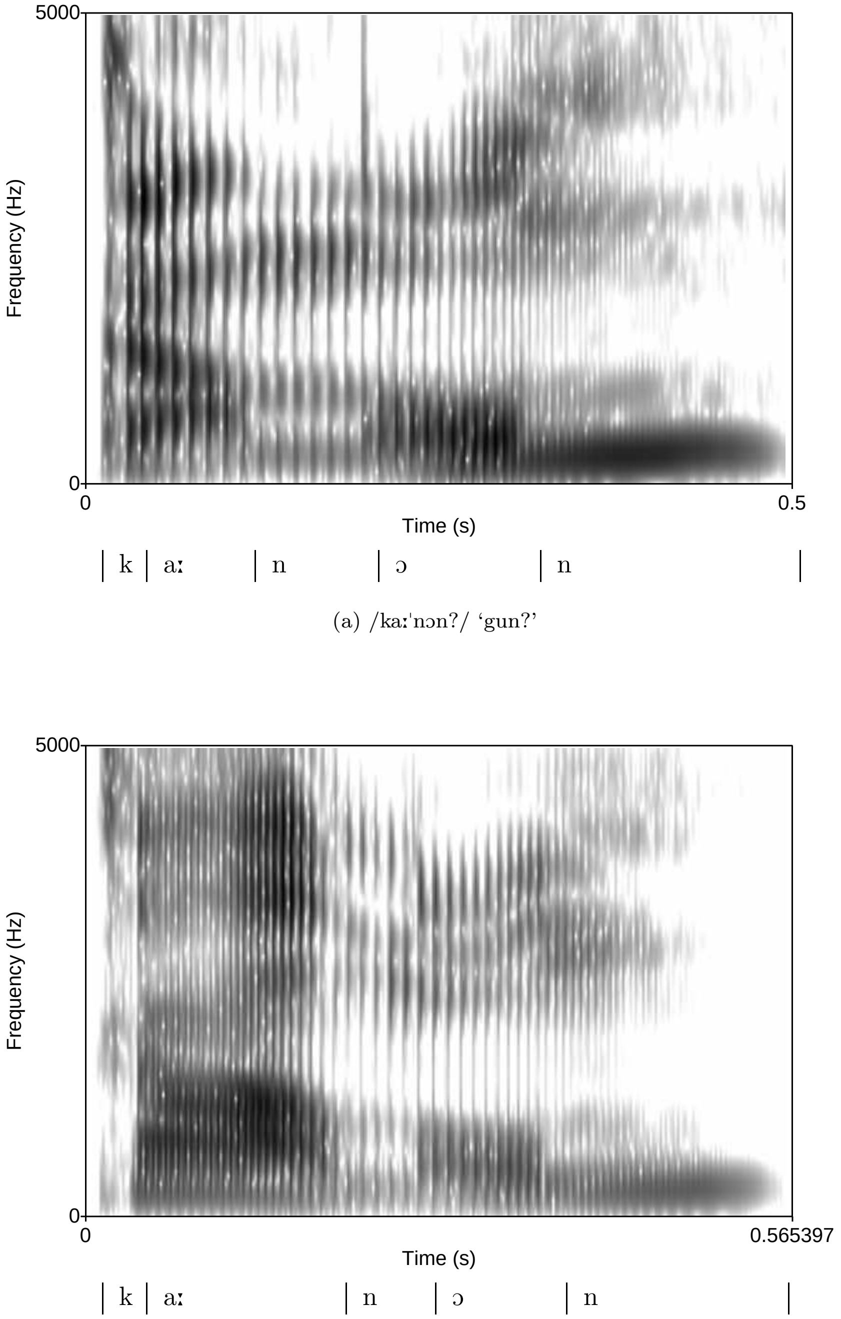
[14] With a phonological term, the most *sonorous*.



(a) /kaːˈnɔn/ ‘gun’

(b) /ˈkaːnɔn/ ‘canon’

Figure 2.4 *Spectrograms for the* kanón – cánon *minimal pair. That stressed syllables have more high frequencies than unstressed ones is easily seen in the spectrum for /ɔ/.*



(a) /kaːˈnɔn?/ ‘gun?’

(b) /ˈkaːnɔn?/ ‘canon?’

Figure 2.5 *Spectrograms for the* kanón? – cánon? *minimal pair. Even with different intonation /ˈɔ/ has more higher frequencies than /ɔ/.*

a newscaster telling anecdotes and afterwards reading out aloud the transcription of what he had produced spontaneously, so that they had two recordings of the same text with different speaking styles. In his spontaneous speech the consonants were pronounced more 'sloppily' in three ways:

1. Reduction in terms of formant frequency lowering. This is essentially the kind of thing we've seen in 2.1.2.4, applied to consonants rather than to vowels.

2. "Consonant realizations shorten like vowels". This is clearly the 'duration' property of 2.1.2.3.

3. "Except for the plosives, all consonants and vowels showed a decrease in [centre of gravity]." This is about the spectral tilt that was discussed in 2.1.2.5 (the "centre of gravity" term has to do with yet another way of looking at the diagrams, but they mean 'spectral tilt'). Plosives are consonants like /k, p, b, t/, for which this feature is apparently not relevant. For all other consonants, like /v, f, ɣ, x, ʋ, r/ (as in ***v**oet*, ***f**out*, ***g**ra**ch**t*, ***w**at*, ***R**ogie**r***), spectral tilt was relevant.

Assuming these features can be used for extracting stress too, it seems like precisely the same features we use for consonants can be used for vowels. This is fortunate because it is not really possible in speech recognisers to use some acoustic features for vowels only, and other features for consonants only.

Now, is consonant reduction in spontaneous speech similar to consonant reduction in unstressed syllables? Is the newscaster's spontaneous speech really relevant for recognition of lexical stress? There is no way to know for sure at this moment. We should hope so, because this would enable the speech recogniser to use the same machinery for consonants as for vowels. And even then it will probably be more difficult to extract stress information for consonants than it is for vowels. However, any extra information we can give the computer may be useful, as long as it can work with the uncertainty. See section 5.6 on page 61 for information about incorporating consonantal features in the recogniser.

2.2 Automatic speech recognition

The assumption underlying automatic speech recognition is that the speech signal is the realisation of a message consisting of a sequence of symbols. The challenge is to perform the reverse operation of this encoding. The speech signal is first converted to sequence of *feature vectors*. The vectors typically contain Mel-frequency cepstral (MFCC) or linear predictive coding (LPC) coefficients. The MFCC transformation is based on the Mel scale, which models human perception of frequencies. LPC "provides a complete model for speech production, and it analyzes the speech signal such that the characteristic effects of the vocel tract and its formants can be separated from the excitation" (Paulus and Hornegger 1998, p. 309). Values are extracted every 10 ms, which is short enough to make the assumption that the feature information is stationary.

2.2.1 *Words and phonemes*

An assumption often made for *continuous speech* is that utterances contain words. This is usually a valid assumption, though the term "word" may not correspond to the meaning it has in spelling. (Does *New York* really consist of two words? Is *can't* really one word?) It also sometimes means relevant morphological information is omitted (e.g. inflections in wordforms like *oak* and *oaks* are different words). Typically, all words are contained in the *lexicon.* The same term is used for the word list in human beings' heads. Note that the lexicon in automatic speech recognition is similar to the Minimalist Program's (Chomsky 1995) lexicon[15]: both contain all inflections of all words. A speech recogniser's lexicon contains entries like (11) (notation adapted from the CGN lexicon; see section 3.3.2 on page 32).

(11)	*aambeeld*	'anvil'	ˈaːmbeːlt
	dankbaar	'thankful'	ˈdɑŋgbaːr
	denken	'to think'	ˈdɛŋkə
	gezaagd	'sawn'	xəˈzaːxt
	incheckbalie	'check-in'	ˈinʃɛgbaːliː
	makkelijke	'easy'	ˈmɑkələkə
	melk	'milk'	ˈmɛlk
	postbode	'postman'	ˈpɔzdboːdə
	projectbureau	'project agency'	proːˈjɛgdbyroː
	rasecht	'genuine'	ˈrɑsɛxt
	weekdag	'weekday'	ˈʋeːgdɑx

The symbols in the third column contain transcriptions for the words whose regular spelling is given in the first column. These consist of strings of phonemes (or something similar) that are used in speech recognition. An assumption that is often made is that these phonemes are always pronounced in a similar way. This assumption is not phonologically correct. Though much of the juggling around with sounds throughout the centuries has had effects on the phonemes (e.g. Dutch *logisch* 'logical' from /ˈloːxiʃ/ has become /ˈloːxis/), how the words are actually formed (their *phonetics*) is decided from their internal representation at the moment they are pronounced.

The transcriptions in the lexicon are the results of half-hearted attempts to encode the words' phonetics. Thus, the proper phonemic transcription for *postbode* is /ˈpɔstboːdə/ (according to Geeraerts 2000), much like the spelling suggests. The transcription in the lexicon through a subset of the rules that may apply has become [ˈpɔ**zd**boːdə], whereas the proper phonetic transcription is usually [ˈpɔ**s**boːdə], which may become [ˈpɔ**z**boːdə]. Something similar happens to *weekdag* /ˈʋeː**k**dɑx/, which is usually pronounced [ˈʋeː**g**dɑx], where /k/ is realised as [g]. However, /g/ is not a phoneme of Dutch,[16] so it cannot be represented in a lexicon with phonemic transcriptions. On the other hand, *rasecht* may become [ˈrɑ**z**ɛxt] (Collins and Mees 1999), but the lexicon does not encode this. *Melk* 'milk' /ˈmɛlk/ is often realised as [ˈmɛl**ə**k] (Warner *et al.* 2001); this

[15] The Minimalist Program is a framework for syntax that emphasises economy contraints more than previous theories. This supposedly means that it is more likely to be correct from a cognitive psychology point of view; the same should then go for lexica in speech recognition.

[16] /g/ may be seen as a *marginal* consonant. Some speakers use it in foreign words like *goal* /goːl/, whereas others use /k/ and say /koːl/.

is not encoded either. Often, phonological rules apply across words: *should you* /ˈʃʊd**ju**ː/ may become [ˈʃʊd**ʒ**uː]. These rules cannot be applied in a lexicon containing only words at all (but see Kessens *et al.* 1999, for an approach where 22 oft-occurring multi-words sequences are explicitly included in the lexicon to yield an impressive increase in performance).

A word for which a full phonetic transcription would be particularly useful is *aambeeld*. This will usually be pronounced as [ãːmbeːɫt], with a nasalised [ãː] (the spectrum of which is noticeably different from the spectrum for [aː]) and a velarised [ɫ] (which sounds more like [ə] than like [l]). A word for which the phonemic transcription may be too much modified is *denken*. Geeraerts (2000) gives /ˈdɛŋk**ə(n)**/. The final /n/ is usually left out in ABN, the Dutch prestige dialect, to form [ˈdɛnk**ə**], but Eastern and Northern dialect speakers may pronounce it [ˈdɛnk**ŋ̍**].

Though phonemic transcriptions are used in speech recogniser lexica, in practice the transcriptions are modified to include some phonetic effects in an arbitrary way. Kessens *et al.* (1999) add pronunciation variants, generated through a few phonological rules, to a Dutch lexicon and see recognition improve. The funny thing is, they use the CELEX lexicon with its /ˈpɔzdboːdə/ transcription, which is probably not effected by their phonological rules: /d/ is not deleted, while /t/ in *rechtstreeks* 'straight away' /rɛxtˈstreːks/ [rɛxˈstreːks] is. On the other hands, psycholinguistic (Pierrehumbert 2000) and phonetic (Chang *et al.* 2000) research suggests that more complex patterns than straightforward phonological rules are used by humans.

2.2.2 *Probabilistic speech recognition*

In a probabilistic sense the goal of speech recognition can be formulated as follows (Jurafsky and Martin 2000):

(12) What is the most likely sentence out of all sentences in the language L given some acoustic input O?

Consecutive slices from the acoustic input (for example, feature vectors), the observations, can be called

$$O = o_1, o_2, o_3, \ldots, o_T \tag{13}$$

The sequence of words of a actually pronounced sentence is called

$$W = w_1, w_2, w_3, \ldots, w_N \tag{14}$$

The probabilistic implementation of the goal (12) can be expressed as

$$\hat{W} = \operatorname*{argmax}_{W \in L} P(W|O) \tag{15}$$

Through Bayes' rule, (15) can rewritten as

$$\hat{W} = \operatorname*{argmax}_{W \in L} \frac{P(O|W)P(W)}{P(O)} \tag{16}$$

$P(W)$ is called the *language model* (often approximated straightforwardly through a word n-gram model). $P(O|W)$ gives the *acoustic model*. $P(O)$ is always the

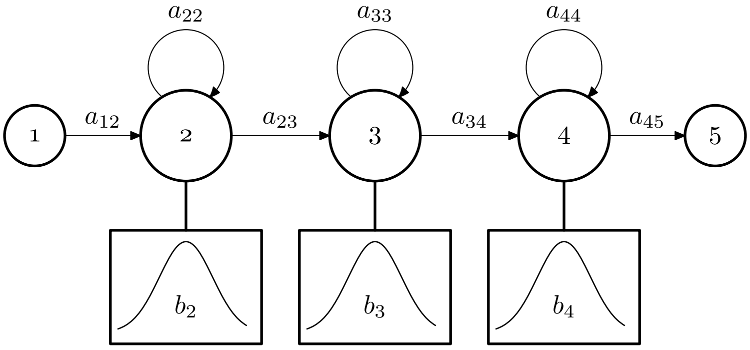


Figure 2.6 *A three-state* HMM*. States 1 and 5 are* non-emitting *states: they exist to facilitate concatenation of* HMM*s and do not themselves model acoustic input.*

same for a given O; thus, $\hat{W}$ can be found by calculating

$$\hat{W} = \underset{W \in L}{\operatorname{argmax}}\ P(O|W)P(W) \tag{17}$$

2.2.3 *Hidden Markov Models*

Having simply, but incorrectly, reduced phonemic and phonetic levels to one, every phoneme can be modelled separately. Phonemes' acoustic properties are usually modelled using *Hidden Markov Models* (or HMMs). These are models that use Markov chains, which contain a number of states with the probabilities of the transition from one state to another. For speech recognition, the actual transition probabilities depend on the input signal.

Figure 2.6 depicts an HMM with three states, let us say for the phoneme /aː/. The model depicted only has transitions from one state to itself and the next, but a more complex topology is possible. The idea of this Markov model is that the properties of the first part of the vowel are captured by state 2, the middle part by state 3, and the final part by state 4. If the middle part of the vowel is the longest, the probability of the next state being state 3 if it is also the current state (i.e. of remaining in state 3), a_{33}, will be large. In this way the probability of lingering longer in one state is increased.

But how is it decided whether an input fits a state? Every state has a function $b_i(s)$ that returns the probability that the observation s matches the state. If a sequence of observations $\mathbf{O}$ in the form of feature vectors $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$ is given, the probability of it being the encoding of the symbol /aː/ may be calculated by summing over the probabilities of all possible state sequences $X = x(1), x(2), \ldots, x(T)$ that start in state 1 and end in state 5:

$$P(\mathbf{O}|M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(\mathbf{o}_t) a_{x(t)x(t+1)} \tag{18}$$

The description has so far used a three-state HMM, though it can be any length. A motivation for using three states may be found in the supposed

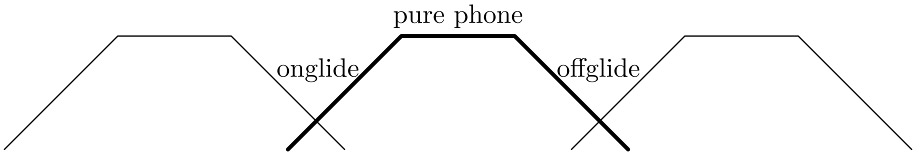


Figure 2.7 *The three phases of a phone, conceptionally.*

structure of phonemes: coarticulation effects may be seen as gestural overlap that results in uncertainty about the exact realisation of the phoneme at the start and end (see figure 2.7, after Wiggers 2001). The canonical form (as seen in textbooks) of an HMM has three states.

A problem is that a particular symbol can have different acoustic realisations in different circumstances (depending on, for example, speaker, mood, gender, environment). Another problem is that the boundaries between symbols cannot be explicitly identified from the speech waveform. Only if the boundaries are known (e.g. when the input is known to consist of exactly one word from a lexicon) can all possible interpretations of the input be iterated over.

2.2.4 *The Viterbi algorithm*

Now that the acoustics of phones on a 10 ms time slice are modelled, how to use the information about all phone models at all times to decode the words of a whole utterance? The problem is that decoding a sequence of phonemes means checking probabilities for *all* possible sequence, with the possible exception of those that can not be obtained through concatenating words from the lexicon. This is basically what the *forward algorithm* does (see Jurafsky and Martin 2000). Using that algorithm for continuous speech, however, is not feasible.

The *Viterbi* algorithm uses dynamic programming: it makes the assumption that if the best path to point Q goes through P, it contains the best path to point P. Applied to the speech recognition, Viterbi finds the maximum likelihood state sequence rather than the total likelihood $P(\mathbf{O}|M)$. Instead of using the sum of probabilities it uses the maximum. The following explanation has been adapted only cosmetically from Young *et al.* (2002):

> For a given model M, let $\phi_j(t)$ represent the maximum likelihood of observing speech vectors $\mathbf{o}_1$ to $\mathbf{o}_t$ and being in state j at time t. This partial likelihood can be computed efficiently using the following recursion:
>
> (19) $$\phi_j(t) = \max_i \{\phi_i(t-1)a_{ij}\} \, b_j(\mathbf{o}_t).$$
>
> where
>
> (20) $$\phi_1(1) = 1$$
>
> (21) $$\phi_j(1) = a_{1j}b_j(\mathbf{o}_1)$$

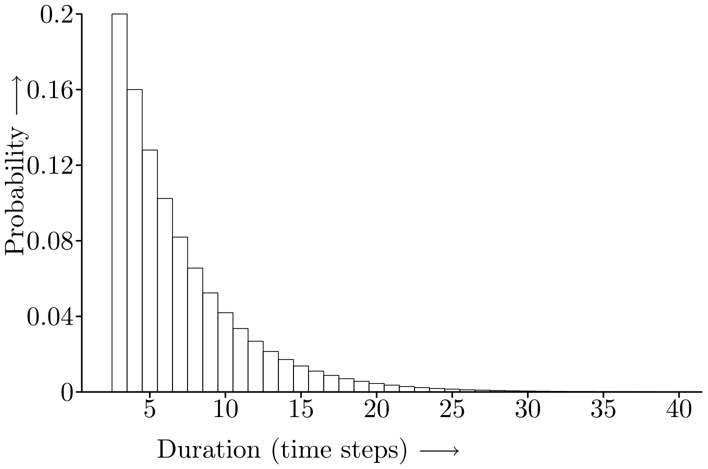


Figure 2.8 *The duration modelled by one HMM state.*

for $1 < j < N$. The maximum likelihood $\hat{P}(O|M)$ is then given by

$$\phi_N(T) = \max_i \{\phi_i(T) a_{iN}\} \tag{22}$$

This recursion forms the basis of the Viterbi algorithm.

2.2.5 *Duration*

Section 2.1.2.3 on page 19 has shown that duration is an important correlate of stress. However, HMMs do not model duration explicitly; the parameters that come nearest are the transition probabilities a_{ij} (as in figure 2.6 on page 26) that influence how often self-transitions occur. Many authors have tried to come up with variants of Hidden Markov Models that describe duration better: Russell and Moore (1985) come up with Semi-HMMs; Ramesh and Wilpon (1992) think of Inhomogeneous HMMs; Sitaram and Sreenivas (1997) after a lucid introduction into the need for modelling duration in HMMs introduce the Trend-HMM and then take the reader on a rambling ride to all kinds of combinations of T-, I-, LC- and S-HMMs. Another overview, one claiming that explicit duration modelling is not needed, is given in Wang *et al.* (1996*b*).

The usual method of depicting duration as modelled by HMMs is by calculating the probabilities of reaching a specific state at a specific time, given that the probability of being in the first state at $t = 1$ is 1.0. A graph for the probability of exiting one HMM state is given in figure 2.8. As the probability of remaining in the state decreases with a constant factor over time, the result looks like a geometric distribution (Ross 1997).

So what is the problem? HMMs are memoryless, but model time through transition probabilities. Sitaram and Sreenivas claim that from statistics it seems that actual phone duration is gamma-distributed (see Wang *et al.* 1996*a*, and section 6.1.1 on page 70 for my measurements). They shrewdly mix up phone duration and state duration in their argument, but *what* they argue — duration of parts of phonemes, as represented by HMM states, should be modelled — seems sensible, especially for diphthongs.

