Cues to Vowels in the Aperiodic Phase of English Plosive Onsets

Kaj Christian Nyman

MPhil

University of York
Language and Linguistic Science

September 2011

Abstract

This thesis addresses the problem of vowel recognition in coarticulatory theory and phonology by assessing how early vowel quality can be recognised from English onset plosives realised with aspiration. Particular attention is paid to aspects of production and perception timing. A gating experiment was used to assess how reliably listeners can recognise English monophthongs.

The treatment of coarticulation distinguishes between phonetic and phonological aspects of production and perception, with a clear demarcation between these levels of representation. The results are interpreted through the lens of prosodic phonology, as this framework constrains the grammar more optimally than segmental-phonemic ones and better exemplifies listeners' sensitivity to the distribution of FPD.

Velar and bilabial onsets give rise to significantly more correct responses than alveolars, which require more precise articulations. High vowels are recognised more reliably than low ones. This result is due to their intrinsically shorter duration, making high vowels less variable through time. This perceptual link is proportionate to the total amount of variation in vowel inherent spectral change (VISC), which corresponds to spectro-temporal variation in formant centre frequencies through time in vowel realisations. Nasal rimes give rise to a smaller proportion of correct responses than non-nasal rimes, especially in the context of high and low front vowels: the VISC and changes in vowel height undergone in the context of such articulations, as well as the phonetic consequences of the overall articulatory constellation shape the resulting percept. CVCs with non-nasal rimes give rise to more correct responses than CVVs, despite there being more articulations on-going: the shortness of the vowel in CVCs compensates for this deficit, making perception more robust. Word frequency does not have a significant effect on recognition for any of the syllable types investigated.

Overall, a much larger temporal window than the phoneme is required for the robust processing and perceptual integration of speech. Phonemes alone cannot adequately define how the relationship between the phonetic co-extensiveness of different sounds and feature sharing is to be accounted for in speech understanding. Since articulators are in constant motion during production, and consonantal gestures have distinctive coarticulatory influences over vocalic ones, the formant frequencies for both types of sound are in constant flux. This variation reinforces perceptual cohesion and has systematic effects on the mapping of FPD, through which larger structures become audible.

Table of Contents

List of Figures

List of Tables

Preface

The aim of the present work is to contribute to coarticulatory and phonological theory. This main aim is accomplished by examining and analysing the results of a perception experiment on vowel recognition from English aspirated voiceless plosives.

Chapter 1 sets up the study and assesses the relevance of the study of the perception of coarticulation and why a polysystemic and mainly non-segmental understanding of coarticulation better exemplifies the complex perception-production mapping that coarticulation requires. In the latter half of the chapter it is shown a) how studying properties of the aperiodic phase of voiceless plosives can contribute to the understanding of coarticulatory and phonological theory and b) what the relationship of this question to previous literature is like. It is also explained why the chosen methodology is appropriate.

Chapter 2 brings together the three main strands of literature relevant to assessing vowel recognition from English aspirated plosives as produced in the context of real English word forms. The first strand focuses on the phenomenon of VISC in vowel sounds and general properties of vowel timing. The second subsection comprises a detailed review focusing on five themes that emerge as guidelines from similar smaller scale studies on vowel recognition timing. The third review section comprises a detailed summary of non-segmental phonological modelling of vowel recognition. The next section contains an evaluation of the previous studies on similar studies of vowel recognition timing in English. The distinctive methodologies of previous smaller scale studies are contrasted with the choices made in this study, the most important of which is working with real

words rather than nonsense syllables. In the final two sections of the chapter, the secondary research questions and hypotheses arising out of the research questions presented in the previous chapter are described and accounted for.

Chapter 3 describes the gating experiment on which the study is based and details the rationale for the choices made in it. The methods implemented in the analysis of the CV(V)/C production data are substantiated and described.

Chapter 4 focuses on the results and describes the findings from the viewpoint of the same themes as detailed in the second literature review section in chapter 2. The results are assessed statistically from each theoretical perspective.

Chapter 5 firstly describes the relationship of the main findings with previous literature and how the study extends those findings, as well as how the results align with the aims and hypotheses outlined in the first two chapters. The second and more substantial part of the chapter outlines a detailed model of phonological processing and vowel recognition timing for English monosyllabic utterances. The final two parts of this chapter exemplify the workings of that model by applying representative results from the previous chapter to it. These results include exemplification of the main trends of phonetic interpretation from both male and female production data for CVVs, as well as a more general presentation of modelling for CVNs. The aim of this chapter is to offer as explicit answers to the primary and secondary research questions as possible.

Chapter 6 sums up the results and their implications for the study of coarticulatory and phonological theory. The body of

the text is drawn to a close by detailing further research questions stemming from this research project.

Acknowledgements

I wish to thank my supervisor Dr. Richard Ogden, who has supported me throughout the preparation and writing of this work. The critical and constructive comments and support received from him have been invaluable in terms of completing the research project. Other key contributors are Paul Foulkes and John Local on my thesis research panel and Barry Heselwood (the University of Leeds). I also wish to thank the following people for additional aid and insightful comments concerning this research: Rachel Smith, Jane Stuart-Smith, Oliver Niebuhr, Sam Hellmuth, Marilyn Vihman, Tamar-Keren Portnoy, Dominic Watt, Carmen Llamas, Gerard Docherty and Huw-Llewelyn Jones.

I wish to acknowledge the contribution made by all the speakers and listeners who took part in the project, for which they deserve special thanks. Lastly, I wish to thank my parents and family, who have acted as a continuous source of support throughout the project.

Author's Declaration

Appropriate references in the text are made to Hawkins and Smith (2001), Hawkins and Nguyen (2001), Fowler and Saltzman (1993), Local (2003), Chapman and Routledge (2005), Sprigg (2005), Hawkins (2003, 2010a), Hanson and Chuang (1999), MacMillan et al (1999), Coleman (1998), Stevens (1998) and Local and Ogden (1997). The study does not contain any substantive amount of material from these works, however. Some of the main ideas discussed in the thesis and the perception experiment have also appeared in an earlier and more concise form as Nyman (2010).

1. Background

1.1 Studying the Perception of Coarticulation within the Context of Vowel Recognition from Aspiration

One of the most interesting things about speech is the way different constituents and features combine together and the kind of fine phonetic detail (FPD) this engenders, especially with respect to the ways in which speech production and perception unfold through time. Hawkins and Smith (2001, p. 107) give a good example of the kind of FPD that occurs in the context of systematic phonetic variation. This kind of variation between structurally similar word forms displaying small distinctions that reflect specific combinations of linguistic properties is shown in figure 1:

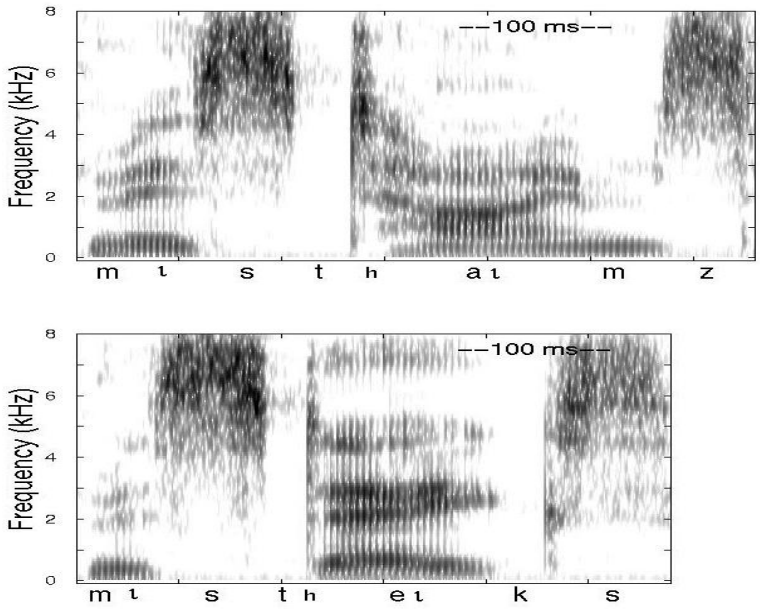
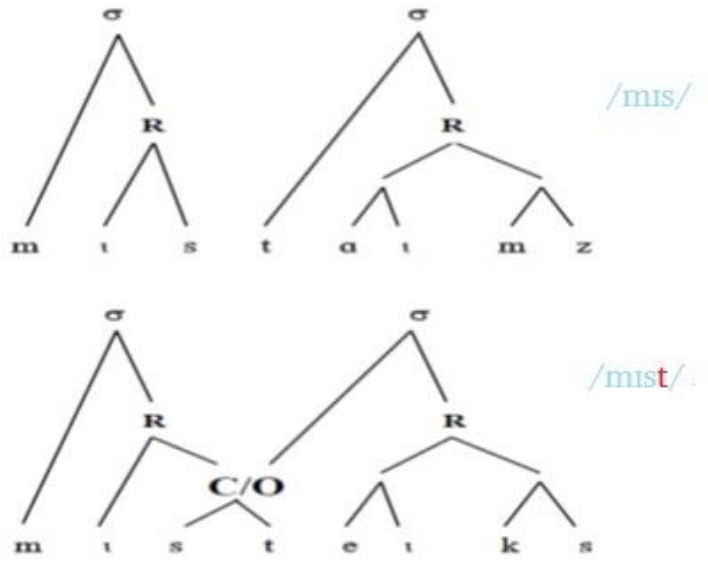


Figure 1: Spectrograms of the words 'mistimes' (top) and 'mistakes' (bottom) spoken by a British English woman in the sentence 'I'd be surprised if Tess _____ it' with main stress on Tess (Hawkins and Smith, 2001, p. 106, fig 1).

In figure 1 can be seen spectrograms of the word forms 'mistimes' (top half) and 'mistakes' (bottom half). A look at

the FPD of the first syllable comprising three sounds (=/mɪs/) in each utterance reveals interesting realisational differences between the two lexemes. For example, the /t/ in ‘mistimes’ has longer aspiration and a longer closure phase, whereas the same sound in ‘mistakes’ has more brief aspiration and a shorter closure. The sibilant /s/ of ‘mistimes’ is shorter, and /m/ and /ɪ/ are of longer duration than in ‘mistakes’: this distinction can be heard as a rhythmic difference, with the first syllable of ‘mistimes’ having a heavier beat than that of ‘mistakes’ (Hawkins and Smith, 2001, p. 108). Such subtle phonetic distinctions can be explained by the structural differences between the two word forms. The structures of these two word forms are described in figure 2:



[1]

Figure 2: The syllable structures underlying the words ’mistimes’ (top) and ‘mistakes’ (bottom)

(Adapted from Hawkins and Smith, 2001, p. 106, fig. 1)

Such realisational distinctions between utterances with

[1] The red colour used in the illustration of ‘/mist/’ is meant to display the structural distinction between ‘mistimes’ and ‘mistakes’.

similar phonological shapes that reflect specific combinations of linguistic properties are considered as FPD[2] in this thesis. The combinations of such linguistic properties make the structures audible. The small distinctions reflecting such phonetic combinations may occur at different levels of linguistic structure. For example, they may be relevant to morphology and stress patterns (e.g. Hawkins, 2003, pp. 390-391). This kind of interaction between different facets of linguistic structure contributing to phonetic exponency and phonological processing is considered as 'polysystemicity'. Polysystemicity can be defined as the interaction between different linguistic systems in language (see e.g. Hawkins and Smith, 2001, p. 112). It reflects a view of language that allows phonetic, phonological, morphological, semantic, syntactic and other linguistic systems to influence each other in the processing of a message from muscle movements into a richly acoustically structured speech signal, which is interpreted as a meaningful utterance.

For this theory to work, rich structures are needed, along with a rich theory of phonetic interpretation. For example, systematic differences in the FPD of /uː/ and /ɑː/ in 'who's sharpened the meat cleaver?' and 'who sharpened the meat cleaver?' may enable listeners to distinguish the grammatical differences between the two utterances (Hawkins and Smith, 2001, p. 109). Figure 3 gives an example of this type of a distinction in the vowel sounds in /uː/ and /ɑː/:

[2] Little attention is paid here to the theoretical associations related to the wording of this term (see e.g. Hawkins, 2010a and Carlson and Hawkins, 2007), since 'fine phonetic detail' best captures the mutual dependency between subtle aspects of perception relating to coarticulation and FPD on the one hand and phonological processing on the other.

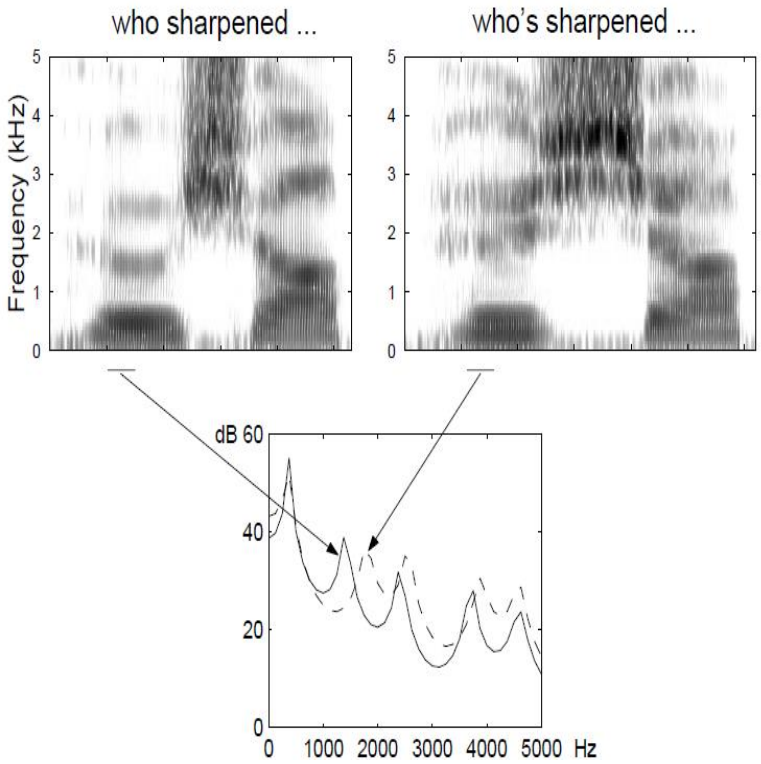


Figure 3: spectrograms of the first two syllables from 'Who sharpened the meat cleaver?' (left) and 'Who's sharpened the meat cleaver?' (right), and LPC spectra of the first part of the realisation of /u:/

Top: spectrograms of the first two syllables from 'Who sharpened the meat cleaver?' (left)and 'Who's sharpened the meat cleaver?' (right). Bottom: 50-ms LPC spectra (18-pole autocorrelation, Hanning window) of the first part of the vowel in 'who' and 'who's', as indicated by the arrows: solid line spectrum from 'who'; dashed line spectrum from 'who's'. The horizontal lines under the spectrograms indicate the 50-ms portions of the signal over which the spectra were made.

(Hawkins and Smith, 2001, p. 109, fig. 2)

Figure 3 shows the spectrograms of the two utterances 'who sharpened the meat cleaver?' (left) and who's sharpened the meat cleaver?' (right), as well as LPC spectra of the first part of the realisation of /u:/, which can be distinguished by inspecting the acoustic detail in the two spectrograms in figure 3 (for example, F2 is higher throughout in the 'who's sharpened...' variant). The significance of the type of FPD displayed in figure 3 is not the fact that the words at the beginning of the utterances contain a different number of phonemes near the beginning parts of the two utterances. Rather, Hawkins and Smith (2001, p. 108) contend that:

> "A conventional analysis would say that the /z/ is fully assimilated to the place of articulation of the following fricative. However, the assimilation is not complete, because the two /u/ vowels before the fricatives are very different, in ways consistent with alveolar versus palatal-alveolar articulations. The panel at the bottom of figure 2 shows lpc spectra from 50-ms windows at the beginning of each vowel, as indicated by the connecting lines to the two spectra. Both F2 and F3 are considerably higher in frequency in 'who's' than in 'who'. That is, an 'underlying /z/' engenders higher F2 and F3 frequencies in the preceding /u/ and, of course, in the /h/ preceding the /u/... Slightly raised F2 and F3 frequencies may not have a strong effect by themselves, and they are not always present in such sequences; but when they are there, and especially when they co-vary with other cues such as duration of the fricative, they could offer good information about the grammatical structure of the utterance".
>
> (Hawkins and Smith 2001, p. 108-109)

In sum for figure 3, a relatively small grammatical difference in two 5/6 word utterances pertaining to only one sound affects the phonetic organisation and detail of the surrounding vowels. Such detail could be used by listeners to distinguish the sounds most closely affected by the coarticulatory-structural difference (i.e. /u ɑ: z ʃ)/. The differences may be used to distinguish upcoming sounds and/or larger structures from the exponents of /u:/: Hawkins and Smith (2001, p. 109) suggest that listeners may use this type of FPD to enhance perceptibility. The example in figure 3 exemplifies the significance of distinctions

which provide evidence for polysystemic accounts in the perception of coarticulation.

Conventional theories such as Motor Theory and Articulatory Phonology (see e.g. Liberman and Mattingly, 1985 and Browman & Goldstein, 1986) pay less attention to the significance of the type of FPD displayed in figures 1 and 3, especially to the extent that it can be used to enhance perceptibility. There is less room in conventional theories for the type of polysystemic thinking advocated by e.g. Hawkins and Smith (2001), since conventional theories are built on models of language and linguistic processing that do not allow the phonology access to other components of the grammar. Conventional theories and frameworks tend to be segmental-phonemic and do not usually allow for perceptually significant long-domain coarticulatory phenomena, especially in structurally non-complex utterances such as CV(V)/Cs.

Conversely, in the examples provided in figures 1-3, listeners need to have access to a sufficient amount of FPD in order to be able to work out the syllabic, morphological and grammatical relationships between phonetically similar but structurally distinct utterances. Having provided a framework for analysis and theoretical discussion, the next section will describe and substantiate the primary research question.

1.2 Research Questions

The main research question of this thesis asks: at what temporal point during the aperiodic phase of an English voiceless plosive can vowels be reliably recognised from English utterance-final CV(V)/Cs? That is, this research examines how early listeners can recognise vowels from aspirated plosives in real English word forms.

On the surface the thesis question may seem too simple. In theoretical terms, however, it is far from simple. This

research targets English as spoken in England, for which this particular aspect of coarticulation has never before been researched in this respect. All previous studies on the topic are on North American varieties and smaller in scale (e.g. Cullinan and Tekieli, 1979 and Winitz et al, 1972).

The reason why this research question is theoretically important is that it intersects research on coarticulation, perception timing, phonological processing and representation. The discussion thus far has shown that coarticulation spans many levels of structure, even in structurally relatively simple utterances. Answering the main research question may be the first step in building a more reliable model of coarticulation which is sensitive to the perceptual implications of polysystemic phenomena.

Since the vast majority of previous studies on coarticulation do not allow for the type of interaction between different linguistic systems as advocated by e.g. Hawkins and Smith (2001, see figures 1-3), coproduction theory can be seen as incapable of fully predicting such results and phenomena. It is argued in this thesis that vowel perception/production timing in utterance-final CV(V)/Cs also displays some of the type of polysystemic properties as advocated by e.g. Hawkins & Smith (2001), even in the absence of interaction between different linguistic systems in the grammar. It is possible to view vowel recognition timing in CV(V)/Cs as a basic indicator of linguistic-phonetic interaction between coarticulation and phonetics on the one hand and the phonology on the other. Conventional theories are usually incapable of providing an adequate account of such coarticulatory and polysystemic phenomena related to speech perception. For example, the findings on ‘led’ and ‘let’ by Hawkins and Nguyen (2001, 2004) are a good example of this theoretical issue, since the duration and darkness of the onset laterals differ significantly. Very small distinctions in FPD in such single word

monosyllabic utterances can significantly affect their recognition, to the extent that given FPD in an articulatory complex onset or coda is more closely associated with one word form than another. Such a question remains very important for this research (especially concerning phonological processing), albeit that not the same types of onsets are investigated as in previous non-segmental polysystemic studies of speech perception. In sum, the approach taken in this research is more comprehensive in terms of modelling such phenomena.

In this study, particular interest is paid to the perceptual features of coarticulation. The thesis question is approached through the lens of a perception experiment: the gating experiment (see e.g. Grosjean, 1996) investigates how early upcoming vowels can be recognised from aspirated plosive onsets. The chosen methodology is appropriate, as gating experiments allow looking at the perception of coarticulation in pinpoint accuracy (Grosjean, 1996, p. 601). Gating allows stimuli of different duration to be prepared, making it possible to show how percepts are updated through time. Gating experiments are easily implemented (Grosjean, 1996, p. 601 and Shockey, 2003) and non-invasive. Other paradigms that might be used, such as eye-tracking and brain imaging might cause participants discomfort and interfere with the ability to interpret and generalise results. Whether considering things from a theoretical or a practical viewpoint, the gating paradigm may be more reflective of online phonological processing (also cf. Grosjean, 1996, p. 601-2) than using an alternative methodology.

Since no previous research (e.g. Cullinan & Tekieli, 1979, LaRiviere et al, 1975, Waldstein and Baum, 1994, and Ostreicher & Sharf, 1976) related to the main research question has looked at both long and short vowels, using real word CV(V)/Cs with both long and short vowels as stimuli may

allow extending and/or supplementing the findings of previous research. It is not clear from previous research whether cues to vowels are as readily available in long vowels compared to short ones. The main research question may allow extending our understanding of coarticulation. Having described the main research question of this study, the generalisations that can be drawn from it are outlined.