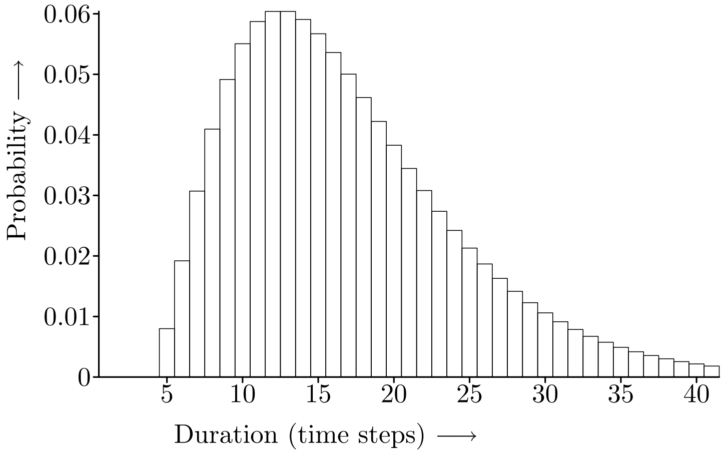


Figure 2.9 *The duration modelled by a three-state* HMM *(see figure 2.6 on page 26 for a depiction) with $a_{ii} = 0.8$, which is binomial-like.*

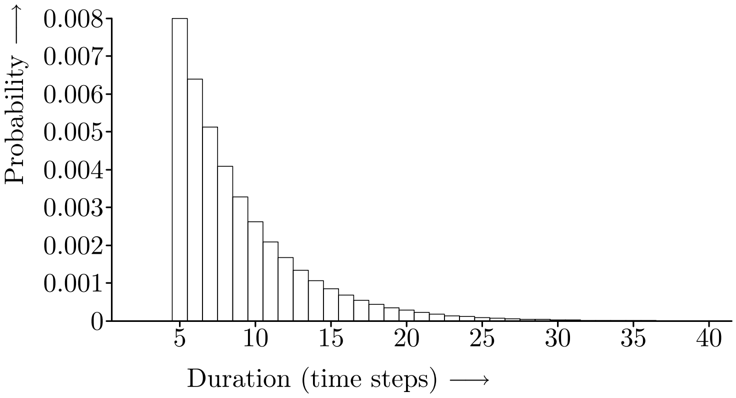


Figure 2.10 *The duration modelled by the same three-state* HMM *as in figure 2.9, now processed by the Viterbi algorithm. It is similar to the one-state case (see figure 2.8 on the preceding page).*

Wang *et al.* (1996*b*) look only at the phone model as a whole and conclude that an HMM is perfectly well able to model duration, as long as their algorithm is used for training rather than the standard Baum-Welch one. Bilmes (2002) shows that a sequence of HMM states is like a negative binomial distribution. This is a discrete version of the gamma function, which seems to be the proper distribution for phone duration modelling. The graph in figure 2.9 shows this as well. The number of paths that can be taken to end up in state 5 of the model is much larger for $t = 13$ than it is for $t = 5$, though the probability of each of these paths decreases. Thus, it seems like an HMM can model any duration by giving it enough states, as Bilmes claims. Neither Wang *et al.* (or Wang 1997) nor Bilmes seem to realise that there is more to this story than HMMs' transition probabilities.

The Viterbi algorithm messes things up. The *forward algorithm* (e.g. see Jurafsky and Martin 2000, p. 172) uses the sums of all possible paths. Viterbi search however uses only the most probable path to a point (see section 2.2.4 on page 27). Figure 2.10 on the page before shows the effect of the dynamic programming assumption on the duration: it is not now binomial anymore; rather, we are back at square one, with an exponential-like duration.

The influence of the phone likelihood estimation is disregarded. This estimation not only influences the duration of one state; if the length of the onset and the offset of a phone (as in figure 2.7 on page 27) is only a frame or so, this will effectively turn the duration modelled by a three-state HMM as a whole back into an exponential-like one, with a constant time added. Moreover, the effects of the state transitions a_{ij} should be assessed relative to the effect of the Gaussians b_i in equation (18) on page 26: if the influence of the Gaussians varies in terms of thousands, transition probabilities in terms of tenths will not be able to change the scene much. This is the scenario that will most often occur in practice.

The speaking rate is not taken into account. While recognising an utterance, the speaking rate of previous utterances or words must be used to properly account for phone duration.

Though some researchers are inclined to believe that HMM transition probabilities can help modelling duration, the Viterbi algorithm and the Gaussians undo the theoretical advantage. Modelling phone duration using only ordinary HMMs and the Viterbi algorithm remains impossible without adding HMM states in numbers in the order of the maximum phone length that is to be modelled.

Chapter 3

Tools

τύνη δ' ὤμοιιν μὲν ἐμὰ κλυτὰ τεύχεα δῦϑι[1] Homer, *Iliad* XVI

3.1 Hidden Markov Toolkit

To implement a speech recogniser from scratch would take too much time; however, there are toolkits that can help constructing a recogniser. I use the Hidden Markov Toolkit, or HTK, which is documented in Young *et al.* (2002). Wiggers (2001) has an introduction to its architecture and purpose:

> HTK is a portable software toolkit for building and manipulating systems that use continuous density Hidden Markov models. It has been developed by the Speech Group at Cambridge University Engineering Department.
>
> HMMs can be used to model any time series and the core of HTK is similarly general purpose. However, HTK is primarily designed for building HMM based speech processing tools, in particular speech recognizers. In can be used to perform a wide range of tasks in this domain including isolated or connected speech recognition using models based on whole word or subword units, but it is especially suitable for performing large vocabulary continuous speech recognition.
>
> HTK includes nineteen tools that perform tasks like manipulation of transcriptions, coding data, various styles of HMM training including Baum-Welch re-estimation, Viterbi decoding, results analysis and extensive editing of HMM definitions. HTK tools are designed to run with a traditional command-line style interface, each tool has a large number of required and optional arguments and most tools require one or more script files. …
>
> Although this style of command-line working results in complex commands, that actually have more resemblance with programming lan-

[1] 'Nevertheless, now gird my armor around your shoulders' (translation Samuel Butler). τεῦχος 'a tool, implement' (<http://www.perseus.tufts.edu/>).

> guages than with commands, it has the advantage of making it simple to write shell scripts or programs to control HTK tool execution. Furthermore it allows the details of system construction or experimental procedure to be recorded and documented.

It is clear that for a work like this, where an off-the-shelf speech recogniser would not do, the HTK is a suitable toolkit.

3.2 Praat

Praat (Boersma 2001) is a program for speech analysis and synthesis. Its user interface provides a variety of functions for speech recordings: the user can perform many types of acoustic analyses, transcribe recordings and even make Optimality Theory grammars. Its functions are also available from a scripting interface; scripts can be recorded. This combination makes it possible to quickly move from editing sound files by hand to an automatic workflow. The program is free and can be found at <http://www.praat.org/>

3.3 Corpora

A speech recogniser must be trained on audio recordings. The audio recordings should come with orthographic or phonetic transcriptions. Two corpora are used in this work: the small DUTAVSC, and the large CGN.

3.3.1 DUTAVSC: *Dutch Audio-Visual Speech Corpus*

The Dutch Audio-Visual Speech Corpus (dutavsc, see Wojdel *et al.* 2002), was originally recorded to provide a combined audio-visual corpus for combining lip-reading with speech recognition. It contains prompts derived from the POLYPHONE dataset. POLYPHONE is telephone-recorded, however, whereas the DUTAVSC has high quality recordings. The corpus contains read words, phonetically rich sentences, digits, and spellings. As it is rather small, and for technical reasons (see section 5.9 on page 64), the corpus was used for initial experiments.

3.3.2 CGN: *Corpus Gesproken Nederlands*

The CGN corpus (Oostdijk 2000) is a nine-million-word corpus of contemporary standard Dutch as spoken in the Netherlands and Flanders. It contains about 1000 hours of speech. The entire corpus has been transcribed orthographically and some phonetic transcriptions, part-of-speech tagging, and prosodic annotations are available for parts of the data. (None of these contain stress information though: the prosodic annotation only indicates sentence-level accents.)

The corpus aims to serve different people's purposes. It could be used to train speech recognisers (as in this work), but it also provides data for linguistic research of various kinds. It contains various kinds of recordings: conversations (face-to-face and telephone), discussions, lectures, radio and television programmes, and read speech.

The transcriptions are linked to the lexicon the CGN provides. The lexicon is produced from various sources, and includes different phonetic transcriptions for Dutch and Flemish. Not all kinds of information are available for all words, however: for example, the only transcriptions encoding lexical stress are from the CELEX corpus. This corpus contains only a subset of the total CGN lexicon.

Chapter 4

Model

Still a man hears what he wants to hear and disregards the rest
Paul Simon, *The Boxer*

The previous chapters have given a more or less uncontroversial overview of the relevant literature. From here on all material will be new — and possibly more controversial. After a short overview of the apparent shortcomings of the literature, a conceptual model will be presented. This will enable a constructive way of deciding how to use lexical stress.

4.1 Previous work

Figure 4.1 on the next page shows the relevant components of a typical speech recognition system. It must be considered *where* lexical stress information can be used in a speech recognition system to increase recogniser performance. Earlier work has not always used the properties of lexical stress to increase the recognition rate. Van Kuijk and Boves (1999) merely examine the acoustic properties of telephone speech. Xie *et al.* (2004) have an automatic recogniser check language learners' stress patterns: the words are known in advance. Bouwman and Boves (2001) use the stress pattern only in a second pass, after the recogniser is done, for utterance verification. They use the syllable length and probabilities of phonemes having been recognised incorrectly to build a confidence measure. This approach is depicted in figure 4.2 on page 37.

Improving reliability modelling is a clear objective. Researchers that try to use lexical stress information in an earlier stage (van Kuijk *et al.* 1996; Wang and Seneff 2001; van den Heuvel *et al.* 2003) do not specify as clearly what their objective is. They probably want to improve recogniser performance, but how this is to be effected is not discussed. It is not clear what information authors envision to have what influence on the recognition at what stages: "distinguishing stressed and unstressed vowel models may have a general impact on recognition results" (van den Heuvel *et al.* 2003) is one of the more enlightening comments in this respect.

What one cannot help noticing when looking at the literature on automatic lexical stress recognition is a total disregard of consonants. It is known from phonology that lexical stress has a profound influence on the realisation of

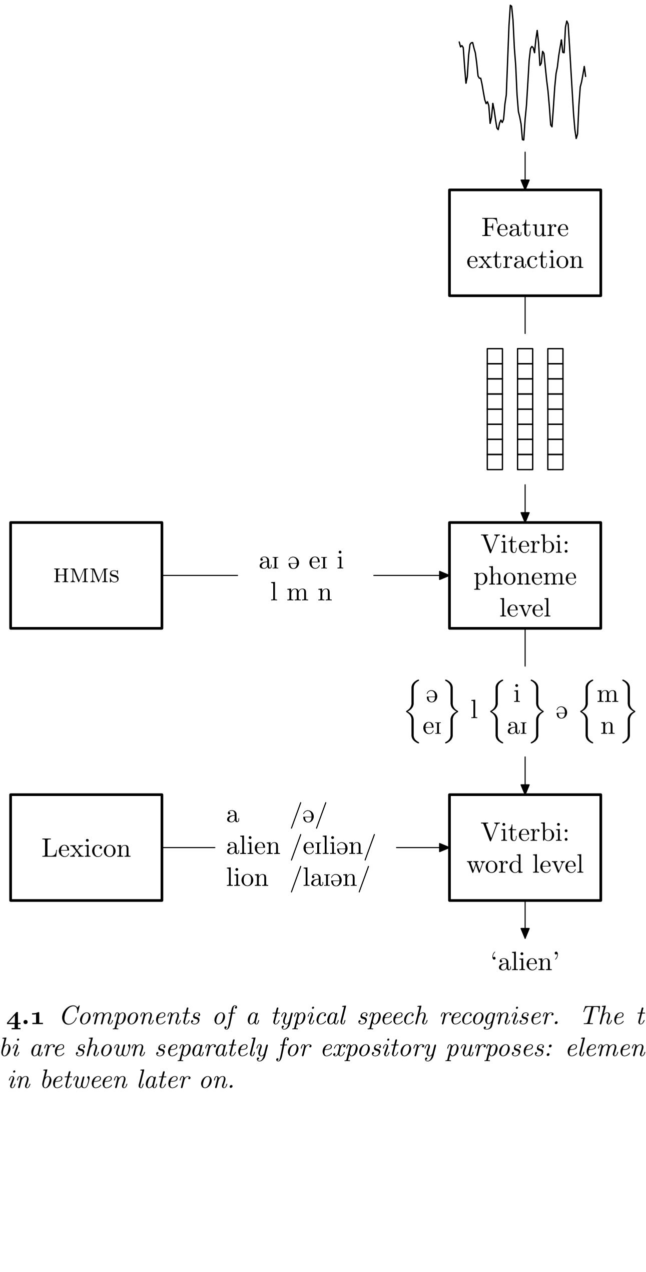


Figure 4.1 *Components of a typical speech recogniser. The two levels of Viterbi are shown separately for expository purposes: elements will be inserted in between later on.*

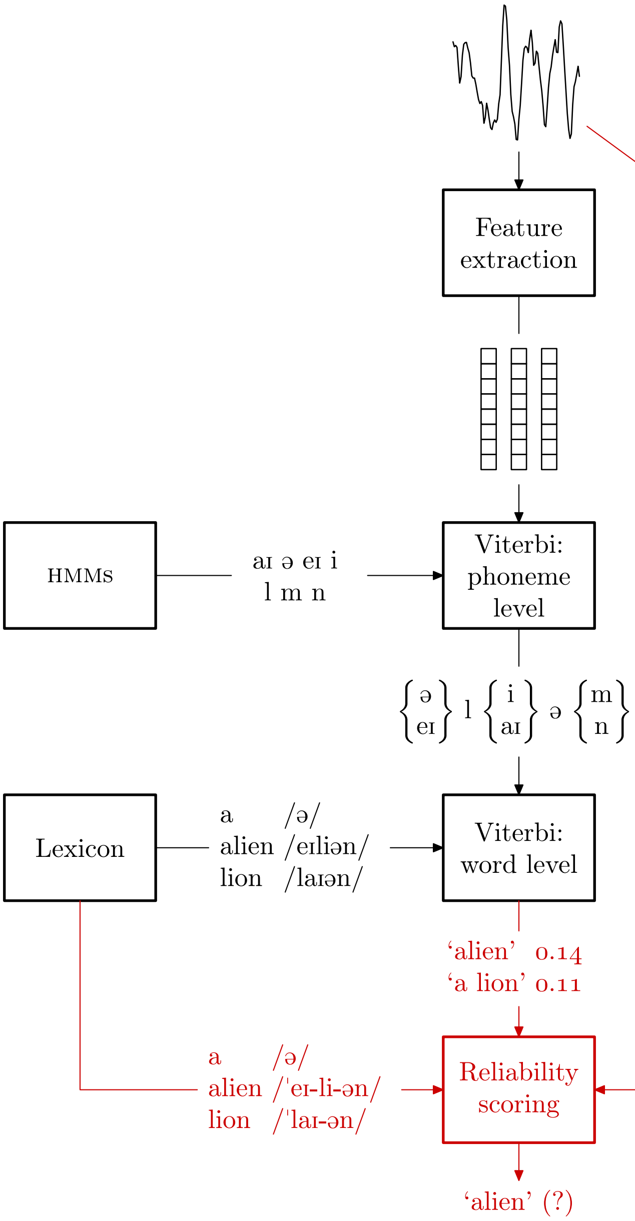


Figure 4.2 *Bouwman and Boves's (2001) approach to using lexical stress. The new parts are coloured red. Lexical stress is used only for confidence measures, after recognition has taken place. In the example 'alien' is probably wrong; hence the question mark.*

phonemes (see Ewen and van der Hulst 2001, and section 2.1.1.1 on page 10); van Son and Pols (1996) show how consonants are influenced by speaking style; Sluijter (1995, p. 33) calculates duration over the whole of a syllable, including consonants; Greenberg *et al.* (2003) show that consonants in stressed syllables have longer durations. It seems this knowledge is not used in lexical stress recognition, nor is it regarded, nor is phonological literature cited.

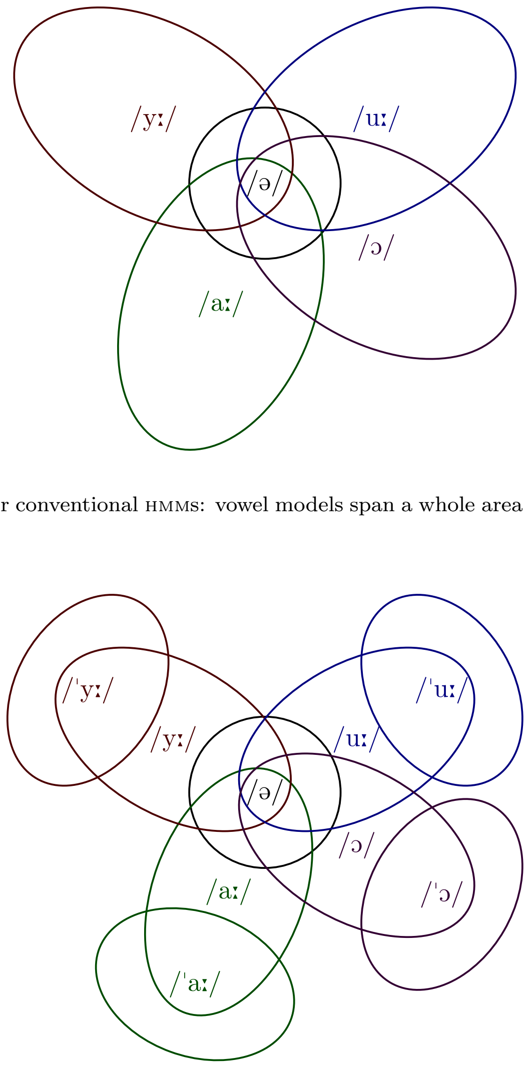
Wang and Seneff (2001) choose a strategy where "[o]nly syllable nucleus vowels are scored by the lexical stress models: for segments that do not carry lexical stress, such as consonants and silences, the stress scores are simply ignored." That consonants should not carry stress is not true. It may be viewed as a half-truth, though: the sonority peak of syllables is on the nucleus (see section 2.1.1.1 on page 11). Still, in other work on automatic recognition of lexical stress not even shaky motivations can be found.

Van den Heuvel *et al.* (2003) use context-dependent models, which appears not to bring about any recognition improvement compared to the context-independent ones. A swap condition (where stress markers are swapped) is used, which does degrade recognition. This is taken to show "that the models indeed capture some stress-related information". When the swap condition yields worse results with context-dependent models than for context-independent models, the obvious conclusion (not made explicitly in the article) is that context-dependent models capture more stress-related information. This is speculated to be because "context-dependent acoustic models may have been better tuned for lexical stress (in a spectral sense)"; no arguments for this are given though. Let me put forward more speculation: context-dependent models take parts of consonants into account; maybe *they* contain relevant information. The article features an abundant amount of discussion about which consonants come with which stressed and which unstressed vowels; why are they not used, and why is there not even speculation on their use?

Earlier efforts to model stress in speech recognisers seem to be driven by the fact that there is stress information that is not encoded anywhere, and thus must be thrown in: the *technology push* at work. Now there is not much wrong with using the amount of information as a motivation for research, but an issue as important as what information is available is how that information can be used to improve recognition. That is the side of the chain where we will start, and then we will link that to what information is needed; an implementation will follow after that. First presenting a model of how stress recognition could work has the advantage of being able to compare the implementation with the model and see what things it can and what things it cannot model.

4.2 Objectives

To determine what aspects of lexical stress are important to use in a speech recogniser, its conceived effects on continuous speech recognition must be clear. These objectives will form the motivation for the model considerations in section 4.3 on page 42. Imagine a speech recognition system that recognises whether phonemes are stressed. This means that stress information is incorporated in the acoustic model ($P(O|W)$ from equation (17) on page 26). What improvements should we expect from such a system compared to a conventional system?



(a) Vowel scope for conventional HMMs: vowel models span a whole area from fully realised to schwa-like.

(b) Vowels are split into stressed and unstressed versions. Stressed vowels, which are not usually reduced have a full realisation, are distinguished from unstressed ones. This does not eliminate overlap, but it does increase accuracy.