1.2.1 Research Questions: Generalisations on this Study and its Relationship with Previous Research

Perception timing/temporal dynamics and the perception of coarticulation for non-standard varieties of English constitutes a significant gap in empirical theory and is in line with aspects of current research into speech perception, phonology and phonological representation, as well as the significance of indexical variation in processing. This statement does not just apply to English varieties spoken in England, but for the English language more generally.

The previous experiments on North American English varieties may not be fully applicable to British English, since there are systematic differences in the phonetic exponency of equivalent sounds between different varieties. In particular, the strategies implemented in the timing of voicing differ between British and North American varieties (see e.g. Docherty, 1992, pp. 25 and 113-114) and speakers of these two varieties may employ different types of FPD in the aperiodic phases of onset plosives (e.g. Wells, 1982). Since it has been established thus far that FPD in consonant-vowel transitions may significantly affect the FPD and processing of upcoming sounds in monosyllabic words (see e.g. Goffman et al, 2008 and West 1999b), this claim can be justified. To briefly assess to what extent the results of this study will be compatible with those of previous studies, figure 4 provides production data on the

acoustic differences between Standard Southern British English (SSBE) and General American (GA) monophthongs (the SSBE values in black are based on the vowels of a female speaker from South London; GA values in red represent GA female values):

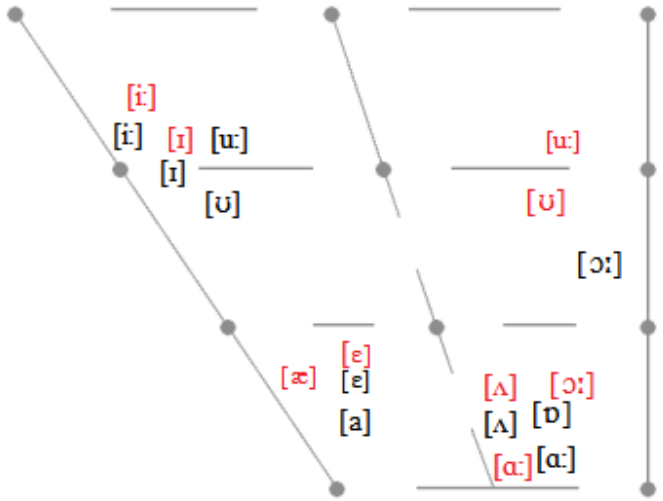


Figure 4: Typical values for monophthongs in SSBE and GA
American values are based on Hillenbrand et al., 1995, p. 3103

Despite the kinds of differences in exponency displayed in figure 4 (cf. e.g. the large acoustic differences between GA and SSBE /uː ʊ/ and /ɔː/, see e.g. Wells, 1982), the two varieties are mutually intelligible and share important features in their phonologies. The results of this study will probably in many ways be similar to those of previous studies (such as Ostreicher & Sharf, 1976, Tekieli & Cullinan, 1979 and Cullinan & Tekieli, 1979). Despite these similarities, the dynamicity and time course of perception may differ across the two varieties, since qualitatively different coarticulatory and listening strategies may be required due to the phonetic differences between vowel sounds in the two varieties (and to a somewhat lesser extent onsets and codas). Differences in vowel duration between GA and English varieties (as spoken in

England) may correlate with the potential differences in the timing of vowel recognition, since duration can affect vowel quality and the time course of vowel processing (see e.g. Rosner & Pickering, 1994, p. 295). Any systematic differences between GA and English varieties with respect to the durational patterning of vowels might affect the timing of vowel recognition in distinctive ways, albeit that it is beyond the scope of this study to give a detailed account of this issue.

Having described the purpose and research questions of this study, a general account of research studies on coarticulation will be given in the next section.

1.3 General Accounts of Coarticulation

Theoretical accounts of coarticulation are varied. There are at least three general schools of thought on coarticulation. The claims made in them are not mutually compatible:

- Some researchers (e.g. Daniloff and Moll, 1968 and Henke, 1966) suggest frameworks where activities of different articulators are viewed as feature look-ahead models: the speech production planning process scans ahead in time in its phonetic implementation. Features such as lip rounding that are not acting in opposition to other features will be inserted as early as possible and carry over as long as they are not met by any other conflicting features along the way.

- Other researchers have emphasised phoneme-specific effects that may bring about changes in the extent of coarticulatory phenomena (e.g. Recasens, 2002 and Günther, 2003). Such effects are broad and sometimes discontinuous, since they can be

interrupted by competing gestures for other phonemes.

- Other researchers take a gestural view of coarticulation. Such models emphasise the role of interaction between qualitatively distinctive consonantal and vocalic gestures (= coproduction). For example, Fowler and Saltzman (1993, p. 185) argue that coarticulatory effects "do not extend very far backward in time from the period of a gesture's own predominant interval". Fowler (2006) and Fowler and Saltzman (1993) argue for effects being more temporally fixed than other researchers do.

Despite some evidence of adaptive effects (e.g. Recasens, 2002) in the types of views of coarticulation described in this subsection, rather few experiments have investigated coarticulatory effects beyond the phoneme. Despite the limited scale of this study in terms of the structures investigated (CVV, CV-plosive and CVN) and the extent to which interactions between different linguistic systems therein might be displayed, it is still possible look at coarticulation beyond the phoneme in CVCs and/or CVNs. This type of a question is relevant to the extent that the phonetic exponents of the aperiodic phase are significantly affected by long-domain coarticulation from the coda portion onto parts/properties of the onset (also see Coleman, 1998, p. 224). For this reason, a key question to ask in this research is whether it is possible to reconcile some of the contrasting findings of e.g. phonemic and gestural studies on coarticulation with ones having a non-segmental framework. The theoretical framework applied therein could e.g. be the type of polysystemic framework described in the previous subsection. The main issue here is not whether vowel targets and other sounds are phonemic or that

polysystemic frameworks often are non-segmental. Rather, the contention made is that vowel perception timing in CV(V)/Cs can be seen as an indicator of linguistic interaction between coarticulation and phonetics contra phonology (for example, if it is found that vowel recognition timing differs for CVVs, CVCs and CVNs). In this particular respect, this research delves deeper than most previous studies on coarticulation, and in particular compared to conventional theories. If long-domain phenomena pervade or are found to be significant in plosive-V(V)-C monosyllabic words polysystemic models will receive further support.

In theoretical terms, the framework presented in this chapter and the rest of this thesis relies more on non-segmental structures and long-domain coarticulatory properties (i.e. contrasts whose phonetic exponents spread over more than one sound) and phenomena than previous models do. This choice can be explained by the fact that many recent pieces of research (such as Hawkins and Nguyen, 2001 and West, 1999b) show that non-segmental attributes of speech can play an important role in the perception of coarticulation. In sum, although the approach taken in this research is largely non-segmental, the ideas behind the applied framework touch upon a much wider range of representational and structural issues than the size and/or types of perceptual targets speaker-listeners aim at in speech processing.

However, in order to be able to show the relevance of the type of systematic phonetic variation for the types of utterances described in figures 1-3, there is a need to adequately demarcate the relationship between articulation and acoustics on the one hand and perception and phonological processing on the other. For instance, is there a direct correspondence between the encoding of phonological features into acoustic cues and the subsequent perceptual mapping reflecting phonological processing?

The variety in the findings of the kinds of coarticulation studies mentioned at the beginning of this subsection stems from a lack of a comprehensive model on coarticulation. There is not yet a model available that provides a mapping between abstract linguistic units contra their phonetic realisation through muscle activity and movements. Goffman et al (2008, p. 1424) show that the acoustic effects of lip rounding may extend across several parts of an utterance and have systematic effects upon its acoustics. The findings of Goffman et al offer support to the hypothesis that there is an initial planning for broad chunks of output by the onset of speech production. This finding on coarticulatory closely reflects the discussions and examples provided in this thesis, since it suggests that if there is a change in a single phoneme, the motor commands to the muscles are altered for the production of the whole utterance. As far as this issue relates to phonological processing, polysystemicity in speech perception (cf. e.g. 'who/who's sharpened the meat cleaver' example in figure 3) is deemed a more important theoretical parameter in this thesis than the non-segmental framework. In answering the main research question, it is very important to transcend the debate on the size and shapes of perceptual units in speech perception (cf. e.g. Goldinger and Azuma, 2003). A stronger focus on polystemicity and an adequate emphasis of how linguistic systems may interact in speech production/perception can help to bridge the divide between phonemic and prosodic models. A stronger focus on polysystemicity and linguistic interactions may be even more significant for reconciling contrasting findings on coarticulation (such as those between coproduction and look-ahead models contra purely phonemic or articulatory models). The findings of e.g. Hawkins and Smith (2001) and Goffman et al (2008), which show that coarticulation simultaneously spans several levels of production/processing, substantiate this claim.

The claims made by Goffman et al (2008) are consistent with prosodic phonology, since the results provide good evidence for the fact that speech production units span multiple levels. Multiple units including syllables, words and phrases are mapped and co-ordinated with and against each other. There is a need to consider the effect of FPD on the perceptibility of linguistic structures. For example, let us assume a sequence of structurally similar sentences which differ only in the features assigned to the vowel sound in the final word in each utterance. The waveforms and spectrograms of ‘core’, ‘car’, ‘coo’ in figures 5-7 display the sounds produced in three plosive-monophthongal vowel utterances: each figure contains a waveform for each word form at the top and a spectrogram at the bottom (each spectrogram includes formant tracks):

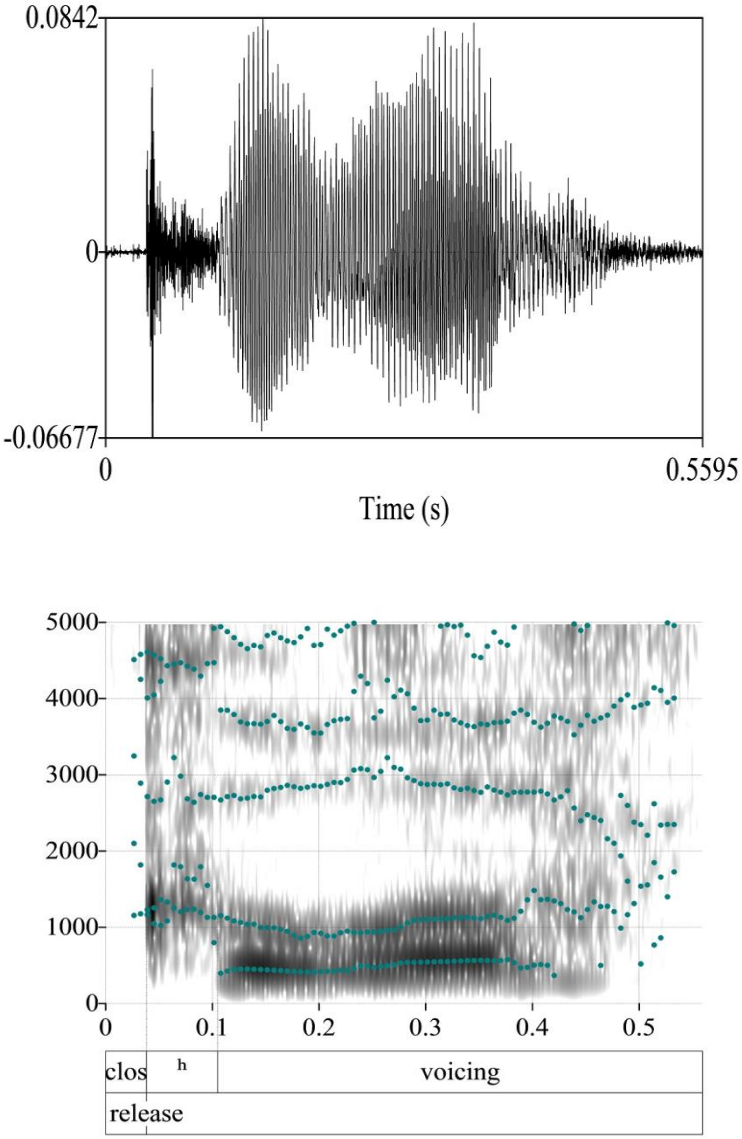


Figure 5: (I think it's a) 'core[3]' as produced by a southern female speaker of English

In figure 5, we can see the aperiodic phase in the onset plosive in 'core' between ca. 0.03 and 0.1 seconds. The aspiration during the aperiodic phase is ca. 70ms in duration, with a relatively strong initial transient at ca. 0.03 seconds (cf. the waveform at the top of figure 5). The individual formant

[3] As can be seen by viewing F1-F2 at ca. 0.2-0.4 seconds on the x-axis, /ɔ:/ has a realisation approaching [oə]. The southern female made variable use of this property in CVVs.

movements during the aperiodic phase in ‘core’ have the following spectral properties:

a) Virtually no trace of F1 can be evidenced in the aperiodic phase of ‘core’, which explains the lack of an estimate given by Praat between ca. 0.03 and 0.1 seconds. This finding is not surprising, given that F2, F3 and F4 are the most typical constituents in aspirated consonants (Stevens, 1998, p. 463).

b) F2, on the other hand, is clearly visible between ca. 0.03 and 0.1 seconds in the aperiodic phase of ‘core’, straddling 1.000-1.200 Hz.

c) F3 in the aperiodic phase of ‘core’ is more variable in its estimated centre frequency compared to that of F2. The third formant fluctuates between ca. 2800 and 3200Hz during the aperiodic phase in ‘core’ with a peak at ca. 0.04-0.05 seconds.

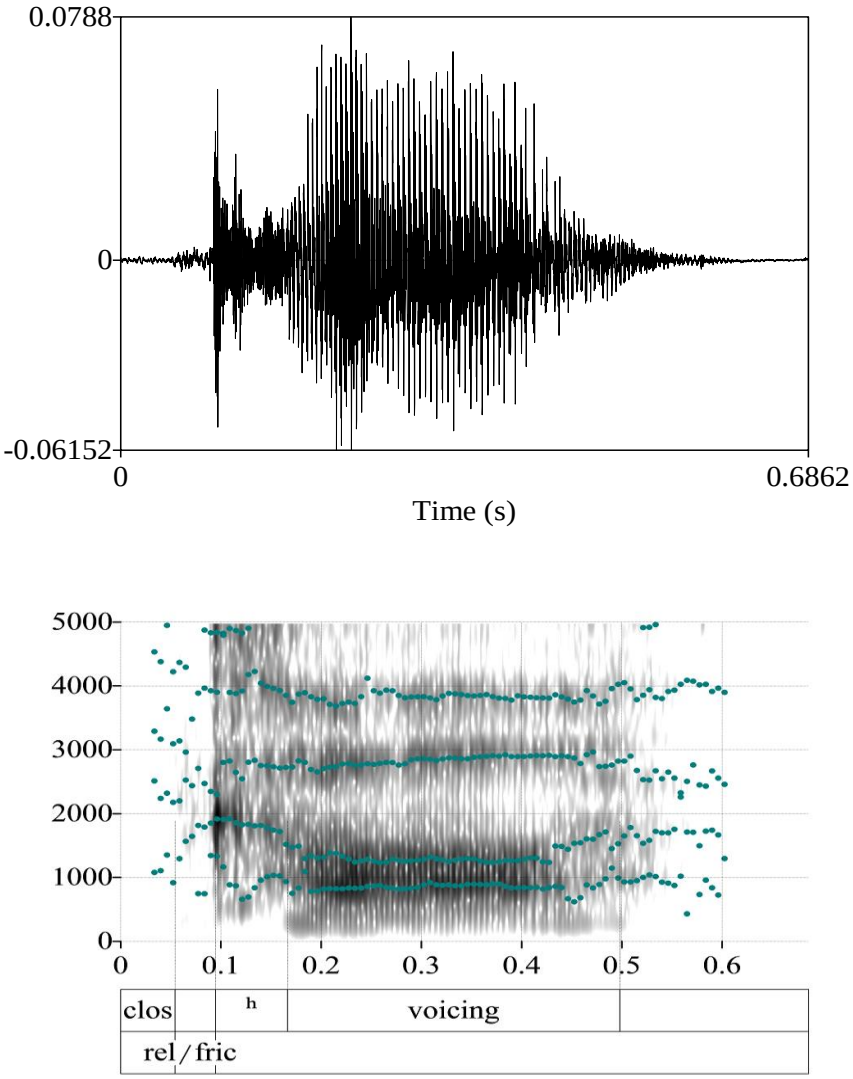


Figure 6: (I think it's a) 'car'[4] as produced by a southern female speaker of English

Figure 6 shows that the aperiodic phase in the onset plosive in 'car' is ca. 110 ms in duration (cf. the x-axis between ca. 0.07 and 0.18 seconds). A ca. 30ms band of frication can be discerned between ca. 0.07-0.1 seconds. The individual formant transitions during the aperiodic phase can be described as follows:

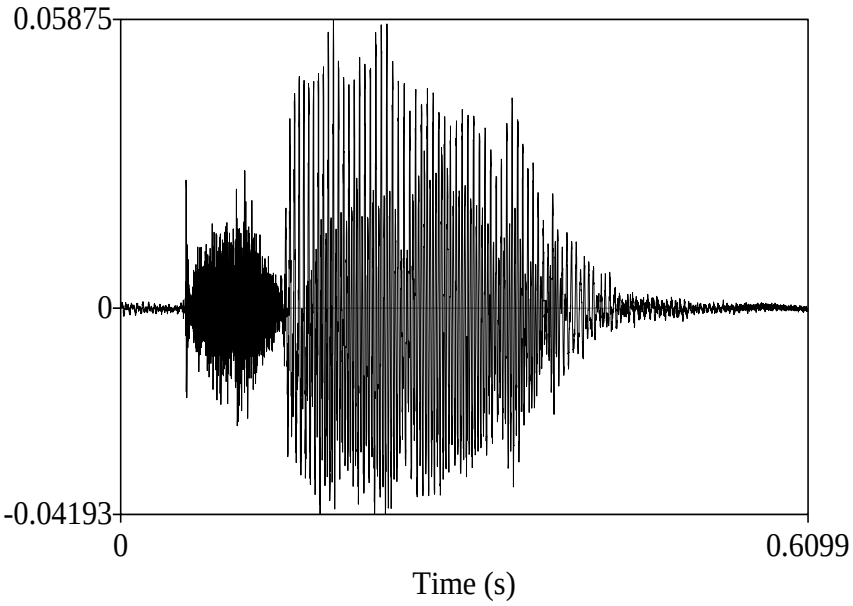
[4] 'clos', 'rel/fric' in figures 5-7 refer to the hold phase closures and releases in each plosive sound, as well as the frication in /ɑ:/, respectively.

a) F1 in the aperiodic phase of ‘car’ fluctuates between ca. 800-1100 Hz (see the bottom green speckled line at ca. 0.07-0.18 seconds)[5]

b) F2 in the aperiodic phase of ‘car’ descends from ca. 1900 to ca. 1400Hz (see the second speckled green line from the bottom between ca. 0.07-0.18 seconds).

c) F3 remains fairly level at around 2900Hz in the aperiodic phase of ‘car’ (see the third speckled green line from the bottom in figure 6 between ca. 0.07-0.18seconds).

[5] F1 is often hard to estimate in aspiration. For /ɑ:/, the values given are more reliable, as they match those produced during the vowel steady state.



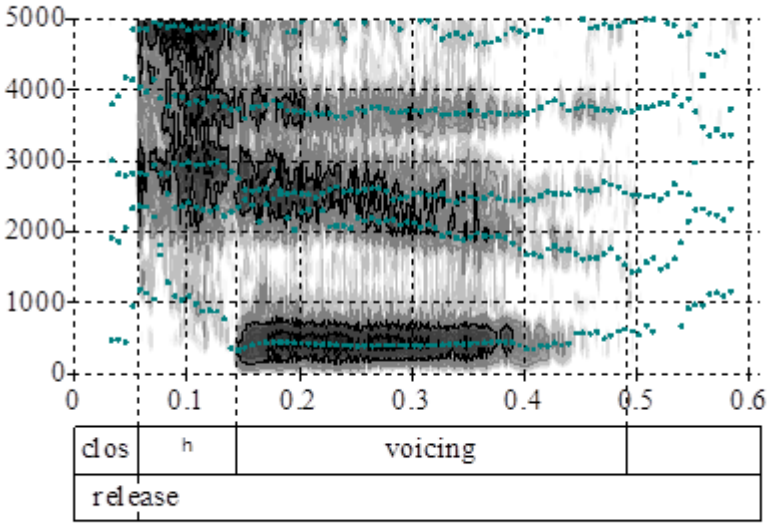


Figure 7: (I think it's a) 'coo' as produced by a southern female speaker of English

Figure 7 shows that the aperiodic phase in the onset plosive in 'coo' is ca. 80 ms in duration (cf. the x-axis between ca. 0.06 and 0.14 seconds). The individual formant movements during the aperiodic phase can be characterised as follows:

a) the estimate of F1 as being on average about 950Hz between 0.06 and 0.14 seconds (cf. the first green line from the bottom near the bottom-left hand corner of figure 7) is almost certainly an error associated with the formant tracking in Praat. According to Stevens (1998: 197 and 463), coupling to the subglottal cavity has its lowest natural resonance at ca. 600Hz, with typical values during aspiration having centre frequencies of around 800 Hz (with higher values for female speakers). Since the bottom green line between ca. 0.06 and 0.14 seconds in 'coo' straddles the area between ca. 600-1100 Hz, coupling to the subglottal cavities may be a good explanation for the high value of F1 in this instance. Were a more typical F1 value being observed, it would almost certainly be lower in frequency (e.g. 350-450 Hz), as for a high back vowel.