Figure 4.3 *Conceptual depiction of the scope of vowel models. Modelling lexical stress makes the phoneme models more accurate.*

Phone model accuracy Several studies in speech processing and phonetics literature have strived to model phonemic and phonetic properties more closely (for example Kessens *et al.* 1999; Greenberg 1999; Greenberg *et al.* 2003). The idea is that this will decrease the scope that the phone models (e.g. the HMMs) have to cover by providing more accurate and more specialised examples for training. The same goes for lexical stress. For example, a recogniser trained on lexical stress should have phone models for stressed /ˈɔ/ and unstressed /ɔ/ that cover two sub-areas of the former /ɔ/ model. The variation within the phone model becomes smaller, which should improve performance. Figure 4.3 on the page before shows this.

Word segmentation Lexical stress information should be used at an early stage — when deciding the word segmentation. For example, the Dutch phrase *aan bód* 'first in line' may be distinguished from *áanbod* 'offer'. The advantage for word segmentation really starts to kick in for languages that have fully regular stress patterns. In Icelandic, Hungarian, and Czech, for example, the stress is always on the first syllable of the word. In languages that have stress on a fixed syllable of the word (be it counted from the first or from the last syllable) stress assignment has a *demarcative* function (Ewen and van der Hulst 2001). English, like Dutch, has strayed from the Germanic Stress Rule, which stressed the first syllable of every word. The *stress domain* is still the word, and it appears that enough English content words have stress on the first syllable to make listeners use the stress pattern for segmentation (see Harley 2001, p. 221). This is something a speech recogniser could also do.

Application of phonological rules Many phonological rules require information about the stress of a syllable. Aspiration of English fortis stops (see section 2.1.1.1 on page 10), which is motivated by stress, exemplifies a more general rule that is particularly useful in speech recognition. Realisations of phonemes in stressed syllables appear to be 'stronger' than in unstressed syllables. Shoup (1980, p. 135) says that English vowels

> becme [*sic*] reduced to a more central vowel position when unstressed or with reduced stress, for example [ritɝn] for *rètúrn* and [rɪtɝn] or [rətɝn] for *rĕtúrn*. Schwas can actually be eliminated when they occur in unstressed positions, such as *chocolate* becoming *choclate*. Syllable-final schwa followed by [l] or a nasal can become syllabic [l̩] or nasal, such as in *bottle* [bɑtl̩] or *button* [bʌtn̩].[1]

Greenberg (1999) analyses the pronunciation variation of both vowels and consonants in the Switchboard corpus and finds that "canonical pronunciation and prosodic stress [...] often travel together". Later research (Greenberg *et al.* 2003) shows that unstressed segments tend to have a shorter duration. Phonemes in unstressed syllables tend to have more reduced realisations than those in stressed syllables. With knowledge about stressed, a speech recogniser can make allowance for more reduced realisations of phonemes in unstressed syllables while requiring fuller realisations

[1] Notation altered slightly to agree with this work's conventions.

in stressed positions. This may especially improve accuracy for recognition of fast spontaneous speech.

Kessens *et al.* (1999) do use phonological models to account for pronunciation variation, but they explicitly state they leave out the distinction between stressed and unstressed segments because of limitations of their phone set. As an example of the different behaviour of /r/ in stressed and unstressed contexts their recogniser cannot model they give *Arnhem* 'Arnhem (city)' /ˈɑrnɛm/ *[ˈɑnɛm][2] as opposed to *Leeuwarden* 'Leeuwarden (city)' /ˈleːʋɑrdən/ [ˈleːʋɑdən].[3]

Word category There are pairs of words with the same segmental content, where the stress pattern signals that they have different categories. Because of the stress assignment rules of English, it has hundreds of verb – noun pairs like *subjéct – súbject* (Higgins 2000). Recognition of those minimal pairs is impossible without knowledge of lexical stress. Note that if the recogniser recognises all other words in the utterance correctly, it might be able to deduce the correct category without knowledge about the stress pattern: in text sentences like "The subject of this thesis is lexical stress" do not usually lead to confusion.

Semantic difference Words with the same segmental content but different morphological structure often have different stress patterns. Thus, we find Dutch *overdríjven* 'exaggerate' – *óverdrijven* 'float by'. If a speech recogniser does not recognise the stress difference in this case, the ambiguity has to be solved at the semantic level. Still, sentences like 'Je kunt ook overdrijven' cannot unambiguously be parsed without proper knowledge of the context, or the stress pattern. Since automatic syntactic and semantic parsing is far from perfect yet, any information about lexical stress should be most welcome.

Stress patterns on higher levels than the word, though not being 'lexical stress' they do not strictly belong in this work, may have an influence on speech recognition as well. Some areas in which stress information can be used in the future are:

Phrasal properties Stress assignment at a higher level than the word level may be used to express both a syntactic and semantic difference that is not clear otherwise. For example, in an earlier version of this thesis this enumeration of objectives used the phrase "sound knowledge". It was removed because in writing it was not clear that I meant 'proper knowledge' /ˌsaʊnˈnɒlɪdʒ/ rather than 'knowledge about sound' /ˈsaʊnˌnɒlɪdʒ/, which could have made sense in the context of this thesis as well. An advanced speech recognition system could interpret stress patterns like this and distinguish between the former version, with *sound* (adj.), and the latter, with *sound* (n.). Apparently, differences like these really do occur. Recognising them in speech, however, does require more sophisticated modelling of stress than simply specifying lexical stress, and is thus outside the scope of this thesis. See section 2.1.1.2 on page 12 for an introduction, Ewen

[2] Asterisks * are used to mark ungrammatical examples in linguistics.

[3] As for the [n] in [ˈleːʋɑdən]: *sic.*

and van der Hulst (2001) to get a better idea of how stress assignment on higher level works in English, and Langeweg (1988) for a theory on stress assignment in Dutch.

Information value Greenberg *et al.* (2003) sees a correlation between canonical realisation and information value. For example, function words like *the* and *are* occur very often, and are almost never pronounced fully (Collins and Mees 1999), while less frequent words have at least one stressed syllable. Greenberg *et al.* say that "the ability to understand spoken language largely depends on the presence of relatively long, highly stressed syllables and words". This suggests that the words with a low probability in an information theory sense are the most prominent and the most canonically pronounced. This is exactly what van Son *et al.* (2004) find: "there is a consistent correlation between redundancy at the *word level*, i.e., word-frequency, and acoustic vowel reduction." Obvious as this may seem from a human speech perception perspective, it could provide useful information for speech recognisers in deciding when to use the whole lexicon and when to select the smaller subset of 'expected words'.

4.3 Considerations

It seems sensible to try and map the segmental information (e.g. phonemes, syllables) that humans use to produce speech onto computer-usable symbols as closely as possible and model the acoustic properties that coincide with them. The symbolic and acoustic representations are much intertwined. In making a speech recogniser spot additional dichotomies (in this case, stressed versus unstressed) these two aspects must be handled at the same time. It does not make sense to try and differentiate between stressed and unstressed /aː/ leaving out information about what the difference sounds like. On the other hand, without the symbolic information additional acoustics merely introduce noise in the mapping from acoustics to symbols. However, even if the symbolic representations are adapted, extracting extra acoustic information that cannot be explained through the symbols might degrade recognition.

The considerations given here combine the properties of lexical stress from chapter 2 on page 9 and objectives in section 4.2 on page 38.

4.3.1 Symbolic repesentation

In making a system that aims at recognising lexical stress the following should be considered:

Model reduction Unstressed phonemes are the first to be reduced; stressed phonemes are, in the words of Lea (1980, p. 169), "islands of phonetic reliability; detected phonetic structure is more nearly identical to underlying phonemic structure in stressed syllables". The system should allow for and anticipate a greater range of reduction for unstressed syllables than for stressed ones. According to the same study, on the other hand, sonorants in unstressed syllables tend to be more often categorised correctly than those in stressed syllables. How come? Unstressed syllables more often have only a sonorant in the nucleus (e.g. /ˈbʌˌtən/ becomes /ˈbʌtn̩/) so that the sonorant gets a fuller realisation than when it is in the coda.

Integration Recognition of lexical stress should begin at an early stage. How stress may interact with phone realisation was depicted in figure 4.3 on page 39. If an unstressed /ɑ/ is reduced to sound much like /ə/, the recogniser should indeed hypothesise that this is unstressed /ɑ/ rather than stressed /ˈɑ/. The models for stressed vowels should not be polluted with reduced variants of those same vowels. Furthermore, if "the stressed vowels of speech are one of the first groups of sounds to be recognized" (Shoup 1980), then it should be a great help to know where the stressed vowels are in the words that may be recognised.

Syllables All phonemes in one syllable have the same specification for stress. This fact must be used by the system to find a consistent hypothesis for the stress of the whole syllable. There seems to have been no use of stress information from consonants in earlier research; this is apparently a silly oversight that is totally unmotivated. Greenberg *et al.* (2003) manually assemble information from the Switchboard corpus about duration of segments, correlating them with (actual) stress. They concludes that vowel length correlates well with stress (this is consistent with Sluijter 1995; van Kuijk and Boves 1999). Although the duration difference for consonants in the onset "is not nearly half as great as observed among vocalic nuclei, the general patterns observed are broadly consistent." The coda shows a mixed picture: some fricatives [f, θ, z, ʃ, ʧ] and fortis stops [p, t, k] show a similar amount of disparity to segments in the onset. The liquids [r, l] pattern with the vowels. Other consonants show little or no difference.

Note that for fully explaining behaviour of consonants with respect to syllables, phonological and morphological modelling is necessary to account for *resyllabification.* For example, *at all* /ət ˈɔːl/ may become [əˈt$^{\mathrm{h}}$ɔːl] and *for eight* /fɔː eɪt/[4] may be realised [fɔːˈreɪt]. On the other hand, there may be a difference between *nitrate* [ˈnaɪtr̥eɪt] and *night rate* [ˈnaɪtreɪt].

Adaptability The recogniser must be able to adapt to speaking conditions. Supposing that lexical stress is realised differently depending on the speaking rate and 'sloppiness', the variability in actual stress may be too great to be handled by a non-adaptive system in practice. This is an issue for speech recognition in general. However, phonemes that are split into stressed and unstressed variants are supposed to model smaller areas. Letting these areas be confused by other factors will undo the advantage at least partially.

4.3.2 Acoustic representations

The acoustic features that may indicate stress for phonemes and must be extracted for detection of lexical stress (see 2.1.2 for a more detailed explanation of the features) are:

Intensity (see section 2.1.2.1 on page 15)

[4] In RP and other non-rhotic dialects.

Pitch (see section 2.1.2.2 on page 15)

Stress is typically thought to be connected to the pitch; this goes for some syllables in slower, more articulated speech in any case. The F_0 slope may be relevant more than its value, especially for consonants: a pitch peak on a stressed nucleus should mean that the pitch rises in the onset and decreases in the coda. However, from Ladd (1996) it is to be expected that the relation between stress and pitch peaks is not straightforward (see section 2.1.2.2 on page 15). Wang and Seneff (2001) find that pitch "yielded the poorest results" for lexical stress classification. Xie *et al.* (2004) are "surprised that the pitch features did not turn out to be particularly useful." Judging from earlier research they should not have expected the fundamental frequency to give much information for lexical stress directly.

Vowel quality (see section 2.1.2.4 on page 19)

The vowel quality is often captured in speech recognisers through MFCC (Mel-frequency cepstral coefficient) features that indirectly model formant frequency. Separating these data for stressed and unstressed vowels is not much of a departure from current-day garden variety speech recognition.

It could however yield quite a performance boost: though MFCCs basically capture formant frequencies, they do so by each measuring the energy in a specific frequency band. On the range of realisations of one phoneme, from full to wholly reduced ones, formants may shift from one coefficient to other ones. When models for stressed and unstressed realisations are separated, both become more specialised: they model a smaller range of formant values, making better use of the properties of MFCCs. See figure 4.3(b) on page 39 for a depiction.

Spectrum (see section 2.1.2.5 on page 20)

Sluijter (1995) uses the frequency bands 0–0.5 kHz, 0.5–1 kHz, 1–2 kHz and 2–4 kHz to capture the 'spectral tilt' she finds is an good correlate of lexical stress. Including this information in a speech recogniser directly is an option; however, MFCC features also contain spectral information. Though Sluijter's bands may be appropriate for expository purposes, a speech recogniser may find enough data in the MFCC features to go by.

If spectral features are added explicitly, it may be necessary to use the difference between the energy for frequency bands rather than the absolute values, so that it is indeed the spectral 'slope' that is measured.

Derivatives Wang and Seneff (2001) find that pitch slope is more relevant than maximum or average pitch over a segment. Indeed, intensity, F_0, vowel quality and spectral tilt are all expected to have a peak on the nucleus of a stressed syllable. Therefore, these features will show some kind of raising in the onset, and lowering in the coda. (Sluijter does not look into subsegmental behaviour of these features.) Derivatives over time could be used for capturing the temporal effects on these features. This should enable the recogniser to recognise those effects both within a phoneme and within the syllable: for example, consonants may be expected to have raising pitch in onsets of a stressed syllables and decreasing pitch in codas.

Also, using derivatives has the advantage of not having to rely on absolute values: sentences with different tunes or spoken with different vocal chords (think differing in sex) are provided for more easily. If the time between stressed syllables remains largely the same under different speaking rates for stress-timed languages, the feature slopes will be similarly normalised.

Duration (see section 2.1.2.3 on page 19)

The length of a phoneme is the primary indicator of its stress according to Sluijter (1995). To use duration during phoneme recognition, however, both the time-independent and temporal properties must be modelled simultaneously.

Another complication may be that duration not only depends on lexical stress. It appears than a context-dependent duration model is more perceptually adequate: not only the position within the word but also the neighbouring segments should be taken into account (Kompe 1997). An example where duration makes all the difference is in the pronunciation of the words *lap* and *lab*. There is only a slight, and often absent, difference in the pronunciation of the final consonant (see Harris 1994), but /æ/ before /b/ is much longer than before /p/ (Collins and Mees 1999, p. 52). In a Viterbi-based recogniser, for example, the difference between the two realisations of /æ/ cannot easily be modelled.

The duration depends on the overall speaking rate, which could be calculated by taking the ratio of the duration of the utterance and the expected duration from the phonemes.

$$\text{Speaking rate} = \frac{\sum \mu_{Dur}(V_i)}{\sum Dur(V_i)} \tag{1}$$

This equation is from Wang and Seneff (2001), but they use V_i, the ith vowel, whereas using the ith segment would be a better approach. (See section 5.8 on page 64 for an off-line implementation.) However, as shown in section 2.2.5 on page 28, it is not trivial to add duration modelling to a speech recogniser that uses HMMs and Viterbi.

4.4 Architecture

In considering speech recogniser architectures that will allow lexical stress recognition, the point of departure will be the current-day garden variety speech recogniser. One that uses HMMs and Viterbi search is depicted in figure 4.1 on page 36. The new elements in the systems that will be introduced will be coloured red, as in figure 4.2 on page 37. This figure shows an approach with verification; the new models proposed here, however, will aim to come up with new hypotheses for utterances.

The aim of this enumeration is not to exhaustively discuss the possibilities for lexical stress recognition; rather, broad sketches will give an idea of models' advantages and drawbacks in terms both of practicalities and the considerations

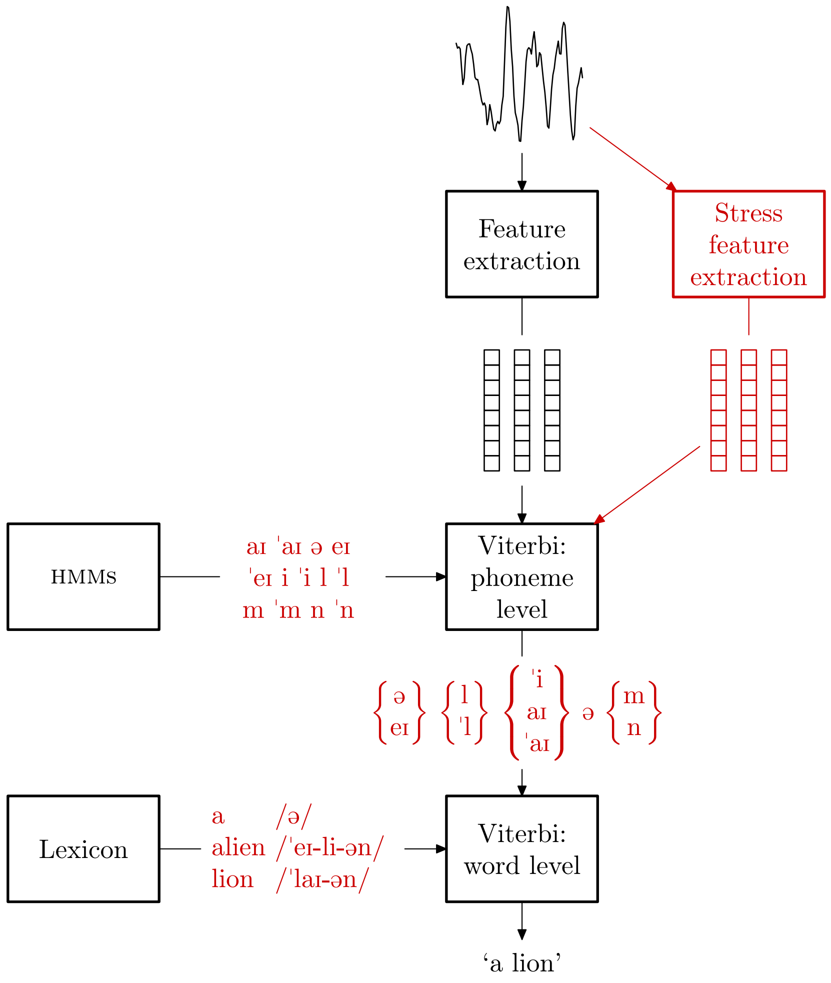


Figure 4.4 *A speech recognition system that uses extra features in the feature vectors.*

from section 4.3. Common to all these models is a split set of phonemes to accommodate the differences between stressed and unstressed ones. The lexicon is modified accordingly: a stress mark is added to all segments in stressed syllables. Thus, the lexicon contains only syllables with matching phonemes. This forces the recogniser to find consistent hypotheses over the segments in one syllable, though syllable stress is not modelled explicitly. For example, assume an English lexicon contains both /ˈs ˈʌ ˈb d ʒ ɛ k t/ and /s ʌ b ˈd ˈʒ ˈɛ ˈk ˈt/; the recogniser could never hypothesise /s ˈʌ b ˈd ʒ ˈɛ k ˈt/.

An architecture for a system that uses acoustic information for a given time are discussed first, then systems will be proposed that explicitly model duration. After that, these two aspects will be combined into one system.

4.4.1 Time-independent

Figure 4.4 on the facing page shows a method of modelling stress that uses only the acoustic information from one time slice, similar to the approaches found in van Kuijk *et al.* (1996); Wang and Seneff (2001). No new algorithms need to be invented compared to conventional speech recognition, but the feature vectors as well as the phoneme set are modified to model stress. These modifications are catered for in many available recognisers. The acoustic properties of stress are extracted from the audio data in the block labelled "Stress feature extraction". The symbolic properties are implemented as a doubling of the number of phones: phone models are separated into stressed and unstressed versions. Accordingly, the transcriptions in the lexicon specify which syllable is stressed. The main difference between earlier efforts and mine is that all phonemes, consonants as well as vowels, are marked for stress.

For capturing the acoustic properties, features that can be inserted into feature vectors are:

- the fundamental frequency (F_0);
- an overall energy measure, if it has not been included yet;
- spectral tilt features as proposed by Sluijter (1995).

The delta of these may be needed as well, effectively as numerical derivatives over time (see the item *Derivatives* on page 44). This should not be difficult to implement, since conventional speech recognisers already calculate derivatives of MFCC features. Derivatives may be useful more than in Sluijter's experiments because we use multiple-state HMMs, which can model different stages of a phoneme (see section 2.2.3 on page 26). For example, the spectral slope may increase more over the first part of a stressed vowel than the first part of an unstressed one.

Adding the new features to the conventional feature vectors may be done by concatenation, as shown in figure 4.5 on the next page. However, HMMs make the assumption that the features are statistically independent. We expect the MFCC features to contain at least some information about spectral tilt, so simple concatation may form a problem for the required statistical independence.

There are methods to reduce the dimensionality of data, removing unnecessary information from feature vectors. Principal Component Analysis (PCA) is often used to reduce the interdependence between feature values. Linear Discriminant Analysis (LDA), on the other hand, clusters data given examples from classes. Since we are looking for a way to use feature data to obtain an optimal separation between stressed and unstressed phonemes, LDA is the more promising method. The classes can be formed by grouping the feature vectors per phoneme from segmented data. The transformation found through LDA can then be used as a preprocessing step for all feature vectors before feeding them into the HMMs, as shown in figure 4.6 on the following page.