In summary for figures 5-7, identifiable differences in the FPD of the three utterances include distinctions in the phonetic exponents of the aperiodic phases in 'core-car-coo' (also see Stevens, 1998, p. 339-375):

i) When examining the left-hand side of the waveforms in figures 5-7, it may be noted that the aperiodic phase in the onset plosive in 'car' is of longer duration than in 'core' and 'coo' (about 110ms vs. 70 and 85ms respectively). The 'car' example displayed in figure 6 contains a ca. 30ms band of frication between ca. 0.07-0.1 seconds.

ii) There are significant differences in the estimated formant transitions into the voiced vocalic portion from the velar plosives in each of the three word forms (cf. left-hand sides of the spectrograms in figures 5-7). In 'car' (cf. figure 6), F2 and F1 extend from the burst at ca. 0.070 seconds to the onset of glottal vibration at ca. 0.18 seconds. The F2 transition during the aperiodic phase in 'car' between ca. 0.070 and 0.18 seconds slopes downwards from ca. 1900Hz to ca. 1400 Hz (in

anticipation of its steady state centre frequency): such a descending F2 is not evidenced in ‘core’ or ‘coo’ at the equivalent time points (cf. the second green line from the bottom in figures 5 and 7). Some differences can also be noted between ‘core’ and ‘coo’. The estimated centre frequency of F2 at 0.05 seconds in ‘core’ has a much sharper upward movement at the equivalent time point than in ‘coo’. F3 in ‘core’ is nearly level at ca. 2900 Hz very early on at ca. 0.04 seconds, whilst in ‘coo’ a change of about 300 Hz can be evidenced (cf. the descent from ca. 2.900 to 2.600 Hz for the third green line from the bottom at ca. 0.15 seconds in figure 7).

In summary, the FPD of each of the three utterances is distinct during the aperiodic phases of the velar plosive sounds. Without having the right types of formant transitions and adequate time to transition from the aperiodic burst to the onset of voiced glottal vibration for the upcoming vowel, the continuity of perception may be distorted. Listeners may otherwise not be able to determine place of articulation of the plosive (Stevens and Blumstein, 1978) and whether the onset is to be recognised as voiced or voiceless (Lisker, 1957). In particular, potential discontinuities or other similar distortions during the aperiodic phase would make it much more difficult to recognise the vowel early on (LaRiviere et al, 1975, p. 475). The perceptibility of each word form and in particular that of the vowel sound described in figures 5-7 will depend heavily on their FPD. The implications of this type of claim concerning FPD shape the definition of coarticulation in this thesis. They will be explored in the next subsection.

1.3.1 A Non-Segmental Structural Definition of Coarticulation: Phonetic vs. Phonological Aspects

The definition given for coarticulation in this thesis is complex. However, it allows for a structurally rich interpretation of

linguistic structure. It sheds light on the complex co-ordination between facets of linguistic structure and FPD contra the implementation of articulatory movements. The definition of coarticulation stems from a review of the existing literature behind

a) previous research on vowel recognition from aspirated plosive CVs, according to which vowels can be reliably recognised from the burst[6] in the aperiodic phase of initial plosives in CVs (see e.g. Winitz et al, 1972 and LaRiviere et al, 1975).

b) general studies on vowel timing in CV(V)/C and CV(V)C-type monosyllables: the most important characteristic in vowel perception is VISC (= vowel-inherent spectral change) in these studies, which can be defined as the momentary spectro-temporal fluctuations in formant centre frequencies through time in vowel sounds (Rosner and Pickering, 1994 and Nearey and Assmann, 1986). Vowel formant trajectories only start to approximate their steady state values no earlier than 30ms post-release, somewhat later than the burst portion.

c) to what extent the results on the significance of similar studies on the perception and production of long-domain coarticulation in English monosyllabic lexemes such as ‘pen’ (see Cohn, 1990) and ‘led’ and ‘let’ (see Hawkins and Nguyen, 2001, 2004) can be generalised to vowel recognition from aspirated plosives.

The findings of all these three strands of studies are in conflict, in that the first are small in scale, few in number and have distinctive methodologies, which have yielded distinctive

[6] The burst portion is considered to last up to a maximum of 20ms (see e.g. Klatt, 1975, p. 690)

results. The first two strands of studies differ as to at what point during the aperiodic phase the vowel can be reliably recognised (cf. e.g. Cullinan and Tekieli, 1979 and Winitz et al, 1972 vs. Nearey and Assmann, 1986). Unlike strand c), the first two strands of the literature lack a detailed account of how polysystemicity and long-domain coarticulatory detail might affect vowel recognition timing in different contexts/structures. For these reasons, a synthesis of the results of all three strands in the literature is required, in order to give a theoretically adequate account and definition of coarticulation.

Coarticulation can be defined as the systematic phonetic influence exerted by the productional and perceptual mapping on constituents of various sizes in the phonological tree at whatever level of representation. Such forms of influences have phonological implications requiring a phonological account. For instance, the two examples from Hawkins and Smith (2001) and those for 'core-coo-car' provided in figures 1-3 and 5-7 (respectively) demonstrate that coarticulation spans multiple levels of structure including morphological, grammatical and syllabic constituents, as well as phonetic vs. phonological properties of different utterances: even structurally quite simple stimuli such as 'core-coo-car' show that multiple phonetic differences can be found in monosyllabic utterances. This kind of detail is particularly relevant to the aperiodic releases of voiceless plosives, which display a high degree of coarticulation with upcoming sounds. The definition of coarticulation given in this subsection encourages giving immediate attention to bottom-up and lower-level articulatory-acoustic properties of coarticulation (see e.g. Catford, 2001), as well as its top-down and higher-level perceptual consequences (see e.g. Hardcastle & Hewlett, 1999). The non-segmental definition of coarticulation can be seen as more polysystemic than that of previous models: more than in previous models, the non-segmental model developed in this research recognises that

a wide range of structures and phonetic properties have significant implications for phonological processing of monosyllabic lexemes.

The definition given for coarticulation in this subsection can be seen as unusual in that it does not lend the bulk of phonetic/phonological influence for the perception of a given sound to its immediately adjacent sounds. Most similar previous research on coarticulation focuses on research for phoneme-sized segments, CV syllables and/or single words, which have been excised entirely out of context without using a carrier phrase. Rather, significant attention is paid to a much wider constellation of properties relating to phonological structure and the phonetic properties of coarticulation. Such factors might include rhythm, intonation and potentially voice quality. To borrow Local's (2003, p. 323) claim: "No order of detail can be dismissed, a priori, as disorderly, accidental or irrelevant".

If the kinds of results on long-domain coarticulation and non-segmental properties by e.g. Goffman et al (2008) and Hawkins and Nguyen (2001) are to be taken seriously, there is a need to entertain a very rich and complex understanding of coarticulation. For example, the claim by Goffman et al (2008) on the articulatory planning that speakers perform prior to the implementation of a sentence cannot be explained on phonetic criteria alone. If motor commands to the muscles are altered for the production of an entire sentence depending on just one sound, such a finding offers good support to the claim that coarticulation is not just a feature of phonetic interpretation and phonetic exponency. Rather, coarticulation reflects features of phonological structure. This claim about the nature of coarticulation applies even more strongly in cases where the changes in the motor commands reflect specific combinations of FPD. For example, studies on the production and perception of English liquids (see e.g. Kelly and Local, 1986 and West,

1999b) have shown that changing the quality of a liquid sound in English sentences has differing effects on the resulting FPD of surrounding vowels in CVCs. This can affect perception of both the liquid sounds and the surrounding vowels so that the distinctions can be recognised several sounds (or even a few syllables) prior to the phonetic implementation of the liquid. Since such words forms with liquid onsets have similar phonetic properties as a whole to the words investigated in this work, and almost the same structural features and properties, this issue remains theoretically important in this thesis.

Coarticulation represents both phonetic and phonological aspects of speech. This claim is justified by the fact that phonetic properties such as voicing and aspiration can take language-specific and even accent-specific shapes and properties. For example, the realisations of French, Thai and English voiceless plosives differ from each other: French only has unaspirated plosives, where English and Thai mainly have voiceless aspirated plosives, which may take specific phonetic shapes. For example, in specific listening situations and across accents the degree and quality of the aspiration may differ: for example, in many varieties spoken in Lancashire and Scotland, plosives have little or no aspiration, which is not true for other varieties (e.g. Wells, 1982, p. 370 and 409). Thai, on the other hand, has both aspirated and unaspirated voiceless plosives and also has fully voiced ones. As an important aside, this discussion does *not* relate to voice onset time (VOT); rather it emphasises the phonological consequences of small but significant phonetic distinctions in plosive-vowel combinations and the complex coarticulatory mapping that monosyllabic utterances require. To summarise, coarticulation *does* represent language-specific (= phonological) properties in the transmission of speech to both listener and speaker. Having covered the definition of coarticulation in this section, now is a good time to consider the influences on this study.

1.4 Influences on this Study and its Theoretical Rationale: Theories and Approaches to Coarticulation

This subsection presents a detailed account of the three linguistic theories that have influenced this study:

i) Firthian prosodic analysis (FPA)
ii) Declarative phonology (DP)
iii) Polysp

Having a descriptive and theoretical account of all three theories allows us to pinpoint their main strengths and weaknesses, in particular with respect to the definition of coarticulation given in the previous subsection.

1.4.1 FPA and DP

According to Plug (2005, p. 22-27), a new line of more radical of thinking about phonology started to emerge at London UCL and SOAS in the early and mid-20th century. The main proponents of this new line of thinking were John Rupert Firth and his colleague Stephen Jones, who discouraged the use of phonemes in structural analysis and supported polysystemicity. Even though Firth saw the potential value of phonemic analysis in broad transcription (Firth, 1934c, p. 2), he understood that phonemes and alphabetic writing share a close connection. Firth highlighted some of the problems that phonemic analysis encounters, and viewed phonology more in non-segmental prosodic terms than from a phonemic viewpoint. According to Chapman and Routledge (2005, pp. 81-82), "Firth considered it perfectly proper to focus on only one very small subsystem of a language, ignoring other subsystems if it made descriptive sense to do so, a principle referred to as 'polysystemicity'. In

polysystemic linguistic analysis, the interaction between different linguistic systems contributes to the formation of the phonetic exponency of different sounds and structures (see e.g. Hawkins and Smith, 2001, p. 112). Firth saw the close connections shared by FPD and phonetic exponency in relation to phonology and representation (Anderson, 1985, pp. 184-185).

Another key concept applied by Firth is that of 'prosodies', which can be equated with idea of contrasts spreading over more than one sound: in opposition to phonemes, prosodies "extend over more than one sound (or segment)" (Sprigg, 2005, p. 125). The phonetic exponents of contrasts are not limited to phoneme-sized segments or words/phrases, according to Firth. For example, the term 'prosody' can apply to junctures, where features are linked syntagmatically so that the structures at the end/beginning of contiguous structures share some features (Sprigg, 2005, p. 125).