4.4.2 Temporal

Various authors have found a correlation between lexical stress and duration (Sluijter 1995; van Kuijk and Boves 1999; Greenberg *et al.* 2003, and see sec-

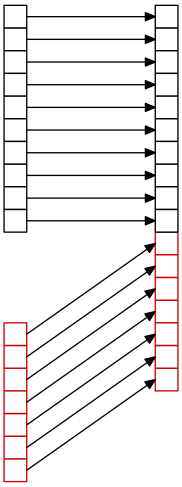


Figure 4.5 *Concatenating feature vectors from different sources. 'Standard' features, e.g.* MFCC *features, are displayed as black items. Those coloured red are added for recognition of lexical stress.*

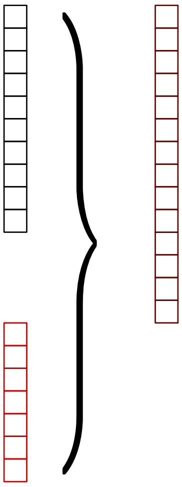


Figure 4.6 *Using* LDA *to combine feature vectors from different sources. 'Standard' features, e.g.* MFCC *features, are displayed as black items. Those coloured red are added for recognition of lexical stress.*

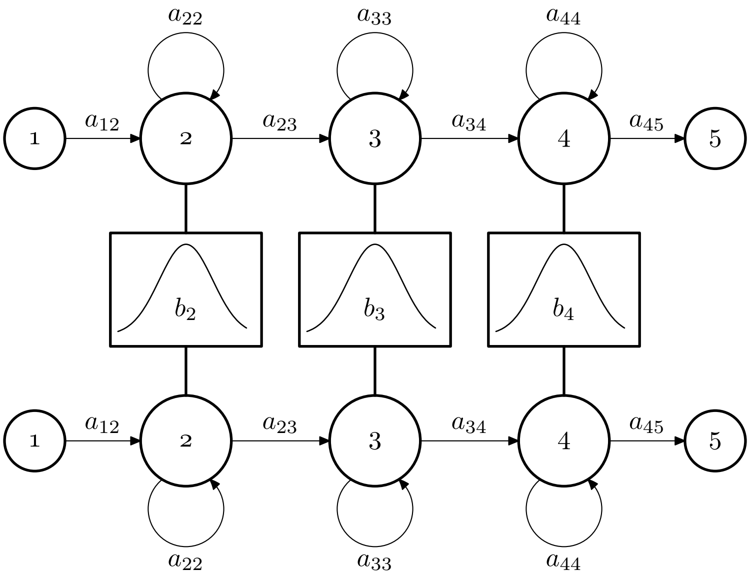


Figure 4.7 *Two HMMs sharing state distributions, but each having their own set of transition probabilities.*

tion 6.1.1 on page 70 for my measurements). However, there is a lack of durational modelling in today's speech recognisers (as proposed by Russell and Moore 1985; Ramesh and Wilpon 1992; Wang 1997, and see section 2.2.5 on page 28 for a discussion). Whereas spectral modelling as discussed in section 4.4.1 on page 47 takes much more processing for every time slice than for conventional speech recognisers, modelling phone duration would not have to take much time. However, there are three problems:

- It is not clear which algorithm must be used; in practice duration modelling either takes much processing or, for the simple variants, does not work (see section 2.2.5 on page 28).
- Duration modelling is all not there is to temporal modelling. As Sitaram and Sreenivas (1997) point out, subphone models should model duration as well.
- Duration is much dependent on speaking rate. A system must keep track of the speaking rate to be able to work out whether a given phone is longer or shorter than its expected rate.

4.4.2.1 State tying

Figure 4.7 shows a naive way to model duration. Stressed and unstressed phoneme models share their Gaussian distributions, but do have their own transition probabilities. This technique is called *state tying*. It has the advantage that training the distributions is as fast and as correct as conventional training (the amount of training data for the feature vectors is still the same), but the transition probabilities are handled differently according to stress. It seems that even

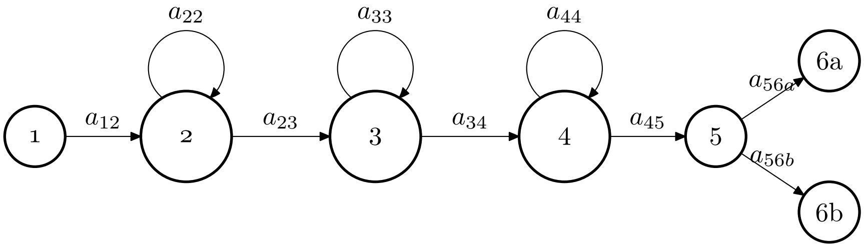


Figure 4.8 *A simple HMM whose end state splits into different phonemes. Here, a_{56_a} and a_{56_b} should depend on the time spent in the model.*

subphone duration would be modelled. However, as shown in section 2.2.5 on page 28, HMM transition probabilities do not work well for modelling duration.

4.4.2.2 *Deferred split*

Rather than introducing many HMMs to model the phonemes, it may be more efficient to defer this split to the final state of the HMM (see figure 4.8). The split should incorporate the duration calculation. That is to say, a_{56a} and a_{56b} should not be constants, but rather functions that depend on the length of the most likely path through the HMM. This strategy has the advantage of modelling phone duration explicitly (as a gamma distribution, for example). However, modelling subphone duration is not possible. Also, the required change to the algorithm, though not too large, is quite pervasive.

The architecture of a system that uses this strategy is depicted in figure 4.9 on the facing page. The figure shows that the difference between stressed and unstressed phonemes is detected through rescoring phoneme hypotheses for duration.

4.4.3 *Combinations*

Lexical stress is signalled by a number of different acoustic features. Considering only time-independent or only temporal features was a simplification mostly for explanatory purposes. How can the two features be combined into one recogniser?

4.4.3.1 *Rescore hypotheses*

Figure 4.10 on page 52 shows a not too pervasive method of integrating stress in a speech recogniser, conceptually similar to the verification approach. However, the Viterbi algorithm here outputs not the one most likely sequence, but the n most likely sequences according to the standard algorithm and phoneme set. When those n sequences, calculated without stress data and using a lexicon without stress marks, have been found, the alternatives are rescored on the basis of feature vector and duration information.

The benefit of using a system that rescores hypotheses in a second pass is that processing time is not unduly increased. Generating the hypotheses usually takes much more time than reevaluating them. Note that this rescoring method

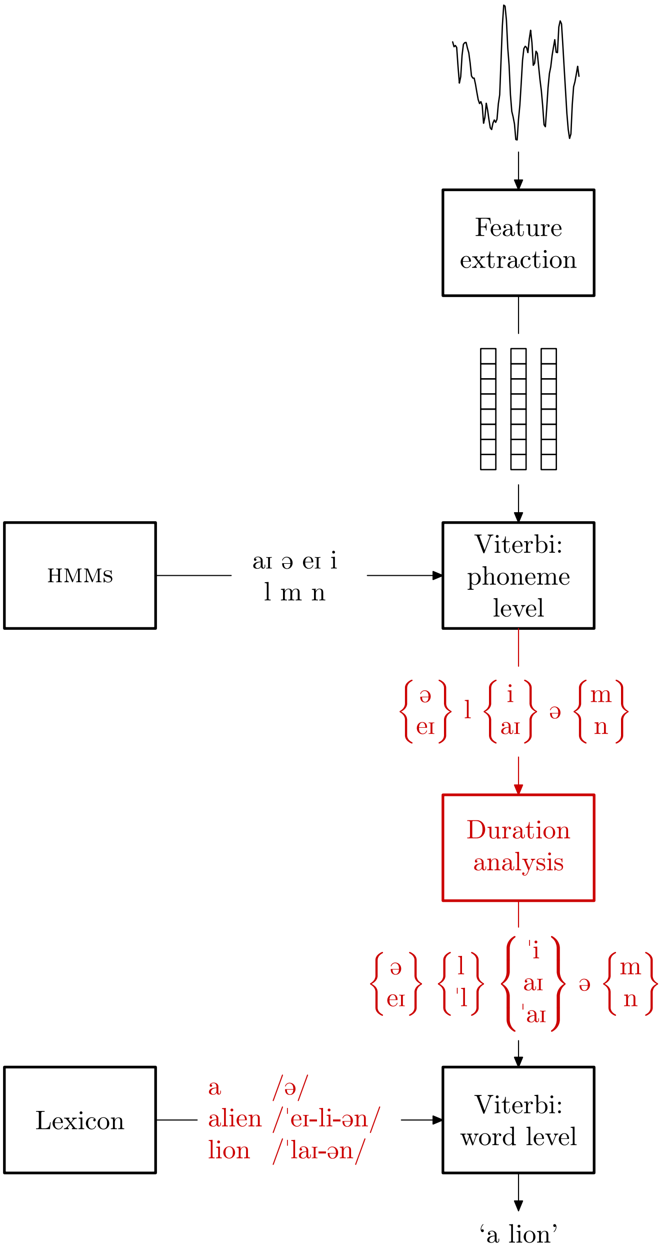


Figure 4.9 *A speech recognition system that uses duration to rescore phoneme hypotheses before they are used at the word level.*

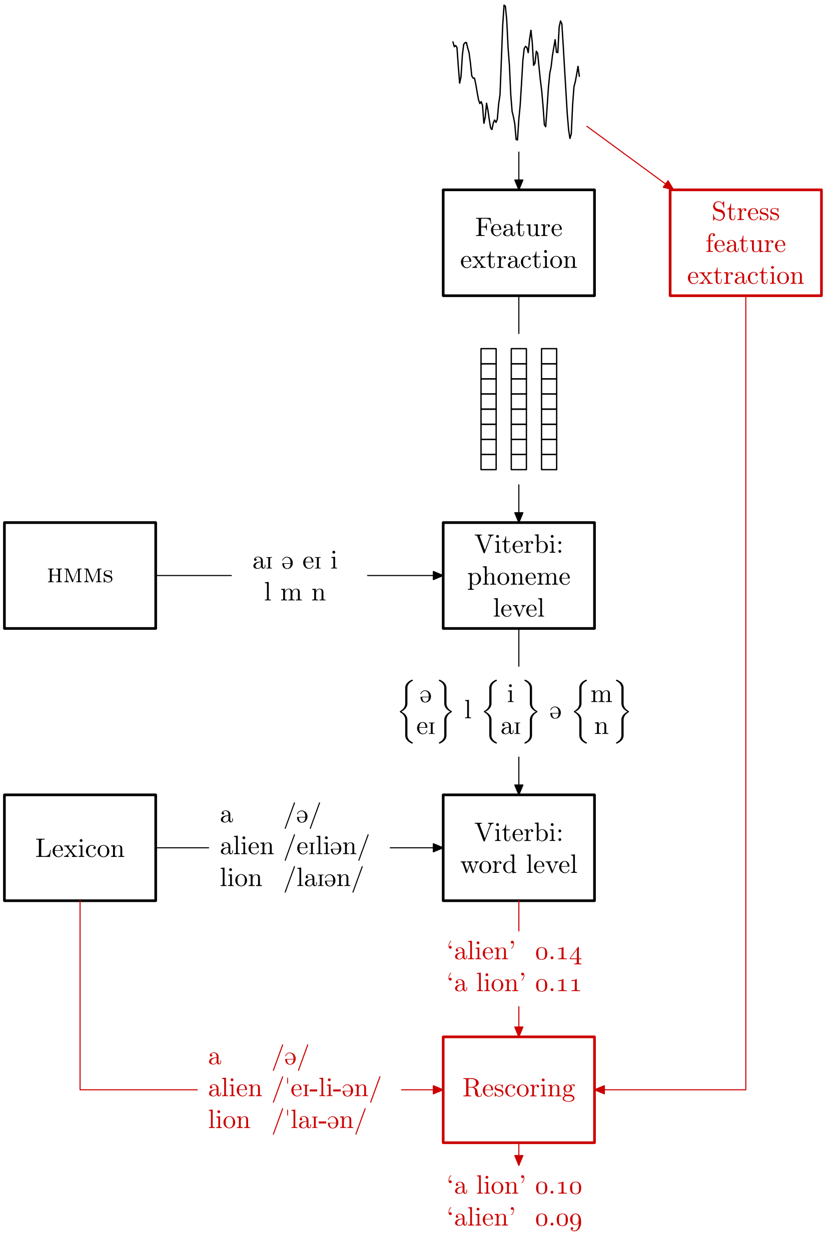


Figure 4.10 *A speech recognition system that uses stress to rescore alternative hypotheses in a final pass.*

would be valid for modelling only temporary properties as well. Another great advantage is that the speaking rate can be calculated by comparing the expected duration of the phonemes to the actual duration of the current utterance rather than previous utterances. This information can be used to properly account for duration differences (see section 5.8 on page 64 for an elaboration).

A disadvantage of modelling stress so late is that it cannot be integrated into processes taking place early, such as segmentation. Furthermore, phoneme models will not account for reduction of unstressed phonemes during actual recognition.

4.4.3.2 *Full stress modelling*

Figure 4.11 on the next page shows a system that uses a stressed – unstressed dichotomy throughout, like in figure 4.4 on page 46. It also uses duration throughout, similarly to figure 4.9 on page 51. The advantage of this system is that no stress information is lost: the integration (see 4.3.1 on page 43) is perfect. However, it does take more processing time because more features have to be considered, and it does have the disadvantages of durational modelling discussed in section 4.4.2 on page 47.

4.4.4 *More advanced*

As discussed in section 2.2.5 on page 28, various methods of including knowledge about phone and subphone duration have been proposed. However, they are not available in standard speech recognition toolkits. This may be due to their computing-intensive methods. However, if it were straightforward to implement more sophisticated temporal modelling than HMMs provide into a stress-enabled speech recogniser, this would help a great deal. The system would look like the one in figure 4.11 on the next page except that temporal modelling and subphone recognition would be intertwined.

Modelling of elision, reduction to $\emptyset$, would be possible with HMMs: a connection from the start state to the end state would do the job. However, it is not straightforward to find for which phonemes this optimisation would go, while incorrectly adding loops may cause more harm than good. To properly account for elision, context-sensitive models should probably be used.

Until now the symbolic part of lexical stress has been modelled by adding stress marks to the lexicon. However, using multiple levels of stress might be more sensible. Greenberg *et al.* (2003) distinguish five levels of actual, phonetic, stress. A corpus with actual stress annotated per phone would make it possible to separate concerns when using stress in a speech recogniser. Now lexical stress is directly determined from the acoustics. A more sophisticated recogniser would recognise the level of phonetic stress more precisely from the acoustics. The level of phonetic stress would be used to find the phonological stress. For two levels of lexical stress, a simple approximation for phonological stress, this would mean a probabilistic mapping. This may make the mapping from acoustics to phonological stress more accurate; it also makes it easier to introduce a more sophisticated system for phonological stress, modelling phrase stress and the like.

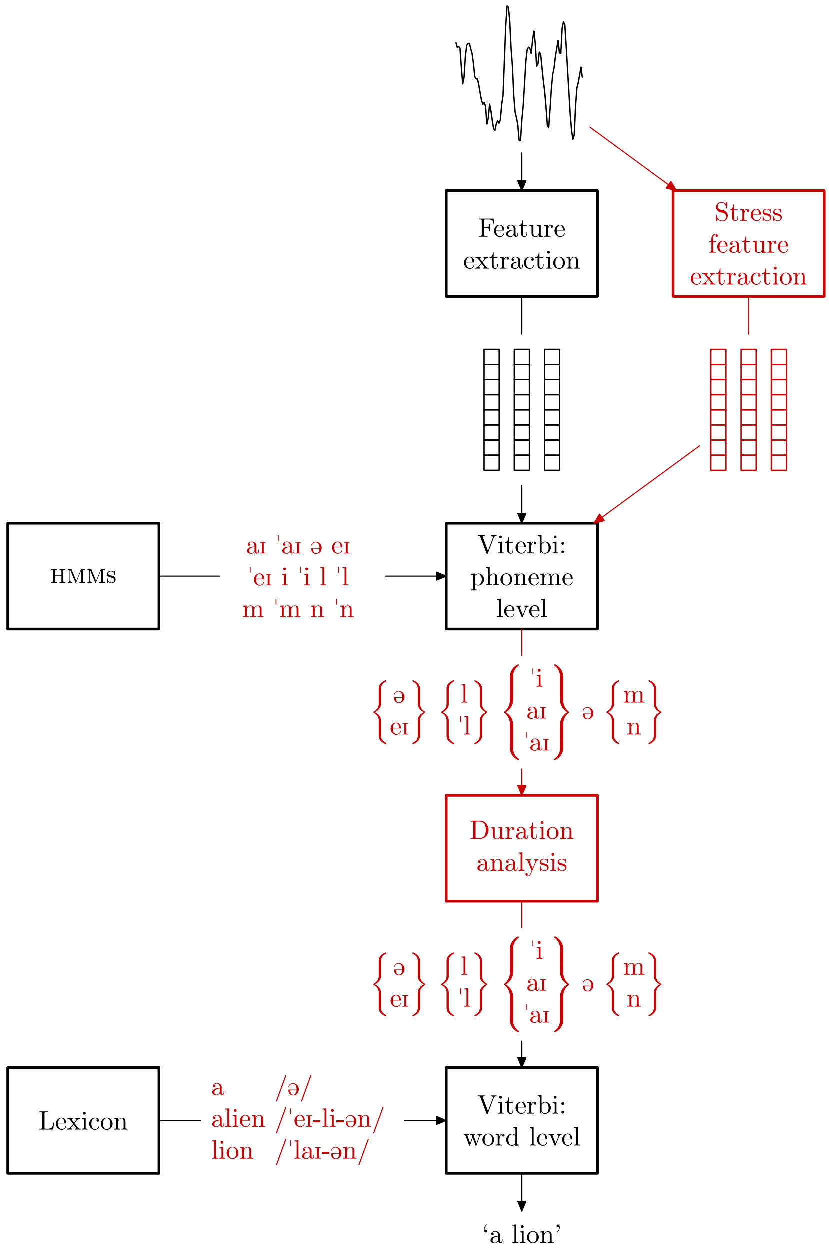


Figure 4.11 *A speech recognition system that uses both extra features and phoneme duration; a combination of the systems in figures 4.4 on page 46 and 4.9 on page 51.*

4.4.5 *System proposal*

Figure 4.4 on page 46 shows the architecture of a system that extracts stress-related features for each time slice and adds them to the feature vectors. This system shows the following properties, when related to the considerations from section 4.3 on page 42:

- Reduction is modelled through MFCCs.
- The stress feature information is already available before the phonemes are recognised, so that it is used at an early stage. This, it can improve reduction and segmentation.
- Consonants are taken into account in the way described in section 4.4 on page 45. The features used for vowels will be used for consonants as well.
- Duration is not modelled, since this would require more advanced HMMs (to measure duration) and an adaptive recogniser, both of which are not available in the HTK; nor are they straightforwardly implemented from scratch.

This architecture will be used for the speech recognition system that implements lexical stress recognition. The implementation of the system is described in the next chapter.

Chapter 5

System

ignotas animum dimittit in artes
naturamque novat. nam ponit in ordine pennas,
a minima coeptas, longam breviore sequenti[1]

Ovid, *Metamorphoses* VIII

To test many of the ideas and claims made thus far, and to give them more practical weight, they are implemented in a speech recognition system. This will be used both to gather statistics from corpora and to see the effect on recogniser performance. A system that does not use stress has been set up as well as one that does. This system is based on earlier work. This chapter will not give a full overview of how the Hidden Markov Toolkit's modules work or how to concatenate feature vectors from various sources. The tools I use for such things have been written earlier (see Wiggers 2001).

What this chapter *will* give is an overview of relevant practical decisions I make, most of them based on findings in the literature and, where the literature did not give plain answers, on common sense.

The implementation structure is given in figure 5.1 on the next page. *HCopy* is an HTK tool that converts the sound recording into feature vectors with MFCCs. *Praat* and *sox* edit sound files. They are used to gather spectral and fundamental frequency features from the sound file. *cvf* (Wiggers 2001) concatenates feature vectors from different sources. *HCompV* initialises the HMMs from scratch; *HERest* retrains them. Finally, *HVite* runs the speech recogniser.

5.1 Base system

The speech recognition system by Wiggers (2001), implemented using HTK (see section 3.1 on page 31), has provided me with the possibility to plug in stress recognition-specific features right away (for the system that uses DUTAVSC). As

[1] 'to uncoth Arts he bent the force of all his wits
To alter natures course by craft. And orderly he knits
A rowe of fethers one by one, beginning with the short,
And overmatching still eche quill with one of longer sort'
(translation Arthur Golding, from <http://www.perseus.tufts.edu/>.)

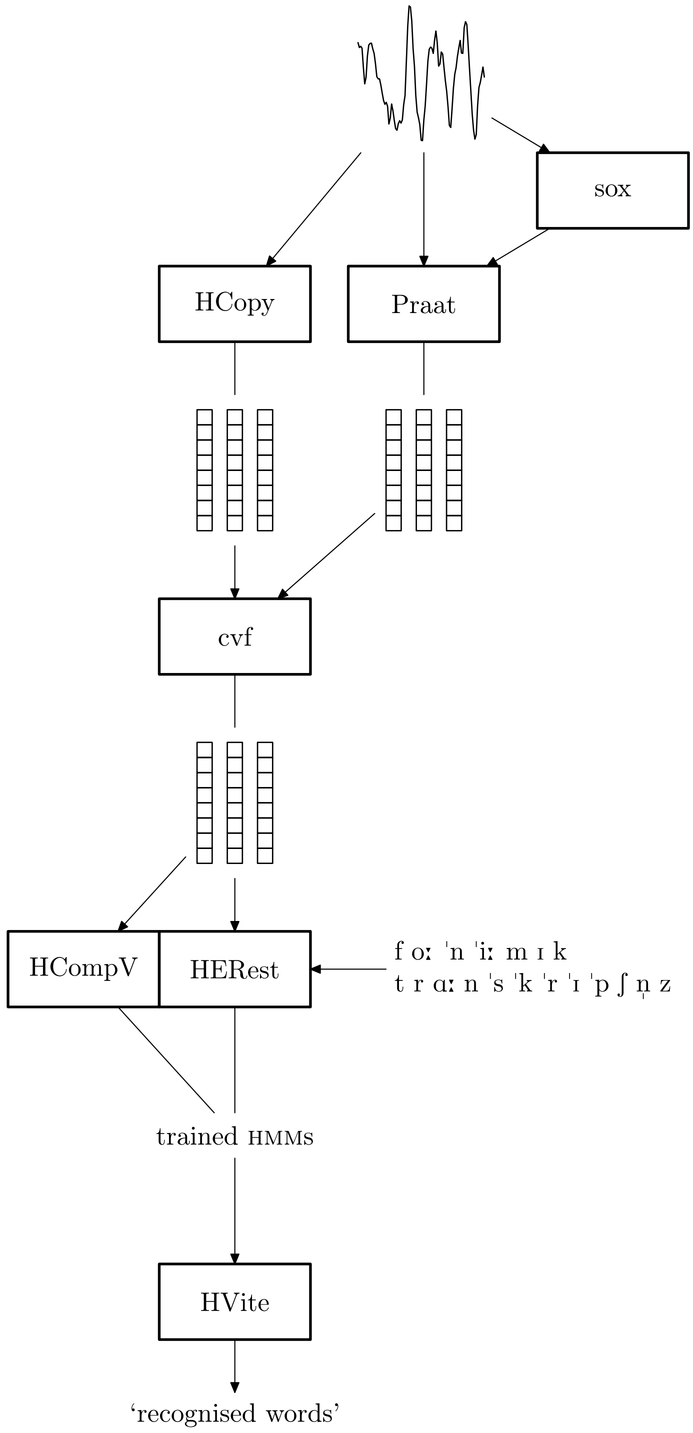


Figure 5.1 *The system architecture.*

his objective was to complement speech recognition with lipreading, even the architecture and the tools to concatenate feature vectors are readily available. A system trained on the POLYPHONE corpus, ready to be retrained on DUTAVSC (see section 5.9 on page 64) is there.

Though many of the shell scripts connecting the HTK tools were available for the DUTAVSC system (see section 5.9 on page 64), all scripts were rewritten for the recogniser based on the CGN (see section 5.10 on page 65), adapting to another operating system, enhancing their customisability and enabling parallel processing. The system that uses CGN (see section 5.10 on page 65) was bootstrapped from a flat start.

Producing the extra feature vectors (so that there actually is something to be concatenated) is done with Praat (see section 3.2 on page 32) and, where that program fails due to preposterous memory requirements, the Linux program *sox* (*So*und E*x*change). The extra features have their derivatives added. The feature values themselves are normalised over one utterance to reduce the influence of factors like microphones and sex on recognition.

The information from the corpora must be converted to HTK-compatible formats. Since the transcriptions my system require stress marks, which are not available in garden-variety systems, the transcriptions and lexica for both the DUTAVSC and CGN systems have to be converted.

5.2 Technical requirements

The memory some programs require is rather too much for current-day personal computers (exceeding 3 gigabytes of internal memory). Praat fails when extracting the energy in spectral bands from sound files over 25 megabytes. To overcome this problem, I add the Linux program *sox* to the toolchain.

The HTK tool *HERest* fails on feature vector files larger than 3 megabytes. This is overcome by cutting up the recordings into smaller pieces; the program that does so is called *cut_ up*. It separates the speech data into smaller phrases, as indicated in the CGN transcriptions by punctuation.

Training one iteration, when six high-end personal computers were used, takes about one hour for the first iterations (where one Gaussian mixture per distribution was used), but eight hours for the final iterations (with sixteen mixtures). Evaluation the recogniser (i.e. trying it out on the evaluation set) takes about four hours at first, and ten hours in the end.

5.3 Levels of stress

Neither in simple nor in more elaborate theories is stress ever seen as an on-or-off thing. Section 2.1.1.2 on page 12 has shown how phonological theories handle levels of phrasal stress. In many dictionaries (e.g. Procter 1995) syllables can have primary stress, secondary stress, or no stress at all. Collins and Mees (1999, p. 229) show that one may hear 5 degrees of stress in the word *eccentricity*. In (1), numbers 1–5 are used: 1 indicates strong stress, 5 weak stress.

(1) ɛk sən trɪ sə tɪ
2 4 1 5 3

However, a practical implementation has to work with limited resources. Time and comprehensive theories on phrasal stress are lacking, and the resource called ‘lexicon’ (see section 3.3.2 on page 32) is limited too — it only knows one type of stress. A consolation may be that Wang and Seneff (2001) do not find any improvement from adding multiple levels of stress to a speech recognition system; also, having only two levels does make the graphs easier to read.

5.4 Fundamental frequency

Section 2.1.2.2 on page 15 has shown that though fundamental frequency may be perceived as an important feature of stress, implementation of this notion is far from straightforward. Notwithstanding, it will be included in the feature vector. There is a problem: not every segment has a fundamental frequency, because not every segment is voiced. This may be a phonological thing (e.g. /p/ is not normally voiced), but it may be a phonetic thing as well (e.g. *lose* /luːz/ realised with final [z̥]). I use the Praat program (see section 3.2 on page 32) to extract the fundamental frequency from the audio recordings. It emits undefined as a value when it cannot find the fundamental frequency. There are two possible courses for mapping undefined unto a feature vector value.

- Setting the feature to zero or some other out-of-range number. This will make a clear separation between slices that have a fundamental frequency and those without. A drawback is that the ability to generalise over these values is crippled, for example when voiced phonemes are devoiced phonetically, or when voiceless phonemes are voiced (e.g. *vijfentwintig* ‘twenty-five’ /**v**ɛi**f**əntʊɪntəx/ [**f**ɛi**v**əntʊɪntəx]).

  Using an out-of-range value would mean adding an extra feature of voicedness too. By this we would inadvertently introduce new data.

  Another drawback is that using an out-of-range number makes it more difficult to normalise the F_0 values. Using derivatives to recognise pitch raising and lowering would become impossible.

- Linearly interpolating the undefined values between neighbouring known values. This approach borders on thinking up values that are not actually there; it will however allow better generalisation. However, its main merit is that it formalises the notion about intonational tune in works like Bolinger (1986; 1989); Ladd (1996). These works show intonation contours as uninterrupted curves, which is supposedly the representation people use internally. The fact that there is no voicing, i.e. no carrier for the intonation contour, in voiceless stops is a mere detail.

  The practical counterpart of this argument is robustness: if by accident Praat found no fundamental frequency where there is one, this would be mitigated because the interpolation would fill in a sensible contour. For example, figure 2.3(a) on page 18 shows such a gap, and clearly linear interpolation would fix up this contour well.

  Furthermore, interpolation may have the practical advantage of enabling the recogniser to detect stress on voiceless consonants in the onset or coda of a syllable that has a pitch peak: the pitch could rise in the onset, and lower in the coda.

Interpolation seems a sensible choice. The fundamental frequency will extracted from the audio data using the custom Praat (Boersma 2001) script `f0.praat` that analyses the pitch and writes it to a file. I wrote a program to interpolate linearly between these values and called it *interpolate*.

5.5 Spectral tilt

According to Sluijter (1995), stressed vowels receive more energy, which is found most prominently in the higher spectral bands. In a graphical representation the spectrum normally has a downward slope, which is smaller for stressed syllables; hence the property is called "spectral tilt". Operationally, this is defined as the relative energy in a number of spectral bands. She uses four spectral bands: 0 kHz – .5 kHz, .5 kHz – 1 kHz, 1 kHz – 2 kHz, and 2 kHz – 4 kHz. The energy of these bands is inserted into the feature vectors by integrating over the energy of each band. This is done with a *Praat* script `spectrum.praat` that extracts spectral information from every 10 ms slice. The difference between the energy in different bands is also included. Together with the deltas and the delta-deltas, this makes an extra 30 features.[2]

5.6 Consonants

In phonological literature a difference is made between realisations in the coda and in the onset (see section 2.1.1.1 on page 10). Phonetic literature (e.g. Greenberg *et al.* 2003) has found more fine-grained differences. A speech recogniser could use this difference by modelling stress in onsets and codas differently.

Another effect that may be seen is different slopes. Stressed syllables often have a pitch peak associated with them; this could yield raising pitch in the onsets, and lowering pitch in the codas of stressed syllables. Spectral features are measures for the effort with which phonemes are pronounced. The speaking effort is a continuous measure: it probably increases over the beginning of a stressed syllable and decreases over the end. We therefore expect that derivatives for spectral features also may be correlated with lexical stress, especially for consonants.

This is where the remarks in section 4.3 on page 42 come in: to properly model the acoustic effects, the symbols must reflect them. In this case, a difference must be made between consonants in the onset and those in the coda. This gives us four models for each consonant. How the specifications for onset, nucleus, and coda for a given syllable are matched through the lexicon is described in section 4.4 on page 46.

[2] I experimented with classifiers to identify the most important features, but they did not seem to favour a certain set of features.

5.7 Function words

It appears that the lexica used, the POLYPHONE and CGN lexica, both have all one-syllable function words specified as stressed. Not only *van* /ˈvɑn/ 'from' and *dus* /ˈdʉs/ 'so, thus' but also reduced forms like *de* /ˈdə/ 'the (m./f.)' and *het* /ˈət/ 'the (n.)' are stressed. This is rather strange, especially for the articles, where the phoneme /ə/, which can only occur in unstressed syllables, is marked as stressed.

To alleviate this problem and to potentially increase recogniser performance, a number of function words are not marked as stressed at all (as was done by van Kuijk *et al.* 1996; van Kuijk and Boves 1999; Wang and Seneff 2001; van den Heuvel *et al.* 2003). It may seem bad practice to remove lexical stress information from function words, but it is backed up by linguistic theory. Syntactic literature (Cook and Newson 1996, p. 187; Poole 2002, Ch. 3) often makes a distinction between *functional* phrases (or categories, phrase types) and *lexical* phrases. According to Cook and Newson, functional phrases form a "closed class"; they have "no descriptive content" but "grammatical features", and are "usually unstressed". Booij (1999) describes function words as clitics that are phonologically dependent on a host word to form a prosodic words, because they are not stressed themselves.[5]

(2) *Zij* *kochten 't* *boek*
$(\text{zɛi})_\omega$ $((\text{kɔx})_\sigma(\text{tə})_\sigma(\text{nət})_\sigma)_\omega$ $(\text{buːk})_\omega$
'They bought the book'

(2) demonstrates how the clitic *'t* 'the (n.)' is embedded in the prosodic word headed by *kochten*. It also demonstrates how the phonological division is different from the syntactical division: *'t boek* is a syntactical constituent, while *kochten 't* is a phonological one. Thus, function words are marked unstressed as a model of effects of the phonological rules that does not require modelling the rules themselves.

Collins and Mees (1999) give a list of Dutch function words that are subject to reduction and seem to fit the bill (see table 5.1 on the facing page). These words are marked as unstressed in the recogniser lexicon.

Some implementation notes:

- The Dutch orthographical system can show reduction for a number of function words, as shown in table 5.2 on page 64. These words have not been selected rationally: centuries of spelling conventions have produced them. However, the CGN corpus does use the reduced forms in the transcriptions, so these can be unstressed as well. (Most of them are already.) Note that this does not mean that canonical versions of function words should be stressed.

- The word *de* 'the (m./f.)' has a reduced orthography already; the CGN lexicon (unlike CELEX) specifies it as unstressed.

- The CGN lexicon differentiates between homographs. The words *haar* 'her' and *was* 'was/were' will be marked unstressed, while *haar* 'hair' and *was* '(I) wash' will not.

[5] σ indicates a syllable; ω a (prosodic) word.

Table 5.1 *Function words that are specified as unstressed, as found as "weak forms" in Collins and Mees (1999, pp. 239–40). Though the* CGN *transcription is documented to be taken directly from* CELEX*, they differ on some words (e.g.* het *'the (n.)' is not stressed in the* CGN *lexicon).*

	Orthography	**Canonical form**	**Weak form**	**CGN transcription**
Determiners	een	/eːn/	/ən/	/ˈeːn/
	het	/hɛt/	/ət, t/	/ət/
Prepositions	aan	/aːn/	/ən, n̩/	/ˈaːn/
	met	/mɛt/	/mət/	/ˈmɛt/
	naar	/naːr/	/nər, nr̩, nə/	/ˈnaːr/
	ten	/tɛn/	/tən/	/ˈtɛn/
	ter	/tɛr/	/tər, tr̩, tə/	/ˈtɛr/
	van	/vɑn/[3]	/vən, fən, fn̩/[3]	/ˈvɑn/
	voor	/voːr/[3]	/vər, vr̩, və, fər, fə, fr̩/[3]	/ˈvoːr/
Personal and possessive pronouns	ik	/ɪk/	/k/	/ˈɪk/
	jij	/jɛi/	/jə/	/ˈjɛi/
	hij	/hɛi/	/i/	/ˈhɛi/
	zij	/zɛi/	/zə/	/ˈzɛi/
	het	/hɛt/	/ət/	/ˈət/
	wij	/ʋɛi/	/wə/	/ˈʋɛi/
	haar	/haːr/	/dər, dr̩, də, ər, ə/	/ˈhaːr/
	hem	/hɛm/	/əm, m̩/	/ˈhɛm/
	mijn	/mɛin/	/mən, mn̩/	/mɛin/
	jouw, jou	/jɑu/	/jə/	/jɑuw, jɑu/[4]
	zijn	/zɛin/	/zən, zn̩/	/ˈzɛin/
Auxiliary verbs	ben	/bɛn/	/bən/	/ˈbɛn/
	is	/ɪs/	/əs, s/	/ˈɪs/
	was	/ʋɑs/	/ʋəs/	/ˈʋɑs/
	heb	/hɛp/	/əp, həp/	/ˈhɛp/
	heeft	/heːft/	/əft/	/ˈheːft/
	had	/hɑt/	/ət/	/ˈhɑt/
	kan	/kɑn/	/kən, kn̩/	/ˈkɑn/
	zal	/zɑl/	/zəl, zl̩/	/ˈzɑl/
Miscellaneous	als	/ɑls/	/əls, əs/	/ˈɑls/
	daar	/daːr/	/dər, dr̩, də/	/ˈdaːr/
	dan	/dɑn/	/dən/	/ˈdɑn/
	dat	/dɑt/	/dət/	/ˈdɑt/
	eens	/eːns/	/əns, əs, s/	/ˈeːns/
	en	/ɛn/	/ən, n̩/	/ˈɛn/
	er	/ɛr/	/ər, r, ə, dər, dr̩/	/ˈɛr/
	maar	/maːr/	/mər, mr̩, mə/	/ˈmaːr/
	of	/ɔf/	/əf, f/	/ˈɔf/
	waar	/ʋaːr/	/ʋər, ʋə, ʋr̩/	/ˈʋaːr/
	wat	/ʋɑt/	/ʋət/	/ˈʋɑt/
	wel	/ʋɛl/	/ʋəl, ʋl̩/	/ˈʋɛl/

[3] Collins and Mees (1999) write /f̬/ rather than /v/

[4] /jɑu**w**/: *sic.*

Table 5.2 *Additional forms for reduced orthography versions of function words in the* CGN *lexicon.*

	Canonical		**Reduced**	
	Orthography	**Transcription**	**Orthography**	**Transcription**
Determiners	het	/ˈət/	’t	/ət/
Personal and possessive pronouns	jij	/ˈjɛi/	je	/jə/
	hij	/ˈhɛi/	ie	/ˈiː/
	zij	/ˈzɛi/	ze	/zə/
	wij	/ˈʋɛi/	we	/ʋə/
	haar	/ˈhaːr/	d’r	/dər/
	hem	/ˈhɛm/	’m	/əm/
	jouw, jou	/ˈjɑu/	je	/jə/
	zijn	/ˈzɛin/	z’n	/zən/
Miscellaneous	daar	/ˈdaːr/	d’r	/dər/
	er	/ˈɛr/	d’r	/dər/
	eens	/ˈeːns/	’s	/əs/

5.8 Duration

Due to such factors as the overall speaking rate, phone duration is not easily measured objectively. As an approximation of the relative length of a phone by a certain speaker in a specific utterance, it is normalised. Define p_j as a phone from the corpus. p_j is the realisation of the phoneme $i = r(p_j)$; n_i is the number of realisations of the phoneme i. $d(p_j)$ is the actual duration of p_j. U_k is one utterance. The expected duration for a phoneme i is defined as

$$\mu_i = \frac{\sum_{p_j : r(p_j) = i} d(p_j)}{n_i}. \tag{3}$$

The normalised duration of p_j is

$$d'(p_j) = d(p_j) \cdot \alpha_k, \quad p_j \in U_k \tag{4}$$

where α_k is the speaking rate of utterance U_k, which is defined as

$$\alpha_k = \frac{\sum_{p_j \in U_k} \mu_{r(p_j)}}{\sum_{p_j \in U_k} d(p_j)} \tag{5}$$

The *duration_ normalise* program calculates $d'(p_j)$ for all phones in the corpus. It excludes non-speech phonemes (sil and sp) from the calculation.

5.9 DUTAVSC system

To get a feel for the kind of data and improvements that are to be expected, a system is built on a small corpus, the DUTAVSC (see section 3.3.1 on page 32). A speech recogniser bootstrapped on the POLYPHONE corpus is readily available

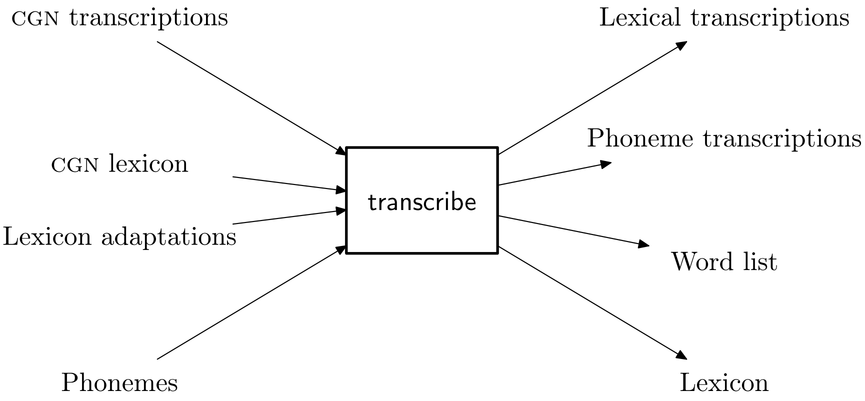


Figure 5.2 *The information flow around the* transcribe *program.*

(see section 5.1 on page 57). Its phone models are copied to stressed and unstressed versions (no difference is made between the onset and the coda yet) and get four training iterations. This system is used to gather statistics (see section 6.1 on page 69).

5.10 CGN system

Two systems are trained on the CGN corpus (see section 3.3.2 on page 32): one baseline system without stress marks and one that does make a difference between stressed and unstressed phonemes. The systems are the same in other respects, so that their performance can be sensibly compared.