In sum, Firth saw one of the main deficits of phonemic analysis, in that phonemes are devoid of context. Firth also recognised the lack of its emphasis of language as an 'enclosed system', which does not fully recognise the interactions between different linguistic systems (such as semantics, phonology and grammar). This view and the term 'prosody' are closely related to the broader scope of language as a polysystemic system in Firth's work. The following four paragraphs discuss the key aspects of DP, which is the other main prosodic theory of phonology that has influenced this research.

First, DP can be considered a child of FPA and was developed by John Coleman and Steven Bird and colleagues in Oxford and Edinburgh in the late 1980s and throughout the 1990s. It is apt to describe DP as a more constrained and systematic version of FPA. DP focuses more specifically on

phonology and representation than FPA, as it is a more phonological theory than one of representation and meaning (like FPA). As far as the theoretical analysis and accounts of coarticulation are concerned, recent research in DP does not go quite far enough in accounting for how strongly feature spreading as a contrastive element in phonological/syllable structure can affect phonetic exponency in mono- and disyllabic utterances. This claim is based on the results of Hawkins and Smith (2001), Hawkins & Slater (1994) and similar pieces of research: small phonetic distinctions in syllable-initial laterals and fricatives can affect the timing and co-ordination of surrounding vowels and codas, for instance. Such results could have implications for phonological representation. For example, although it is suggested in Coleman (1998, p. 224) that rime exponents spread over the whole duration of CVCs, it is not explained why vowel length is represented at the nucleus level.

Second, the strength of DP lies in how it highlights the weaknesses and overly powerful procedural rewriting rules of conventional phonological theories, such as generative phonology (e.g. Chomsky and Halle, 1968) and autosegmental phonology (Goldsmith, 1976). The rules in generative phonology do not adequately constrain the forms that phonetic and phonological representations can take. Since any phoneme can be deleted or inserted anywhere in a structure (as in the [t] segment in ‘next door’), this kind of a claim results in an unconstrained grammar. The main argument concerning insertion here is a mathematical one. The generative analysis does not allow a listener to have a sufficient understanding of what kinds of parsing strategies to use. This claim can be substantiated by the fact that i) such a theory would not allow a listener to comprehend where one thing ends and another begins (e.g. 1 + 1 phonemes may equal either more and/or less

than one sound). In sum, the point that DP makes is that it is not possible to reliably parse a grammar that includes the deletion/insertion of sounds.

Last, DP recognises the need for explicit temporal and parametric interpretation (parametric interpretation refers to the realisations of phonetic exponents being sensitive to structural properties, cf. Coleman, 1998). For instance, if labiality can be observed throughout a syllable, it must be determined where the phonological representation for labiality is located (e.g. at the syllable level or at a lower node). From the syllabic level, the syllable length of the spreading follows 'for free', whereas at lower nodes the extended duration of this feature must be specified in phonetic interpretation by temporal constraints.

In summary, DP is an extension of FPA that contributes to phonological analysis and relates it more optimally to phonological concepts and phenomena, the most important of which are domain of contrast and phonological representation

1.4.2 Polysp

Polysp, as its name suggests is a polysystemic theory, looking at more subtle aspects of speech perception and speech understanding (see e.g. Hawkins and Smith, 2001, Hawkins, 2003 and Smith et al, 2012):

> "…Polysp (for POLYsystemic SPeech understanding) that combines a richly-structured, polysystemic linguistic model derived from Firthian prosodic analysis and declarative phonology, with psychological and neuropsychological approaches to the organisation of sensory experience into knowledge. We propose that the type of approach exemplified by Polysp promises a fruitful way of conceptualising how meaning is understood from spoken utterances, partly by ascribing

> an important role to all kinds of systematic fine phonetic detail available in the physical speech signal and by rejecting assumptions that the physical signal is analysed as early as possible into abstract linguistic units".
>
> Hawkins and Smith (2001, p. 99)

According to Hawkins & Smith (2001, p. 99) Polysp makes a more fruitful effort in detailing the kinds of processes involved in linguistic abstraction. Episodic multi-modal sensory experiences are seen to be at the heart of the perception/production process, so that the emphasis is on interaction and understanding of meaning rather than constructing a thorough structural understanding of any given utterance at different successive and compulsory levels of formal linguistic analysis. Speaker-listeners rarely build up complete and formal analyses of different utterances, especially in online speech production and perception. It is important to take a step away from the structural-linguistic properties of utterances if we are to give an adequate and detailed account of perception and production.

There is one other difference between FPA and Polysp that needs to be addressed, and which relates to the relationship between FPD and larger structures:

> "In FPA, a difference in FPD is reflected in different prosodic/grammatical structures: when the linguistic structures that describe two utterances differ, then their sounds differ. Polysp retains the polysystemicity but reverses the logic, so that in perception, a reasonable hypothesis is that if the sounds in two utterances differ, then one or more things in their structures differ. Thus in Polysp, small parts of the sensory signal (such as

> acoustically distinct segments) can only be processed in terms of their wider context".
>
> Hawkins (2010a, p. 482)

Where FPA sees larger structures (e.g. phrases or larger parts of a sentence or utterance) as being realised with specific phonetic exponents at particular and structurally smaller points, Polysp can be seen to emphasise the importance of FPD even more strongly. Its stance comprises an affirmation of how speaker-listeners construct meaning through subtle phonetic changes at specific points in structure without necessary recourse to the properties of larger constituents. The 'mistakes/mistimes' example given at the beginning of this chapter comprises an apt illustration of this stance. The fact that Polysp recognises the significance of how larger structures unfold from what may be very small chunks of FPD is a particularly important claim in this thesis.

Polysp comes closer than FPA in characterising speech understanding. For example, Hawkins (2003, p. 373) shows that Polysp emphasises changing the focus of enquiry in linguistic analysis. Polysp takes the view that a detailed analysis of more global aspects of speaker-listeners' communicative situation (of which speech is only part) forms a key approach in theoretical linguistic research:

> "…one may interpret the meaning of an utterance directly from the global sound pattern; reference to formal linguistic units of analysis, such as phonemes, words, and grammar, is incidental; circumstances dictate whether such reference takes place at all, and if it takes place, whether it does so after the meaning has been understood, before it has been understood, or simultaneously with the construction of meaning. The implications of this position are that speech perception

> does not demand early reference to abstract linguistic units, but instead, to flexible, dynamic organisation of multi-modal (and modality-specific) memories; and that models of speech perception should reflect the multi-purpose function of phonetic information, and the polysystemic nature of speech within language".
>
> Hawkins (2003, p. 373)

Polysp views phonetics and phonology from a much wider perspective than conventional theories of speech perception/understanding, which often emphasise the study of lab-speech and isolation forms. The importance of this is issue is that although Polysp could potentially be seen as a child of FPA, it takes a wider view of linguistic analysis and representation, incorporating neural and other physiological detail into its analyses. Such claims also apply to this thesis in terms of, for example, physiological and phonetic-articulatory detail: for example, details are given in chapter three on the movement velocities of the articulators involved in bilabial contra alveolar and velar plosives. In contrast, physiology and sensory processing are areas which FPA largely detaches itself from (Plug, 2005, p. 40-41).

Having characterised the main differences and similarities between three related non-segmental linguistic theories, it is time to show what aspects of coarticulation and phonological processing FPA/DP and Polysp have not modelled in sufficient detail. In part, what is discussed in the next subsection reflects the progress and the partial lack of development in the three theories over time. However, the following subsection does highlight some of the limits of phonological and coarticulatory phenomena that even the most modern of the three influencing theories on this research (i.e. Polysp) has not closely addressed. The next subsection therefore affirms what the results of this research will look like

if these three non-segmental theories are in the right with respect to the perceptual and productional significance of polysystemicity and FPD.

1.4.3 The Contributions of Previous Theories: Explaining Vowel Timing Non-Segmentally

The purpose of the non-segmental approach in this research is twofold: (1) to strengthen the claims made in Polysp as a more phonetic theory on the one hand and DP and FPA as more phonological ones on the other and (2) to show that there are phenomena relating to the perception of coarticulation which can be modelled by these theories which have not yet been considered. This work extends FPA, DP and Polysp.

Since it is known that there are sometimes substantial differences in the secondary articulations[7] and phonological-metrical structure of different varieties of English (see e.g. Wells, 1982 and Pierrehumbert, 1980), this conclusion suggests that caution should be exercised in claiming that coarticulatory strategies between different forms of English do not differ significantly in phonetic and perceptual terms. For example, it seems quite likely that the kinds of distinctions between laterals in northern and southern English varieties (mostly dark laterals in the north of England and more clear palatal ones in the south, see e.g. West, 1999b) are not limited to resonant sounds. Rather, they might indicate the importance of resonance as a wider phenomenon in different sounds and structures.

Secondary articulations and phonological-metrical structure are considered as the key features and structures across which coarticulation is implemented in speech. This thesis sheds new light on the analysis and representation of

[7] Secondary articulations can be defined as secondary strictures in speech sounds, which affect a given part of their exponents. Examples in English include palatalisation and velarisation of liquids and affrication of plosive onsets (Wells, 1982).

coarticulation in speech. The rest of this subsection discusses certain weaknesses and gaps relating to coarticulatory and phonological theory in FPA, DP and Polysp.

First, FPA is very specific as a non-segmental theory and makes useful claims and predictions about the structure, representation and realisation of speech sounds in context. However, FPA is in need of some theoretical enhancements, especially as far as modelling the relationship between phonetic exponency and domain of contrast is concerned. This aspect of exponency is most notable for the potential interaction between carryover and anticipatory coarticulation in the context of CV(V)/Cs having complex phonetic properties, such as nasal rimes and lateral onsets (see e.g. Hawkins and Stevens, 1985 and Hawkins and Nguyen, 2001, 2004). The influence of how duration is encoded in short contra long vowels is not addressed in detail in FPA, especially in terms of how it might affect the coarticulatory properties of monosyllabic utterances. For example, FPA has very little to say about the implications for differences in phonetic exponency in CVVs contra CVCs (even for oriental languages): the fact that stressed CVs do not occur monosyllabically in English is only briefly addressed in existing FPA work on English, such as commentaries on Eileen Whitley's study of English Phonology (Simpson, 2005).

Second, DP can be seen to come farthest of the three theories in the modelling of coarticulation. There is a much stronger focus on phonological factors and on contrast in general in DP than on phonetic-temporal properties. DP must be enhanced in two respects. The first concerns the bidirectionality of coarticulation contra the relationship between phonetic exponency and contrast. The second relates to how such bidirectionality may have differing consequences for the planning and realisation of FPD depending on i) the mutual relational properties of English sounds and ii) how that

affects the demarcation of domain of contrast for various phonological features, such as length and nasality.