772 recordings are selected for their degree of preparation, which are divided into 54 842 files containing a phrase each (see section 5.2 on page 59). These comprise almost 53 hours of recording with 775 034 words. The recognition system cannot adapt to varying speaking styles and speaking rates (see sections 4.3.1 on page 42 and 2.2.5 on page 28). To get the kind of speech in the selected recordings as consistent as possible, only those the corpus marks as "scripted" are used. They are split into groups: 80 % are used as training data, 10 % for testing, and another 10 % for evaluation. No group has data from speakers that also occur in other groups.

5.10.1 Data preparation

To convert the transcriptions that come with the CGN to a form with stress information and readable by the HTK components, I write a program called *transcribe*. Its main purpose is to convert transcriptions from CGN's XML files into a HTK-readable MLF file. The problems it deals with are phonemes of which not enough examples are found, unknown words, unavailable or invalid transcriptions, and recordings that are too long to be processed by HTK (see section 5.2 on page 59 for the last problem). Figure 5.2 shows the inputs and outputs of the program. The information going into it consists of information that comes with CGN and information that is added.

The CGN transcriptions contain the words in all recorded utterances. The words are described through their orthography and a word ID.

The CGN lexicon contains different pronunciation transcriptions for all words. *transcribe* maps the lexical transcriptions onto transcriptions consisting of separate phonemes with stress marks. The lexical transcriptions taken from the CELEX lexicon have stress marks, but are not available for all words.

The adaptations to the lexicon are necessary for marking function words unstressed (see section 5.7 on page 62) and for adding hesitation fillers *eh* and *ehm*.

The phonemes contain the mapping of phonemes in the CELEX transcription to phoneme symbols suitable for processing with HTK. This is where the difference between the conventional and stress-enabled recognisers is made. For the former, all phonemes are mapped to symbols with stress marks; for the latter, the destination symbols are different in terms of stress marking.

The entries in the transcriptions are first matched with the ones in the lexicon by word ID. This enables transcribing the difference between for example *haar* ‘her’ /haːr/ and *haar* ‘hair’ /ˈhaːr/. The orthography is used only for words without an ID or CELEX transcription. sil phonemes are appended and prepended to the transcriptions to account for silence at the start and end of recordings. Non-speech sounds are transcribed as ggg in the CGN transcription; these are retained. CGN transcribes unknown words with xxx. Some words do not pass through the system and are replaced by xxx.

- The CELEX transcription, which is the only pronunciation transcription in which syllables are marked for stress, is not available for 4.5 % of the transcribed word instances of the training set.[6] These words are transcribed as xxx. (To retain consistency this also goes for the non-stress-marked system, even though more comprehensive transcriptions are available.)

- The lexicon contains words with transcriptions with syllabic consonants (*’s* ‘once’ /s/, *sst* ‘sh’ /ṣ̩t/ and *pst* ‘psh’ /pṣ̩t/) and words with invalid transcriptions for Dutch (*pass* ‘pass (soccer)’ /pɑːs/). For the reduced form *’s* for *eens* ‘once’ (but used much more often than English ‘once’) the pronunciation should be /əs/, notwithstanding the orthography. Assigning a stress valuation to /s/ in context is not trivial, so /əs/ (unstressed) is used. The other words (*sst*, *pst* and *pass*) are *hapax legomena*: there are too few instances of the phonemes to be trained; furthermore, their number is fairly low compared to the words already transcribed as xxx. Adding those few words will not significanty add to the number of unknown words.

- The CGN lexicon includes the marginal nasalised phonemes /ɔ̃/ and /ʉ̃/, of which the training data has no examples, and /ɛ̃/, of which the training data contains only two examples (*point* (French) /ˈpwɛ̃/). These phonemes are deleted from the list of phonemes; the *transcribe* program, seeing that /ɛ̃/ does not exist, shrewdly replaces *point* by xxx as well.

[6] To be precise, 29 904 words out of 665 353.

Table 5.3 *Outline of the structure of the speech recognisers through the training iterations.*

Iteration	Structure
1...4	
5...19	sp phoneme introduced; sil and xxx phonemes altered
20...24	2 mixtures per phoneme
25...29	4 mixtures per phoneme
30...34	6 mixtures per phoneme
35...39	8 mixtures per phoneme
40...44	10 mixtures per phoneme
45...49	12 mixtures per phoneme
50...60	16 mixtures per phoneme

The outputs of the *transcribe* program are the following.

Transcriptions that come in two flavours: lexical transcriptions (i.e. a list of word IDs per recording) and phoneme transcriptions (a list of phonemes per recording). The former is basically the same as the CGN transcriptions; why not simply convert this one? This is not done because the phoneme transcriptions need to be consistent with the lexical ones, especially since there are a large number of changes (due to word ID processing and replacing unknown words by xxx).

Lexical transcriptions are used to compare the recogniser output to when testing. Phoneme transcriptions are used for training.

The lexicon contain the words (as word IDs) with the new phonemic transcriptions.

The word list contains all words, which, though not fundamentally different from the lexicon, some HTK tools need.

Though this is not shown in figure 5.2 on page 65, the lexicon and the transcriptions come in different flavours as well, depending on whether an sp phoneme is needed in between words (see the next section).

5.10.2 Training

The systems are initialised from a flat start (using the HTK tool *HCompV*). To account for silence at the beginning and end of recordings, a *silence* (sil) phoneme is introduced. I set both this phoneme and the xxx phoneme to a straightforward three-state HMM to prevent them from eating up all the recording through their general nature. After four re-estimations with the HTK tool *HERest* I introduce a *short pause* (sp) phoneme between words to account for the optional pause between words. This phoneme contains one optional state, which it shares with the sil phoneme. The sil and xxx phones get an extra loop from the fourth to the second state to take care of their variable length.

From the 20th training iteration Gaussian mixtures are introduced in the phone likelihood estimation for more precise modelling of the feature vectors' properties. Table 5.3 shows the number of mixtures used for which iterations.

The number of mixtures is increased in steps of two. (The HTK tools cannot determine the number of mixtures per phone model automatically. However, when increasing the number globally the mixtures with the highest variance are split up.) The minimum number of examples for a phone model to be retrained is increased when more mixtures are included, to prevent the models from losing their general quality.

Chapter 6

Experiments

> *[T]he saddler's next end is to make a good saddle, but his further end to serve a nobler faculty, which is horsemanship.*
>
> Sir Philip Sidney, *The Defence of Poesy*

Sir Philip Sidney would have agreed: good ideas are not enough; they must be put into practice. If the theories on lexical stress and their implementation have enough bearing on speech recognition, this should be reflected in the statistics from the recogniser. This chapter contains four different types of results.

- The statistics from collected data on vowels, which are obtained from forced alignment, will be used to corroborate the results from Sluijter (1995) and van Kuijk and Boves (1999).
- The statistics from similarly collected data for consonants will be used to find how much stressed and unstressed consonants' acoustics differ. This may also shed light on whether Greenberg *et al.*'s (2003) results (on duration differences between stressed segments in English) are valid for lexical stress in Dutch as well.
- Statistics on the separability by the speech recogniser of stressed and unstressed phonemes will tell whether the phoneme models' Gaussians are able to model stress from the acoustic data.
- The statistics comparing the baseline speech recogniser performance with the stress-enabled recogniser's will tell whether the recogniser is able to use lexical stress.

6.1 DUTAVSC

A system trained on the DUTAVSC corpus, which was described in section 5.9 on page 64, is used to gather statistics from a relatively small collection of data. The results are assembled by collecting feature vectors from phones that are segmented with forced alignment. (Forced alignment is a technique where the speech recogniser is run on the recordings, giving it the correct transcriptions.

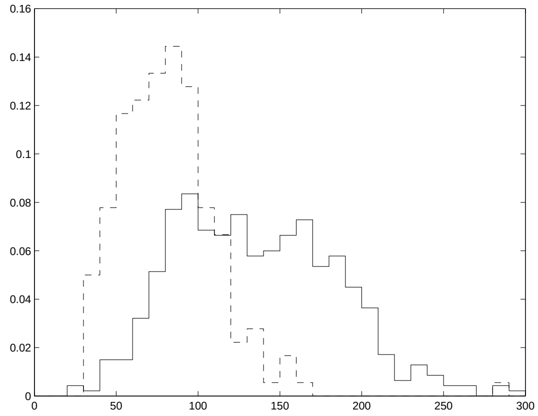


Figure 6.1 *Distributions of durations of /ˈiː/ (solid line) and /iː/ (dashed line).*

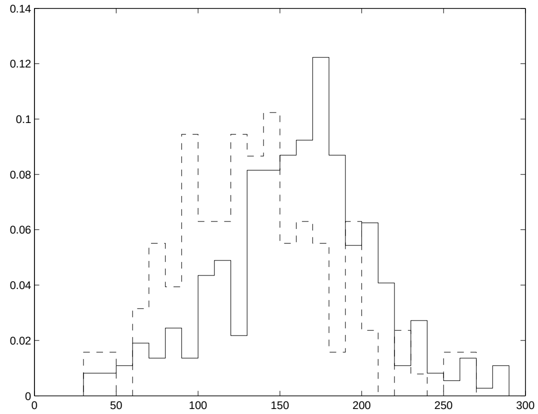


Figure 6.2 *Distributions of durations of /ˈaː/ (solid line) and /aː/ (dashed line).*

If done with a trained recogniser this should give a pretty accurate phone alignment.) No significant result from the fundamental frequency data is found; duration and spectral features, however, do show much separation. This will be demonstrated with histograms that have the probability on the vertical axis and the feature value on the horizontal one.

6.1.1 Duration

The duration statistics are extracted from the forced alignment data. The normalisation procedure is described in section 5.8 on page 64.

Similarly to Sluijter (1995); van Kuijk and Boves (1999), I find that duration is in general a good indicator of stress. Stressed vowels are quite consistently longer than their unstressed counterparts (see figure 6.1). Still, the distribution of /aː/ (figure 6.2) is more muddled than the one van Kuijk and Boves (1999) have. Not all consonants are, as shown in figure 6.6 on page 72: for stops the

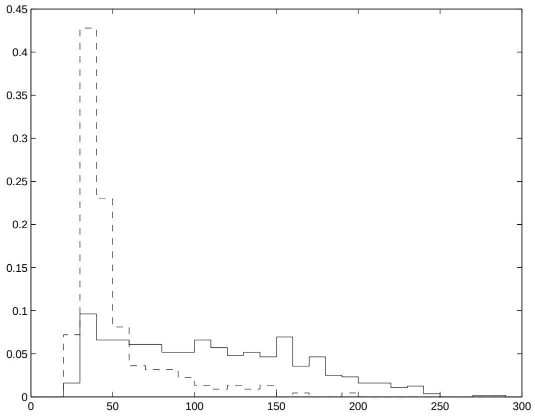


Figure 6.3 *Distributions of durations of /ˈl/ (solid line) and /l/ (dashed line).*

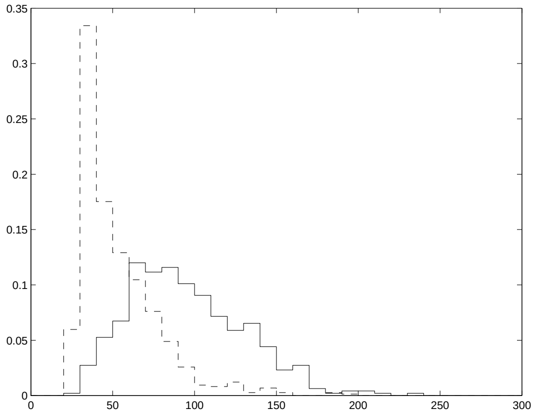


Figure 6.4 *Distributions of durations of /ˈn/ (solid line) and /n/ (dashed line).*

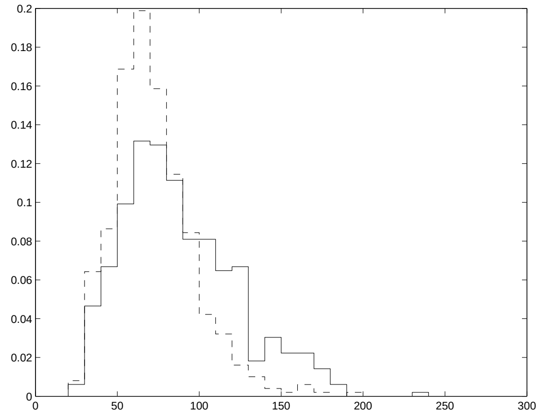


Figure 6.5 *Distributions of durations of /ˈv/ (solid line) and /v/ (dashed line).*

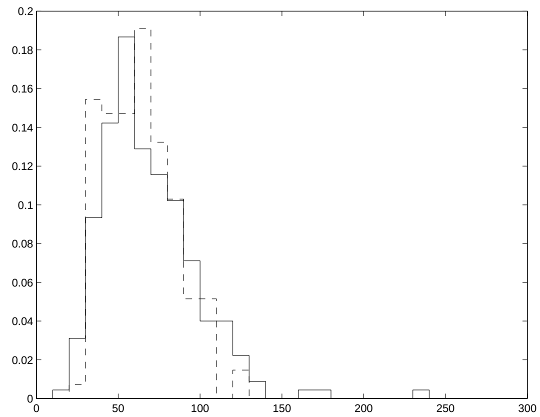


Figure 6.6 *Distributions of durations of /'b/ (solid line) and /b/ (dashed line).*

duration does not seem to differ at all, probably because stops' complete closure makes it difficult to produce lengthened ones sensibly: only the silence would be longer. Liquids consistently show a large difference, as exemplified by /l/ and /n/ in figures 6.3 and 6.4 on the page before, while fricatives are in between (figure 6.5 on the preceding page).

6.1.2 *Spectral tilt*

Different spectral tilt features appear to apply for different phonemes. For many phonemes stress correlates well with some spectral tilt measure. This may be why Sluijter (1995) found clear correlations on a limited set of phonemes, while van Kuijk and Boves (1999) had troubles finding correlates with a limited set of features (two). My results show much more difference than the latter found; this may also be due to the telephone speech they used being spectrally impoverished. Spectral features correlate with stress for vowels, both for long vowels (figure 6.7 on the next page) and for short vowels (figure 6.8 on the facing page).

Most interestingly, the features that work for vowels give similar results for consonants. Figures 6.9 on the next page and 6.10 on page 74 show how stressed and unstressed consonants differ in terms of their spectrum. On the other hand, /n/ (figure 6.11 on page 74) does not show spectral disparity at all. Two factors could play a role here:

- The effect of speaking effort for fricatives and stops on the spectrum may be greater due to their friction-based realisation.

- From a perception perspective, stressed and unstressed /n/ already differ greatly in duration (see figure 6.4 on the preceding page) so the difference in spectrum is not as necessary to distinguish the two.

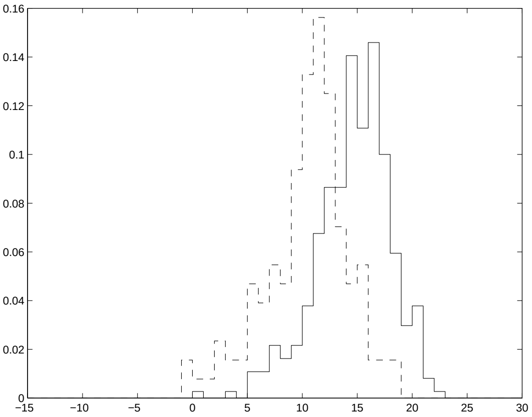


Figure 6.7 *Distributions of the energy in 0.5 – 1 kHz for /ˈaː/ (solid line) and /aː/ (dashed line).*

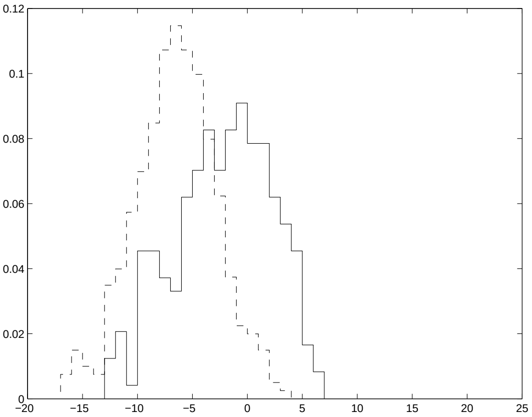


Figure 6.8 *Distributions of the difference between the energy in 1 – 2 kHz and in 2 – 4 kHz for /ˈɪ/ (solid line) and /ɪ/ (dashed line).*

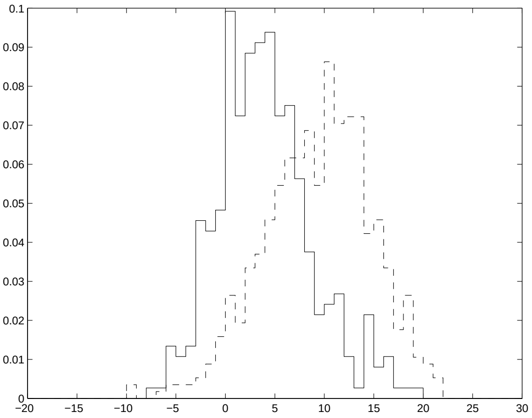


Figure 6.9 *Distributions of the energy in 0 – 0.5 kHz for /ˈd/ (solid line) and /d/ (dashed line).*

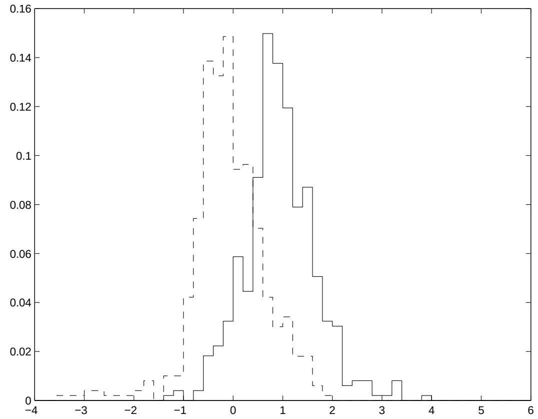


Figure 6.10 *Distributions of the energy delta in 1 – 2 kHz for /ˈv/ (solid line) and /v/ (dashed line).*

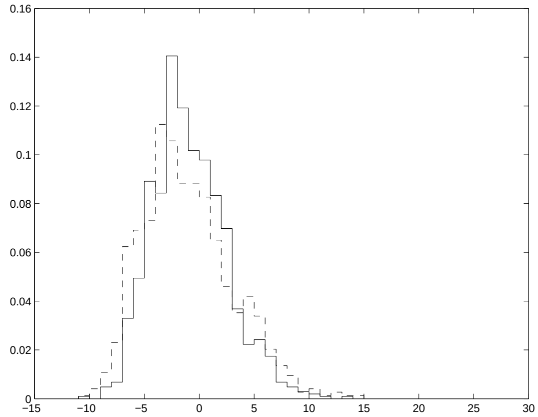


Figure 6.11 *Distributions of the energy in 0.5 – 1 kHz for /ˈn/ (solid line) and /n/ (dashed line).*

Table 6.1 *The confusion between phones of the phoneme-only speech recogniser on* DUTAVSC. *Stress is recognised incorrectly for only very few phones.*

	Short vowels	Long vowels	Consonants	Total
Correct	81.1 %	84.1 %	82.1 %	82.1 %
Incorrect stress	0.9 %	1.5 %	1.9 %	1.6 %
Deleted	5.8 %	3.8 %	1.9 %	5.6 %
Others	12.2 %	10.6 %	10.1 %	10.7 %

6.1.3 *Conclusion*

Lexical stress has been demonstrated to influence acoustically not only vowels, but also consonants. The same features that are canonically associated with stressed vowels (duration, spectral tilt, intensity) are correlates of stress for consonants. Various spectral tilt features apply to various phonemes.

To test whether they help a speech recogniser in determining whether a segment is stressed, a phoneme recogniser without a language model is tested on a subset of the DUTAVSC corpus. The recogniser has a deplorable 43 % overall error rate; but this is mostly due to insertions. Table 6.1 contains the number of substitutions, split into short vowels and long vowels. The recogniser does rather well: only 1.6 % of the correctly recognised phones' stress is classified incorrectly.

6.2 CGN

As discussed in section 5.10 on page 65, two systems are trained. One is a conventional recogniser, which does distinguish between consonants in the onset and in the coda. It will be labelled the "baseline" recogniser throughout. The other is a stress-enabled recogniser, which distinguishes two types of vowels (stressed vs. unstressed) four types of consonants (stressed vs. unstressed and onset vs. coda).

The speech recognisers are not state-of-the-art ones: the HMMs have only three states, no biphones or triphones are used, and they are trained on a corpus with 4.5 % untranscribed material. Furthermore, the feature vectors are not as lean as they could have been: some features may not not help phoneme classification but provide mere noise, and LDA (see section 4.4.1 on page 47) is not used. Therefore, the baseline recogniser is expected to have a pretty low recognition rate. Its purpose in life is not to recognise speech well, though, but to be compared to the stress-enabled speech recogniser.

Similar set-ups have been used by others. Some have not been able to effect an improvement in recognition performance at all (van Kuijk and Boves 1999; van den Heuvel *et al.* 2003). However, Wang and Seneff (2001) made the word error rate decrease from 7.6 % to 7.2 %, which is a relative improvement of 5.3 %.

Notwithstanding the greater number of phoneme models to be trained, the stress-enabled recogniser set up for this work performs better than the baseline recogniser at all points during training. Figure 6.12 and table 6.2 show the recognition rates after every fifth training iteration. The recognition rate of

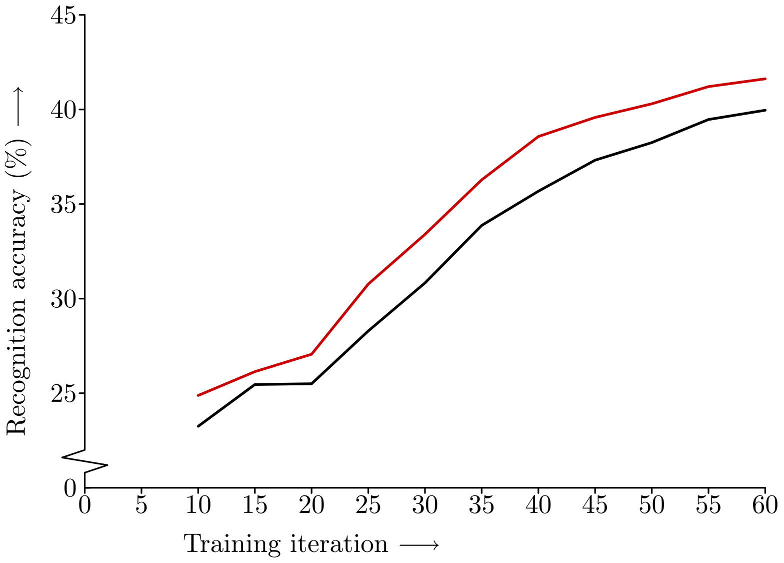


Figure 6.12 *Recognition rates on the evaluation set while training. The black lines describe the baseline system's progress; the red lines the stress-enabled system's.*

Table 6.2 *Recognition rates on the evaluation set while training.*

Iteration	Baseline	Stress-enabled
10	23.25 %	24.88 %
15	25.46 %	26.14 %
20	25.50 %	27.06 %
25	28.29 %	30.77 %
30	30.83 %	33.40 %
35	33.87 %	36.28 %
40	35.68 %	38.57 %
45	37.32 %	39.58 %
50	38.25 %	40.30 %
55	39.47 %	41.21 %
60	39.96 %	41.62 %

Table 6.3 *Recognition rates for the test set.*

	Baseline	Stress-enabled
Correct	30875	31694
Deletions	4647	3499
Substitutions	35820	35659
Insertions	13921	20885
Recognition rate	43.28 %	44.73 %

the stress-enabled system is consistently over 1.5 % higher than the baseline system's. At some points during training, the difference is almost 3 %. Table 6.3 shows the recognition rate and details from iteration 60 on the test set. The difference in recognition rate is 1.45 % as an absolute value. The word error rates are 56.72 % and 55.27 %; this is a relative decrease of 2.6 %. To be able to compare the rates objectively, such parameters as the word insertion penalty were kept at their default values. However, the difference between the numbers of insertions in table 6.3 suggests that there is room for improvement by tuning the word insertion penalty.

Chapter 7

Conclusion

> *Now when this was noised abroad, the multitude came together, and were confounded, because that every man heard them speak in his own language.*
>
> *The Bible*, Acts 2: 6

This last chapter will answer the research questions as formulated in section 1.4 on page 6. This overview will be short; the references in the margin point to places where extensive descriptions can be found. The objective of the thesis work was to model lexical stress in a speech recogniser, and thus to improve recognition accuracy. I made a model of how lexical stress could be used and a proof of concept implementation. The part of the model that was implemented is discussed first; section 7.3 on page 81 deals with future improvements.

7.1 Model

The final objective was to improve recogniser performance by modelling human speech more accurately. Gains that were expected from modelling lexical stress are *1.1 on page 2* *4.2 on page 38*

- better segmentation of continuous speech into words;
- better recognition both of minimal pairs (*óverkomen – overkómen*) and words with similar parts (*octóber – óctopus*);
- more accurate modelling of phone properties.

Properties that are relevant at the symbolic level are *4.3.1 on page 42*

- modelling of reduction, because stressed syllables are less likely to be reduced than unstressed syllables;
- integration of stress recognition with phoneme recognition, as stressed phonemes are pronounced differently; *Figure 4.3 on page 39*
- stress on consonants; hypotheses for stress should be consistent for syllables.

7.1.1 *Acoustics*

6.1 on page 69

The acoustics of lexical stress were found to be consistent with Sluijter (1995) in general. Differences may be ascribed to the laboratory setting in which her research was conducted. The acoustic correlates for consonants had never before been looked into. It appeared that similar features were relevant for consonants.

- Duration was found to be most correlated with lexical stress. This goes for consonants as well, and is quite in line with Greenberg *et al.*'s (2003) results (who looked into phonetic rather than lexical stress, though).
- Spectral tilt was found to be related to lexical stress. Unlike Sluijter's results, the exact frequency bands appeared to vary. Varying frequency features showed better separation than van Kuijk and Boves's (1999) fixed features from telephone speech. Furthermore, consonants showed similar results to vowels.
- The fundamental frequency appeared to be not directly correlated with lexical stress.
- Deltas of feature values had not been discussed by anyone yet. Still, these appeared to be relevant for the spectral tilt features.

7.1.2 *Implementation*

5.5 on page 61

Spectral features are straightforwardly collected from audio data. Energy features are pretty standard in speech recognisers. Phoneme reduction is modelled partially: the MFCC features capture formant frequency differences. However, duration differences and deletion are not modelled.

4.4 on page 45
2.2.5 on page 28
4.3 on page 42
4.4 on page 46

Current-day speech recognition systems typically use feature vectors and HMMs. The features that correlate with lexical stress can be concatenated to feature vectors, except for duration. Duration needs more elaborate modelling. The speech recogniser must also distinguish between stressed and unstressed phones at the symbolic level. All phonemes can be marked as such in the recogniser lexicon, vowels as well as consonants. The latter had not been used in earlier efforts. By limiting the entries in the lexicon to valid syllable structures, the recogniser is limited to matching hypotheses for syllables.

7.2 Discussion

6.2 on page 75

The relative decrease in word error rate was 2.6 %. Thus, the answer to the question "Does it help recognition in practice?" is affirmative. Using lexical stress in speech recognition improves recognition performance.

Table 7.1 on the next page compares my result to that of earlier efforts. Three factors that I believe helped in improving recognition by using stress in a speech recogniser for Dutch are:

- because Sluijter's (1995) results appeared not to generalise directly to natural continuous speech, I picked an array of spectral features;

Table 7.1 *The performance gain of modelling lexical stress from different researchers.*

Work	Language	Improvement
Van Kuijk *et al.* (1996)	Dutch	0 %
Wang and Seneff (2001)	English	5.6 %
Van den Heuvel *et al.* (2003)	Dutch	0 %
This work	Dutch	2.6 % %

- the data were not spectrally impoverished, as I speculate the POLYPHONE data used by the previous efforts using a Dutch-language corpus were;
- I have used consonants, which other researchers have not done.

That Wang and Seneff (2001) got such a large recognition rate improvement may be due to the limited domain of the JUPITER corpus they use, which is from a telephone dialogue system that answers questions about the weather. The POLYPHONE corpus that van Kuijk *et al.* (1996) and van den Heuvel *et al.* (2003) used is more general. However, whereas the part of the CGN I used contains scripted texts, the POLYPHONE corpus has read speech only. Therefore, the difference cannot be explained away through the kind of speech used. *5.10 on page 65*

7.3 Future

In this work a proof of concept implementation of the model has been provided. This section will tie the loose ends about future improvements together.

7.3.1 Acoustics

The data that is input to the speech recogniser may need more analysis. Initially, the spectral features in this work were included for observation. *5.5 on page 61* Experiments with classifiers did not result in any conclusive information on which features were relevant and which were not. Therefore, all features were kept, which may be comparable to using a cannon to kill a mosquito. Using LDA to find a linear dependency and to get rid of the high dimensionality may be possible, though. It may also be useful not to use any stress marks for phonemes for which there is not much difference acoustically between stressed and unstressed versions. This should probably be done using some kind of automatic analysis.

Explicit modelling of reduction in the lexicon has been known to lead to better recognition performance (Kessens *et al.* 1999). My system does not explicitly model reduction, though it is in theory possible to adapt the HMMs' topology to the kinds of reduction the phones can be subjected to. *4.4.4 on page 53* For example, it might be useful for unstressed vowels that can be reduced to /ə/ to optionally share a Gaussian. Sounds that can be reduced to /∅/ could optionally be skipped by linking the first state directly to the final state. These strategies require either a large amount of manual tuning, though, or a tool that automatically performs such analyses. Furthermore, since reduction often depends on the context, explicit reduction modelling may make much more sense when using context-dependent phoneme models. The current system does account *4.3.1 on page 42* for reduction of unstressed phonemes through the normal generalising nature of HMMs.

2.2.5 on page 28 Duration modelling with HMMs and the Viterbi algorithm is not possible. Also, speech recognisers that use duration must keep track of the speaking rate. This requires another architecture than current systems have. However, lexical stress modelling will increase performance more if duration is modelled; and duration-enabled recognisers should gain more when stress recognition is built in. In the future other ways of modelling speech than using HMMs and Viterbi may be explored that make it more beneficial to model the correlation of stress and duration.

7.3.2 *Phonology*

Modelling lexical stress can be seen as inspired by Greenberg's (1999) warning that "the contents of the lexicon needs [*sic*] to accurately reflect the range of phonetic variation observed, otherwise, performance is impaired". However, lexical stress will never be quite the same as phonetic stress. In the current system the acoustic models for stressed and unstressed phonemes have to cover all phonetic variation. *4.4.4 on page 53* If phonetic stress could be recognised separately and through a more or less sophisticated system be transferred to phonological stress, accuracy could be much improved. However, modelling phonetic stress will take explicit training. This will require rigorous manual phonetic transcriptions, like Greenberg *et al.* (2003) have made for part of an English-language corpus. Recognising phonetic stress could be done on a syllable level.

The question is, then, how to derive phonological stress from phonetic stress and how to relate this to the lexicon and the syntactic structure. In my system, *5.7 on page 62* stress is copied directly from the lexicon. Function words may be seen as an exception to that rule; and their not being stressed could be called a poor man's *2.1.1.1 on page 10* prosody model. Proper prosodic modelling could make all the world's difference for the mapping of the phonological and the phonetic levels. To find the prosody from the syntax, though, something about the syntax needs to be known; thus, a multi-pass system would be appropriate.

7.3.3 *There's no two ways about it*

You may have noticed a recurring theme throughout this work; it was introduced in section 4.3 on page 42, and it has occurred in this chapter a number of times. This is the need to model things on two sides at the same time, because modelling only one side is senseless. Those two sides often are the acoustic and the symbolic level. For example: mathematical duration modelling (e.g. Wang 1997) may not improve recognition without building knowledge about stress into the recogniser. Building knowledge about lexical stress into a speech recogniser with very limited duration modelling, while duration is the most important acoustic correlate of lexical stress, will not yield results as good as they could be. Yet I have tried the latter; and I have been able to effect quite an improvement in recogniser performance, solely because there are yet other correlates of stress than duration. Still, recognition would benefit from a more elaborate phonological model, and from a more accurate acoustic model, but more than any of the two, from the two at once.

A similar issue stems from my modelling an extra property of phonemes. Using more accurate phoneme models enhances rather than alleviates the inappropriateness of using one phoneme model for all kinds of speech and speakers — more adaptive and therefore more accurate models could yield better recognition performance. Another area where adaptation is important is duration, which depends much on speaking rate. Clearly, the whole is more than the sum of the parts.

4.3.1 on page 43

7.3.4 *Science fiction*

What will the future bring for speech recognition? I think automated language processing in general will need much more sophistication. The theme I outlined in section 7.3.3 on the facing page — let's call it "Rogier's Law" — will be heard over and over again at different levels, I believe. For example, while processing spontaneous speech syntax may not have much practical use today because no speech recognition system is good enough to figure out all words correctly, such processing is needed just to fill in the words that cannot be heard correctly.

Tying in with that, what parts of utterances are the ones that are pronounced most canonically? Obviously, the important parts, semantically and syntactically; the latter coincide with what I have called "stress". Stressed syllables are perceived clearest; they are the "islands of reliability" (Lea 1980) in a stream of reduced phonemes. Should not a speech recogniser try to reconstruct those words and syllables correctly, and then try to fill in the gaps between? This seems to be what humans do when they perceive "conduct ascends uphill" as "the doctor sends the bill". Probably, the stressed syllables are decoded first: "[..] duct [..] cends [..] hill" becomes "[..] doct [..] sends [..] bill". The unstressed syllables must be reconstructed mainly from other factors than the acoustics: "con [..] as [..] up [..]" does not straightforwardly map onto "the [..] or [..] the". I believe that copying this human behaviour to speech recognisers will enable major recognition improvements. For this, much theory and experience not only with stress will be necessary, but also with spontaneous speech syntax.

1.1 on page 2

A problem preventing speech recognition research from implementing heaps of linguistic features has been the availability of resources. Splitting up phonemes into groups of stressed and unstressed, or groups of those preceded by consonants and those preceded by vowels, or Dutch /r/s pronounced with a uvular fricative and those with a Randstad approximant, or using syntactic knowledge to handle higher levels of prosody, all require data. Training speech recognisers on more groups with more sophistication on more data requires faster computers. The latter requirement is seemingly automatically fulfilled with the passing of time. As for the former: a full-blown corpus of Dutch containing many hours of recording with many different speakers, speaking styles, and conditions, the CGN, was quite recently released. I trust that this release will enable the speech recognition community to take more sophisticated models or linguistic theories, like I have done, and use them to improve speech recognition of Dutch.

Appendix A

Phonemic symbols

The conventions for phonemic symbols used in this work are similar to those in Ewen and van der Hulst (2001) and Collins and Mees (1999). The vowels are in tables A.1 and A.2 on the following page. The consonants follow the obvious conventions. I have used /v/ in Dutch where Collins and Mees use /f̬/.

Table A.1 *The vowels of Dutch (Algemeen Beschaafd Nederlands).*

Checked	**Keyword**	**Free steady-state**	**Keyword**
ɪ	zIt	iː	zIE
ɛ	zEt	yː	nU
ɑ	zAt	uː	mOE
ɔ	zOt	eː	zEE
ʉ	nUt	ø	bEU
ə	werkElIJk	oː	zO
		aː	lA
Free diphthongs	**Keyword**	**Free vowels sequences**	**Keyword**
ɛi	mEI	aːi	sAAI
œy	lUI	oːi	mOOI
ɑu	kOU	uːi	bOEI
		iːu	nIEUW
		yːu	rUW
		eːu	mEEUW

Table A.2 *The vowels of English (Received Pronunciation).*

Checked	**Keyword**	**Free steady-state**	**Keyword**
ɪ	kIt	iː	flEEce
ɛ	drEss	ɑː	pALm
æ	trAp	ɔː	thOUGHt
ɒ	lOt	uː	gOOse
ʊ	fOOt	ɜː	nURse
ʌ	strUt		
ə	bonUs		
Free diphthongs	**Keyword**		
eɪ	fAce		
aɪ	prIce		
ɔɪ	chOIce		
əʊ	gOAt		
ɑʊ	mOUth		
ɪə	nEAR		
ʊə	cURE		
ɛə	squARE		

Appendix B

Paper

This paper was submitted to the Text, Speech and Dialogue conference 2005 as Rogier C. van Dalen, Pascal Wiggers, and Leon J. M. Rothkrantz, “Modelling Lexical Stress”. The layout is changed; this appendix shares the bibliography with the rest of the thesis.

Abstract Human listeners use lexical stress for segmentation and disambiguation. We look into using lexical stress for speech recognition by examining a Dutch-language corpus. We propose that different spectral features are needed for different phonemes and that, besides vowels, consonants should be taken into account.

B.1 Introduction

Prosody is an important part of the spoken message structure. The foundation of prosody of *stress-timed* languages is laid by *lexical stress* (Ewen and van der Hulst 2001). Higher prosodic levels attach to the words at stressed syllables (Ladd 1996).

Lexical stress may be used by listeners to identify words. Though the orthography does not normally encode stress, English has many minimal noun – verb pairs like *súbject – subjéct*, but also pairs like *thírty – thirtéen* or *digréss – tígress* that differ very little except in the stress pattern.

Even though in English and Dutch stress is not on a fixed syllable of the word, in many cases content words do start with a stressed syllable. Listeners use this for segmentation of speech into words (Harley 2001). English-hearing children appear to associate stressed syllables with word onsets at the age of seven months already (Thiessen and Saffran 2003).

Dutch listeners use the stress pattern to identify words before they have been fully heard as well. When hearing the beginning of a word *octo-*, Dutch listeners will decipher whether it is *octó-* or *ócto-* and reconstruct *octóber* or *óctopus* (Cooper *et al.* 2002).

Garden-variety speech recognisers do not use lexical stress, useful though it may be. This paper will describe how it can be automatically detected whether phonemes are stressed. It will be determined what features correlate most strongly with lexical stress, with an eye on how this can benefit speech recognition.

B.2 Related Work

There has been research on the acoustic correlates of lexical stress. Sluijter (Sluijter 1995) in fundamental linguistic research on the acoustic properties of stress minimal pairs demonstrated that lexical stress in English and Dutch is signalled mostly through duration, formant frequencies, intensity, and *spectral tilt*. The latter is a feature that denotes the energy in high frequency bands relative to the energy in low frequency bands. Van Kuijk (van Kuijk and Boves 1999) examined the acoustic properties of a larger corpus of Dutch telephone speech and found similar results: a combination of duration and spectral tilt was the best predictor for lexical stress.

Lexical stress has been used to generate a confidence metric (Bouwman and Boves 2001). From those that have actually used lexical stress recognition in a speech recogniser (van Kuijk *et al.* 1996; Wang and Seneff 2001; van den Heuvel *et al.* 2003), only Wang and Seneff (Wang and Seneff 2001) have been able to effect a performance gain. This is probably what the other authors are after as well; but how this is to be done is not discussed. Van den Heuvel hopes "distinguishing stressed and unstressed vowel models may have a general impact on recognition results."

Notably, none of the authors model lexical stress for consonants, though even textbooks show that stressed and unstressed consonants are realised differently (Ewen and van der Hulst 2001) and though stressed consonants have a longer duration (Greenberg *et al.* 2003). Consonants are influenced by speaking style in the same ways vowel are: duration, spectral tilt and formant frequencies (van Son and Pols 1996). This suggests similar effects can be found for lexical stress on consonants. The closest thing to a rationale for not regarding consonants in automatic lexical stress recognition is the claim that consonants do not carry lexical stress in (Wang and Seneff 2001). This claim is not further motivated, and it will be demonstrated to be incorrect.

B.3 Model

B.3.1 *Objectives*

Since humans use lexical stress in processing speech, modelling it could help speech recogniser performance. We expect the following advantages from using lexical stress.

Phone model accuracy Current speech recognition systems have severe problems coping with speech that is pronounced much faster or slower than the speech it is trained on. Phonemes in unstressed syllables are less often realised canonically than those in stressed syllables. Therefore separating phone models into stressed and unstressed versions may increase predictive strength of the models, improving recognition. For example, unstressed vowels tend to become /ə/[1]. Because the range /ə/–/aː/ is split into into /ə/–/aː/–/ˈaː/, the phone models may become more accurate.

[1] In both Dutch and English.

Word segmentation English hearers, when presented with a faint recording "conduct ascends uphill", will reconstruct words starting at stressed syllables, for example, "the doctor sends a pill" (Harley 2001). Humans use stress for segmentation; a speech recogniser could use this strategy too.

Word recognition Lexical stress signals differences between:

1. words with the same segmental content and different meanings (e.g. Du. *vóorkomen* 'happen' – *voorkómen* 'prevent');
2. words of different categories (e.g. En. *récord* – *recórd*);
3. similar words with different stress patterns (e.g. En. *portráy* – *pórtait*).