Last, Polysp makes some useful claims that go further with respect to phonetic exponency than either FPA or DP. Hawkins and colleagues offer much in the way of the perceptual importance of FPD, which is of considerable aid in modelling the mutual dependency between FPD and larger structures. Polysp also makes several useful predictions on higher-level processing and hemispheric lateralisation (see e.g. Hawkins, 2003, 2010a). However, all this discussion is done at the cost of modelling phonological processing, representation and domain of contrast adequately. In fact, Polysp has very little to say about domain of contrast or about the necessity and requirements set by anticipatory and carryover coarticulation, even for more complex sounds such as lateral onsets and voiced fricatives (e.g. Heid and Hawkins, 2000, Hawkins & Slater, 1994 and Hawkins and Nguyen, 2001, 2004).

Drawing from the improvements and enhancements that FPA/DP and Polysp may require, one of the main aims of this research is to show whether these three theories of phonetics/phonology, coarticulation and representation can be brought together by investigating vowel recognition from onsets realised with aspiration in CV(V)C monosyllables. The question whether FPA, DP and Polysp can be brought closer together by making specific predictions about the phonetic and phonological modelling of English CVCs is also investigated in this thesis. These predictions can in large part be derived from the kinds of long-domain coarticulatory effects exemplified in subsections 2.2 and 2.3 on vowel timing and its phonological treatment. The previous studies on vowel recognition for North American varieties specifically have very little to say about long-domain coarticulation in vowel recognition (see e.g. Tekieli and Cullinan, 1979, Cullinan and Tekieli, 1979 and LaRiviere et al, 1975).

If the production and perception results described in chapter 4 give rise to additional similar forms of variation and/or effects as the previous non-segmental studies exemplified in 1.4.1-1.4.2, such works receive additional support. How the results of this study may affect the way perception evolves through time in the way onset plosives coarticulate with upcoming vowels is deemed most important. For example, if phonetically complex realisations of rimes or onsets require listeners to hear comparatively longer portions of the vowel than for structures having articulatorily simpler exponents (i.e. in order to recognise vowel quality), non-segmental findings and frameworks will be supported. The same applies to the encoding of vowel duration. If it is shown that structures with complex phonetic exponents have significant effects on the temporal evolution of vocalic information in e.g. CVCs contra CVVs, but do not delay or distort perception temporally, neither segmental nor non-segmental theories receive strong support. If no such effects can be shown to exist, phonemic models are supported.

Having defined coarticulation in detail in this subsection, it is time to consider some of the key terminology in this research and to what extent the application of the main terms differs from that in previous studies.

1.5 The Application and Use of Terminology in this Study

In order to fully understand and appreciate the use of key terms and the claims made in this thesis, there is a need to discuss, define and illustrate three things:

i) how the type of polysystemic FPD exemplified for ‘mistimes’ and ‘mistakes’ and ‘who(’s) sharpened the meat cleaver’ in 1.1 relates to coarticulation and its bidirectionality in CV(V)/Cs: the bidirectionality

of coarticulation can be defined as the two-way phonetic influence between carryover (left-to-right) and anticipatory (right-to-left) coarticulation on CV(V)/C monosyllables.

ii) how the terminology in this research ties together the bidirectionality of coarticulation and the temporal dynamics of aspiration/coarticulatory timing with phonological terms and phonological processing in an innovative fashion.

iii) to what extent the relationship between phonetic and phonological aspects of the temporal dynamics of speech perception exemplifies coarticulation as a more complex phenomenon than conventional theories suggest. For example, the mapping of features to sounds (see Goffman et al, 2008 and Hawkins and Nguyen, 2001) has been shown to be more complex than many conventional studies suggest, in terms of perception and production as well as acoustics and phonological processing.

Now is a good time to briefly highlight and illustrate the main phonetic and coarticulatory terminology used in this study, and how the polysystemic non-segmental definition of coarticulation given in 1.3.1 supports the claims made on timing, temporal dynamics and structural differences in subsequent chapters.

First, the use of the term ‘bidirectionality of coarticulation’ is meant to facilitate the understanding of the definition of coarticulation given in 1.3.1. Of particular importance here is the use of the terms ‘anticipatory’ and ‘carryover’ coarticulation as reflecting the desire to understand coarticulation from a wide perspective, which involves complex mapping between several levels and elements of structure. For example, there may be a complex interaction

between different structural levels or nodes in a syllable. An example of such interaction might be the spreading of one element or feature affecting daughter structures and/or properties (e.g. terminal nodes) which also influences the phonetic exponency of its sisters (e.g. if it is the case that rime nasality significantly affects properties of the onset).

Second, a discussion and two illustrations of the spectro-temporal variation related to VISC are given. This type of variability has implications for the definition of coarticulation given in 1.3 and for gaining a full understanding of the application of phonetic and phonological terminology in this thesis. VISC cannot be seen as a coarticulatory or phonological effect (cf. Rosner and Pickering, 1994: 291). Since it is inherent to vowel production/perception (even in isolated vowel productions), it is considered as a feature of phonetic exponency in this thesis. Speakers cannot avoid producing VISC in vowels because the articulators are in constant motion.

The two examples used in this subsection to illustrate properties having to do with VISC are ‘cap’ and ‘cat’ as produced by a northern male speaker from Lancashire:

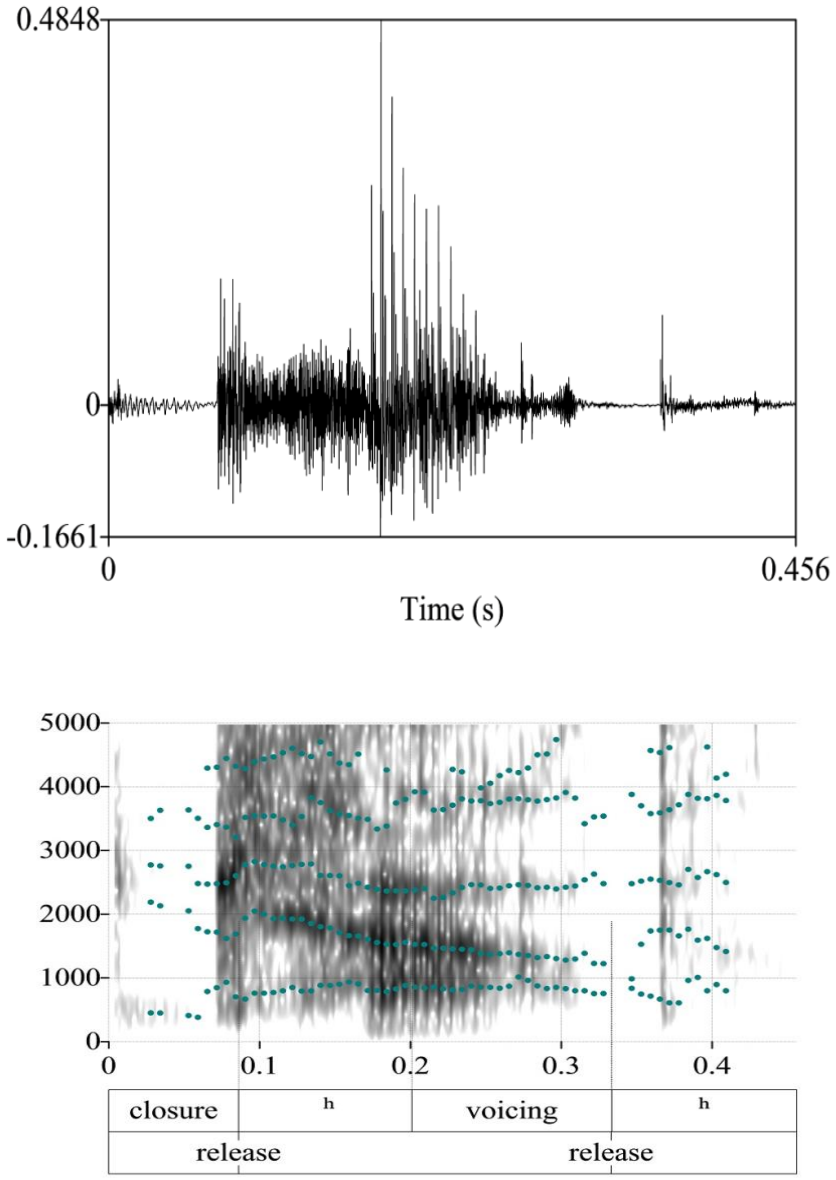


Figure 8: Spectro-temporal moment-to-moment variation in formant centre frequencies in ‘cap’

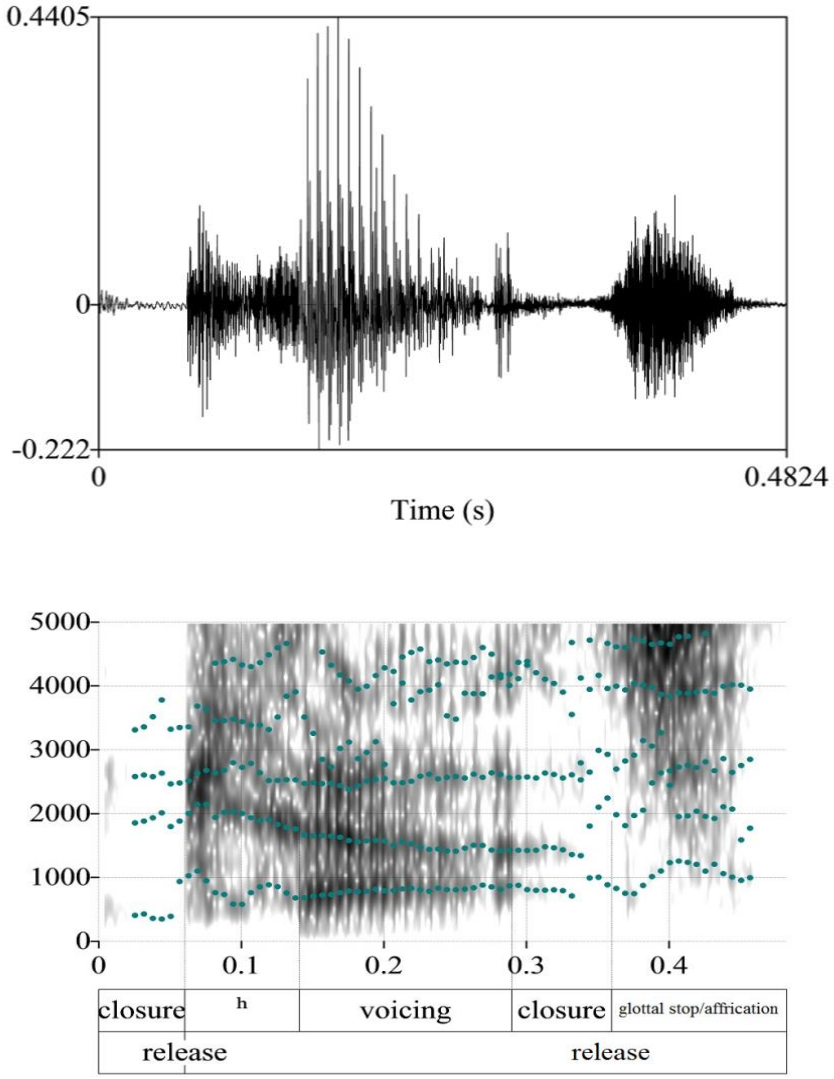


Figure 9: Spectro-temporal moment-to-moment variation in the formant centre frequencies in ‘cat’.