B.3.2 Syllables

Lexical stress is specified for syllables as a whole. This poses a problem for speech recognisers, which typically use phonemes as units. Earlier approaches have circumvented this problem by using only vowels for stress detection. When consonants are included as well, their specification must match the vowels' in the same syllable. This can be done by using a consistently stress-marked lexicon: if it contains both /ˈs ˈʌ ˈb d ʒ ɛ k t/ and /s ʌ b ˈd ˈʒ ˈɛ ˈk ˈt/, the recogniser would never hypothesise /s ˈʌ b ˈd ʒ ˈɛ k ˈt/.

In the phonological literature a difference is made between realisations in the coda and in the onset. For example, English /t/ is pronounced as [t^{h}] in *táil*, but as [t] in *rétail* and *líght* (Ewen and van der Hulst 2001): /t/ is only aspirated in the onset of a stressed syllable. We expect that similar effects can be found in the acoustics of lexical stress.

B.3.3 Acoustic representation

To integrate recognition in a speech recogniser, stress can be modelled a phoneme at a time. We look into acoustic correlates of lexical stress that can be fed into a speech recogniser, for example by including them in the feature vectors.

Fundamental frequency Stress is typically thought to be connected to pitch. However, from linguistic literature (Ladd 1996) and literature on automatic stress recognition (Xie *et al.* 2004) it is expected that the fundamental frequency is not straightforwardly correlated with lexical stress. It can straightforwardly be included in a speech recogniser's feature vector though.

Formants Unstressed phonemes can have more reduced realisations than their stressed counterparts; this is visible in the formant values. Standard MFCCs should be able to capture this difference. Note that MFCCs do not directly model formants, but frequency bands. Separating MFCC-based phone models into stressed and unstressed models, whose formant values are confined to a smaller area, will therefore increase MFCCs' ability to recognise the phonemes.

Spectrum The energy in a number of frequency bands can be extracted from the waveform to yield information about the spectral tilt.

Intensity Overall intensity is generally thought to be associated with lexical stress. However, (Sluijter 1995) claims that what is often perceived as loudness variation may actually be spectral tilt: speaking effort would be the common cause.

Duration Lexical stress is generally found to be correlated with phoneme duration (Sluijter 1995; van Kuijk and Boves 1999; Greenberg *et al.* 2003). However, information about phoneme duration is not available during first-pass recognition. Standard HMMs can encode duration through transition probabilities, but this does not work well in recognition. A number of alternatives have been proposed though (Wang 1997; Russell and Moore 1985; Ramesh and Wilpon 1992; Sitaram and Sreenivas 1997).

Derivatives In (Wang and Seneff 2001) it is found that fundamental frequency slope is a better predictor of stress than the raw fundamental frequency. Spectral features are measures for the effort with which phonemes are pronounced. The speaking effort is a continuous measure: it probably increases over the beginning of a stressed syllable and decreases over the end. We therefore expect that derivatives for spectral features also may be correlated with lexical stress, especially for consonants.

B.4 Experimental set-up

We bootstrapped a speech recogniser, made with HTK (Young *et al.* 2002), from Wiggers' system (Wiggers *et al.* 2002) and did measurements on the Delft DUTAVSC corpus (Wojdeł 2003). We used the stress marks from the CELEX lexicon. All phonemes in stress syllables were marked as stressed, except for function words, which were marked as unstressed. All features were normalised over the whole of one utterance. The intensity measure was included in the feature vectors by HTK. For the energy in spectral bands we used the Linux program *sox* and Praat. (Sluijter 1995) chooses spectral bands so that the formants least influence the results; we use the same bands: 0 – 0.5 kHz, 0.5 – 1 kHz, 1 – 2 kHz, and 2 – 4 kHz.

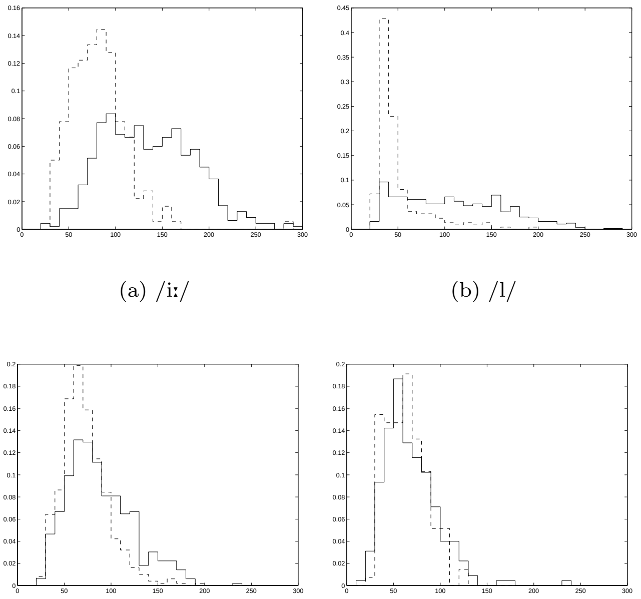
The fundamental frequency was extracted with Praat (Boersma 2001). Where Praat did not find the fundamental frequency, it was linearly interpolated. This has a number of advantages over using an out-of-range value:

- It formalises the notion of the intonational tune in the linguistic literature (Bolinger 1986; 1989; Ladd 1996), where it is pictured as a non-interrupted curve.

- From the linguistic literature, a pitch peak on or near the stressed syllable is expected. Through interpolation, even voiceless phonemes will include pitch information, so that a pitch peak at the onset or the coda of the syllable will be noticed.

- If Praat does not find voicing where there is, linear interpolation provides a reasonable workaround. This increases the algorithm's robustness.

- An out-of-range value, rather than giving the recogniser information about stress, would inject inappropriate information about apparent voicedness.



(a) /iː/

(b) /l/

(c) /v/

(d) /b/

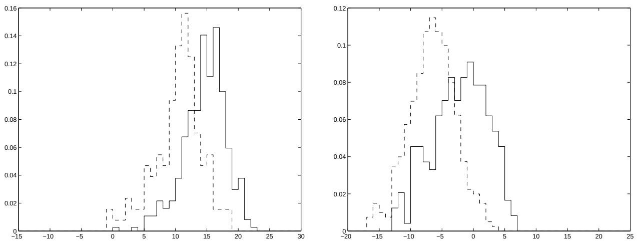
Figure B.1 *Distributions of durations stressed and unstressed (dashed lines) phonemes in ms.*

B.5 Results

The results were assembled by collecting feature vectors from phones that were segmented with forced alignment. The most discriminating features are candidates for inclusion in a speech recogniser that aims at recognising lexical stress. We did not find any significant results from the fundamental frequency data; duration and spectral features, however, do show much separation.

B.5.1 Duration

Similarly to (Sluijter 1995; van Kuijk and Boves 1999), we found that duration is in general a good indicator of stress. Stressed vowels are quite consistently longer than their unstressed counterparts (see Fig. B.1(a)). Not all consonants are, as shown in Fig. B.1(d): for stops the duration does not seem to differ at all, probably because stops' complete closure makes it difficult to produce lengthened ones sensibly: only the silence would be longer. Liquids consistently show a large difference, as exemplified by /l/ in Fig. B.1(b), while fricatives are in between (Fig. B.1(c)).



(a) /aː/: 0.5 − 1 kHz

(b) /ɪ/: The difference between 1 − 2 kHz and 2 − 4 kHz

Figure B.2 *Distributions of spectral tilt features for stressed and unstressed (dashed lines) vowels in ms.*

B.5.2 Spectral tilt

We find that different spectral tilt features apply for different phonemes. For many phonemes stress correlates well with some spectral tilt measure. This may be why (Sluijter 1995) found clear correlations on a limited set of phonemes, while (van Kuijk and Boves 1999) had troubles finding correlates with a limited set of features (two). Our results show much more difference than the latter found; this may also be due to the telephone speech they used being spectrally impoverished. Figure B.2 shows how spectral features correlate with long vowels (as in Fig. B.2(a)) and with short vowels (as in Fig. B.2(b)).

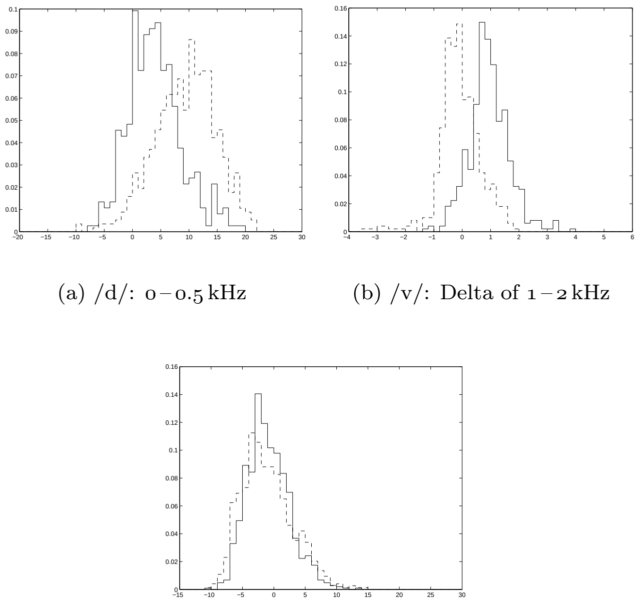
Most interestingly, the features that work for vowels give similar results for consonants. Figures B.3(a) and B.3(b) shows how stressed and unstressed consonants differ in terms of spectrum. On the other hand, /n/ (Fig. B.3(c)) does not show spectral disparity at all. We suspect two factors play a role here:

- The effect of speaking effort for fricatives and stops on the spectrum may be greater due to their friction-based realisation.
- From a perception perspective, stressed and unstressed /n/ already differ greatly in duration (similarly to Fig. B.1(b)) so the difference in spectrum is not as necessary to distinguish the two.

B.6 Conclusion

This paper has described the importance and the feasibility of detecting lexical stress in speech. That stress works on the syllable level can be modelled effectively by adding stress marks to the phonemes in the lexical entries of a speech recogniser.

Lexical stress has been demonstrated to influence acoustically not only vowels, but also consonants. The same features that are canonically associated with stressed vowels (duration, spectral tilt, intensity) are correlates of stress



(a) /d/: 0 – 0.5 kHz

(b) /v/: Delta of 1 – 2 kHz

(c) /n/: 0.5 – 1 kHz

Figure B.3 *Distributions of spectral tilt features for stressed and unstressed (dashed lines) vowels in ms.*

for consonants. Various spectral tilt features apply to various phonemes. Using the duration of a phoneme while it is being recognised is not well possible with Viterbi and standard HMMs. Another algorithm should be used if duration modelling is considered important.

Given the fact that many consonants will participate in the decision whether a syllable is stressed, we hope that implementing lexical stress recognition, even without extensive duration modelling, will improve general recognition performance on three accounts: general phone recognition, word segmentation and word recognition.

Bibliography

M. P. Aylett (2000), *Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech*, Ph.D. thesis, University of Edinburgh.

J. K. Baker (1975), 'The Dragon system — an overview', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1), pp. 24–29.

J. Bilmes (2002), 'What HMMs can do', Tech. Rep. UWEETR–2002–0003, Department of Electrical Engineering, University of Washington.

P. Boersma (2001), 'PRAAT, a system for doing phonetics by computer', *Glot International*, 5, pp. 341–345.

D. Bolinger (1986), *Intonation and its Parts*, Edward Arnold, London.

D. Bolinger (1989), *Intonation and its Uses*, Ph.D. thesis, Stanford University.

G. Booij (1999), *The Phonology of Dutch*, Oxford University Press, Oxford.

A. G. G. Bouwman and L. Boves (2001), 'Using information on lexical stress for utterance verification', in *Proceedings of ITRW on Prosody in ASRU, Red Bank*, pp. 29–34.

P. Carr (1999), *English Phonetics and Phonology: An Introduction*, Blackwell, Oxford.

S. Chang, L. Shastri, and S. Greenberg (2000), 'Automatic phonetic transcription of spontaneous speech (American English)', in *Proceedings of the International Conference on Spoken Language Processing*, pp. 521–524.

N. Chomsky (1995), *The Minimalist Program*, The MIT Press, Cambridge, Massachusetts.

B. Collins and I. Mees (1999), *The Phonetics of English and Dutch*, Brill, Leiden.

V. Cook and M. Newson (1996), *Chomsky's Universal Grammar: An Introduction*, Blackwell, Oxford.

N. Cooper, A. Cutler, and R. Wales (2002), 'Constraints of lexical stress on lexical acces in English: Evidence from native and non-native listeners', *Language and Speech*, 45(3), pp. 207–228.

A. Cutler and S. Butterfield (1992), 'Rhythmic cues to speech segmentation: evidence from juncture misperception', *Journal of Memory and Language*, 31, pp. 218–236.

A. Cutler and D. Norris (1988), 'The role of strong syllables in segmentation for lexical access', *Journal of Experimental Psychology: Human Perception an dPerformance*, 14, pp. 113–121.

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

D. Fenning (1770), *The universal spelling-book, or a new and easy guide to the English language*, S. Crowder, London.

D. Geeraerts (ed.) (2000), *Groot woordenboek der Nederlandse taal op CD-ROM, versie 1.0*, Van Dale, Utrecht.

S. Greenberg (1999), 'Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation', *Speech Communication*, 29, pp. 159–176.

S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang (2003), 'Temporal properties of spontaneous speech — a syllable-centric perspective', *Journal of Phonetics*, 31(3–4), pp. 465–485.

T. Harley (2001), *The Psychology of Language: From Data to Theory*, Psychology Press, Hove.

J. Harris (1994), *English sound structure*, Blackwell, Oxford.

H. van den Heuvel, D. van Kuijk, and L. Boves (2003), 'Modelling lexical stress in continuous speech recognition', *Speech Communication*, 40, pp. 335–350.

J. Higgins (2000), 'Homographs', URL: <http://myweb.tiscali.co.uk/marlodge/wordlist/homogrph.html>.

B. Hyde (2002), 'A restrictive theory of metrical stress', *Phonology*, 19, pp. 313–359.

O. Jespersen (1952), *Growth and Structure of the English Language*, Basil Blackwell, Oxford.

K. J. de Jong, M. E. Beckman, and J. R. Edwards (1993), 'The interplay between prosody and coarticulation', *Language and Speech*, 36, pp. 197–212.

D. Jurafsky and J. H. Martin (2000), *Speech and Language Processing*, Prentice-Hall, New Jersey.

J. M. Kessens, M. Wester, and H. Strik (1999), 'Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation', *Speech Communication*, 29, pp. 193–207.

R. Kompe (1997), *Prosody in Speech Understanding Systems*, Springer.

D. van Kuijk and L. Boves (1999), 'Acoustic characteristics of lexical stress in continuous telephone speech', in *Speech Communication*, vol. 27, pp. 95–111.

D. van Kuijk, H. van den Heuvel, and L. Boves (1996), 'Using lexical stress in continuous speech recognition for Dutch', in *Proceedings of the International Conference on Spoken Language Processing*, vol. IV, pp. 1736–1739.

D. R. Ladd (1996), *Intonational phonology*, no. 79 in Cambridge Studies in Linguistics, Cambridge University Press, Cambridge.

P. Ladefoged (1975), *A Course in Phonetics*, Harcourt Brace, New York.

S. J. Langeweg (1988), *The stress system of Dutch*, Ph.D. thesis, Leiden University.

W. A. Lea (1980), 'Prosodic aids to speech recognition', in W. A. Lea (ed.), *Trends in Speech Recognition*, chap. 8, pp. 166–205, Prentice-Hall, Englewood Cliffs.

M. Liberman and A. Prince (1977), 'On stress and linguistic rhythm', *Linguistic Inquiry*, 8(2), pp. 249–336.

N. Oostdijk (2000), 'The Spoken Dutch Corpus: Overview and first evaluation', in *Proceedings of the International Conference on Language Resources and Evaluation*, vol. II, pp. 887–894.

N. E. Osselton (1984), 'Informal spelling systems in Early Modern English: 1500-1800', in N. F. Blake and C. Jones (eds.), *English Historical Linguistics: Studies in Development*, pp. 123–137, CECTAL, Sheffield.

N. E. Osselton (1985), 'Spelling-book rules and the capitalization of nouns in the seventeenth and eighteenth centuries', in M.-J. Arn and H. Wirtjes (eds.), *Historical & Editorial Studies in Medieval & Early Modern English*, pp. 49–61, Wolters-Noordhoff, Groningen.

D. W. Paulus and J. Hornegger (1998), *Applied Pattern Recognition: A Practical Introduction to Image and Speech Processing in C++*, Vieweg, Wiesbaden.

J. B. Pierrehumbert (2000), 'Exemplar dynamics: Word frequency, lenition and contrast', in J. Bybee and P. Hopper (eds.), *Frequency effects and emergent grammar*, John Benjamins, Amsterdam.

G. Poole (2002), *Syntactic Theory*, Palgrave, Houndmills.

R. F. Port (2003), 'Meter and speech', *Journal of Phonetics*, 31(3–4), pp. 599–611.

P. Procter (ed.) (1995), *Cambridge International Dictionary of English*, Cambridge University Press, Cambridge.

P. Ramesh and J. G. Wilpon (1992), 'Modeling state durations in hidden Markov models for automatic speech recognition', *Proceedings of ICASSP*, 1, pp. 381–384.

S. M. Ross (1997), *Introduction to Probability Models*, Academic Press, San Diego.

M. J. Russell and R. K. Moore (1985), 'Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition', *Proceedings of ICASSP*, 10, pp. 5–8.

J. E. Shoup (1980), 'Phonological aspects of speech recognition', in W. A. Lea (ed.), *Trends in Speech Recognition*, chap. 7, pp. 125–138, Prentice-Hall, Englewood Cliffs.

R. N. V. Sitaram and T. Sreenivas (1997), 'Incorporating phonetic properties and hidden Markov models for speech recognition', *Journal of the Acoustical Society of America*, 102(2), pp. 1149–1158.

A. Sluijter (1995), *Phonetic Correlates of Stress and Accent*, Ph.D. thesis, Leiden University.

R. van Son, O. Bolotova, M. Lennes, and L. C. Pols (2004), 'Frequency effects on vowel reduction in three typologically different languages (Dutch, Finnish, Russian)', in *Proceedings of INTERSPEECH 2004, Jeju Island, South Korea*, pp. 1277–1280.

R. J. J. H. van Son and L. C. W. Pols (1996), 'An acoustic profile of consonant reduction', in *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, pp. 1529–1532.

J. Taglicht (1998), 'Constraints on intonational phrasing in English', *Journal of Linguistics*, 34, pp. 181–211.

E. D. Thiessen and J. R. Saffran (2003), 'When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants', *Developmental Psychology*, 39(4), pp. 706–716.

C. Wang and S. Seneff (2001), 'Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain', URL: <citeseer.nj.nec.com/453327.html>.

X. Wang (1997), *Duration modelling in HMM-based speech recognition*, Ph.D. thesis, University of Amsterdam.

X. Wang, L. C. W. Pols, and L. F. M. ten Bosch (1996*a*), 'Analysis of context-dependent segmental duration for automatic speech recognition', in *Proceedings of the International Conference on Spoken Language Processing*, vol. II, pp. 1181–1184.

X. Wang, L. C. W. Pols, and L. F. M. ten Bosch (1996*b*), 'Integration of context-dependent durational knowledge into HMM-based speech recognition', in *Proceedings of the International Conference on Spoken Language Processing*, vol. III, pp. 1073–1076.

N. Warner, A. Jongman, A. Cutler, and D. Mücke (2001), 'The phonological status of Dutch epenthetic schwa', *Phonology*, 18, pp. 387–420.

P. Wiggers (2001), *Hidden Markov models for automatic speech recognition and their multimodal applications*, Master's thesis, Delft University of Technology, Delft.

P. Wiggers, J. C. Wojdel, and L. J. Rothkrantz (2002), 'Development of a speech recognizer for the Dutch language', *Proceedings of 7th annual scientific conference on web technology, new media, communications and telematics theory, methods, tools and applications (EUROMEDIA)*, pp. 133–138.

J. Wojdel, P. Wiggers, and L. Rothkrantz (2002), 'An audio-visual corpus for multimodal speech recognition in Dutch language', in *Proceedings of the International Conference on Spoken Language Processing*.

J. C. Wojdeł (2003), *Automatic Lipreading in the Dutch Language*, Ph.D. thesis, Delft University of Technology, Delft.

H. Xie, P. Andreae, M. Zhang, and P. Warren (2004), 'Detecting stress in spoken English using decision trees and support vector machines', in *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation*, pp. 145–150, Australian Computer Society, Inc.

S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland (2002), 'The HTK book (for HTK version 3.2.1)', URL: <http://htk.eng.cam.ac.uk/docs/docs.shtml>.