The spectrograms in figures 8-9 represent speech by a northern male. In figures 8-9 can be seen the type of variation in formant centre frequencies hinted at by e.g. Ogden (1992, p. 91), according to which small acoustic distinctions can be achieved by speakers in the production of words like ‘cat’ and ‘cap’.

Before discussing the differences between ‘cap’ and ‘cat’, it is necessary to briefly refer to the differences in F1 in comparison with the spectrograms of ‘core-car-coo’ in figures

5-7. The traces for F1 are much clearer in figures 8-9 than for F1 in ‘core-car-coo’ in figures 5-7. This difference is related to the availability of F1 information in formant tracking in Praat:

“spectral tilt is an especially significant parameter for differentiating male and female speech. These findings are consistent with fiberscopic studies which have shown that males tend to have a more complete glottal closure, leading to less energy loss at the glottis and less spectral tilt”.

Hanson and Chuang, 1999: 1064

The important point about this established difference between male and female speech is that it may help to explain the lack of a clearly discernible F1 in the aperiodic phases of ‘core and coo’ (see figures 5 and 7), since the larger open quotient in female CV(V)/C productions may lead to more critical damping of F1. The more clearly discernible F1 in ‘cap-cat’ (figures 8-9) do not have such spectral properties for F1. Having given a brief account of this difference on spectral tilt, we will compare F1 in the male productions of ‘cat’ and ‘cap’.

The starting point for F1 subsequent to the onset of fold vibration in [a] is ca. 100Hz higher in ‘cap’ than in ‘cat’ and there are differences in the F2 transitions at ca. 0.08 seconds as well as a more descending F3 in ‘cap’ than in ‘cat’ at ca. 0.16 seconds. The differences in the transitions for F1 may reflect the type of complex coarticulatory mapping and influence as discussed earlier in this chapter (for e.g. Goffman et al, 2008). For example, since the articulation of the coda consonant in ‘cap’ lacks a front cavity and has no intrinsic tongue posture requiring an airtight closure in the alveolar/dental region (as in the coda in ‘cat’), F1 in the preceding vowel and aspiration in ‘cat’ produced by a northern male may be excited more relative to higher formants than in ‘cap’. This example shows that the

claims on subtle distinctions in the FPD of plosive-V-plosive monosyllables made by Ogden (1992, p. 91) enjoy good validity. The FPD in this small but theoretically significant example helps to show that in order to account for the temporal dynamics of consonants and vowels and the bidirectionality of coarticulation, a recognition and account of the mapping of FPD at *all levels* of representation is needed. While some of the examples given (e.g. 'core-coo-car' in figures 5-7) do not exemplify polysystemicity to the same extent as some of the examples described in Polysp (see e.g. Hawkins, 2003, and Smith et al, 2012), they do highlight the perceptual role of the same type of FPD. The remainder of this subsection will briefly highlight and define how the application of some of the phonological terms differs from previous research (e.g. Coleman, 1998 and Hawkins and Nguyen, 2001).

The structural phonological terms 'mother/parent node', 'daughter' and 'sister' which are applied non-segmentally in in this thesis, refer to the hierarchical relationships between different nodes in the syllabic representations of CV(V)/Cs and whether a given node dominates another (= mother/parent node), stands in a subordinate relationship to it (= daughter) or is located at the same level of representation (= sister). The significance of this theoretical phonological terminology is that it is applied to an innovative and explicit structural definition of coarticulation, which displays sensitivity to the mapping of FPD onto richly defined structures, and which requires a rich and complex theory of phonetic interpretation This issue is important for the perception of coarticulation, since speech perception and speech production are linked at certain levels (Moore, 2008).

In sum, though the terms used in this research exist in the previous literature, they have not been applied in a similar fashion before. The new definition for coarticulation which underlies the terms discussed in this subsection can be seen as a

theoretical *compilation* of a wide set of terms that fulfil a specific purpose in the study of speech perception timing.

At this point, the background behind this research has been fully reviewed. The next chapter describes the literature behind vowel recognition timing and the secondary research topics that this study is based on.

2. Literature Review

In the first chapter, the groundwork for this research on vowel recognition was laid by arguing that there is a lack of an available and comprehensive model of coarticulation and its timing aspects. That is, in chapter 1, we assessed and described the literature on vowel recognition timing. This chapter examines the existing literature behind vowel recognition timing in English from three perspectives:

i) vowel timing (for both production and perception)

ii) segmental-phonemic studies on the recognition of vowels from plosive-V monosyllables and other related literature relevant to vowel recognition in CV(V)/Cs as well as disyllabic utterances (such as 'berry' and 'belly') including the production of long-domain coarticulation and vowel nasalisation. The related literature investigated a wider range of syllable structures with similar phonetic exponents (e.g. 'led-let', Hawkins and Nguyen, 2001, 2004 and 'pen', Cohn, 1990). Their findings can be seen to relate closely to those on vowel recognition timing in North American varieties of English, albeit that the approach taken in studies on real words is distinctive methodologically.

iii) non-segmental phonological modelling of vowel recognition

The latter parts of the chapter (2.4-2.5) evaluate the methods and findings of previous similar studies on vowel recognition from plosive-V CVs and offer a set of secondary research

questions that emerge from the background literature. Subsection 2.6 offers a set of hypotheses relating to both the primary and secondary research questions. In the next section, properties of vowel timing and the FPD relevant to the timing of vowel recognition are considered in more detail.

2.1 Vowel Timing

The first subsection on the temporal dynamics of vowel sounds summarises the literature behind VISC. The final three subsections detail some of the main literature behind more general aspects of vowel timing, including articulatory-phonetic timing, the way durational information is encoded in vowel sounds and how order effects may affect vowel perception depending on the sequence in which distinctive vowels are heard. We begin by reviewing the literature on VISC.

2.1.1 Timing Information: VISC

According to Rosner and Pickering (1994, p. 283), the articulatory-perceptual facts associated with coarticulation are in many ways opposed to the widely held theoretical notion of vowel targets. When approaching the issue from the speech production viewpoint, vowel targets are normally associated with specific and ideal articulatory targets. Before describing the theoretical problems associated with this issue, some references to constraints on vowel articulation and their timing are made.

Rosner and Pickering (1994, p. 281) show that three principles govern vowel articulation, of which the first two are physiological. These principles are known as the synergy and rate constraints, out of which the synergy constraints have an

impact upon the relations between different articulators by limiting how static they can be in their spatial movements. Such constraints impose limits on the degree of how static different articulators can be at particular moments in time. Rate constraints limit the velocity of articulators in the sense of how rapidly they can move from one configuration to the next.

These two constraints are important factors in accounting for the articulatory-perceptual requirements set by vowel articulations in vowel timing: speech sounds should not be viewed as ideally definable static articulatory targets. Rather, they should be seen as dynamically variable targets.

This fact is particularly relevant within the context of the human speech apparatus, which is constantly adopting different states and positions depending on the requirements of the particular settings in which speech unfolds in time. This property can be seen as a direct consequence of synergy constraints. The rest of this subsection considers this issue, i.e. how vowel perception and production evolve through time with respect to the individual formants that vowel sounds comprise.

Rosner and Pickering (1994, p. 291) argue that VISC does not constitute a coarticulatory effect. VISC *is inherent* to voiced periodic open approximation, and speakers cannot avoid producing it in vowels because the articulators are in constant motion. Isolated vowels have paths in auditory vowel space (i.e. the range of audible differences for vowel sounds regarding the firing of auditory nerves on the basilar membrane in the inner ear). These paths can be seen as direct derivatives of the changes that typify VISC.

The moment-to-moment variation in VISC in isolated vowels leads to the same theoretical issues in tackling the problem of recognition as in consonantal contexts. For example, is vowel categorisation dependent on values around the steady state, or do listeners perceive some kind of momentary or transitional averages from vowel realisations?

Tackling such questions requires a closer examination of the productional detail associated with VISC.

Figure 10 illustrates the kind of acoustic-temporal variation associated with the ways in which VISC alternates in two English syllables having lexical meaning (the word ‘bee’ and the letter {d}):

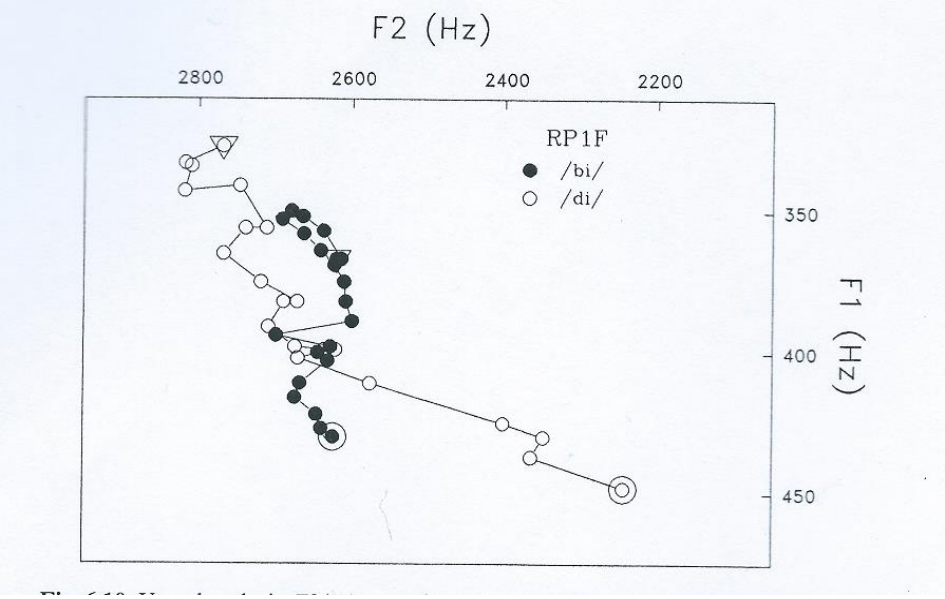


Figure 10: Vowel paths in *F*2/*F*1 space for /i/ in two different /CV/ contexts
(Rosner and Pickering, 1994, pp. 280, fig. 6.10)

Figure 10 shows the auditory vowel paths (AVP) for F1 and F2 associated with [i:] in two CV contexts. F1 can be found on the y-axis, whilst F2 is displayed on the x-axis (top).

The AVP shown in figure 10 corresponds directly to the associated acoustic variation in VISC in [bi:] (cf. black dotted circles) and [di:] (cf. white circles, also see Rosner and Pickering, 1994, p. 291). It can also be seen in figure 10 that there is more extensive moment-to-moment variation in [di:] compared to [bi:]. This difference is in response to the fact that alveolar consonants have more complex articulation than bilabials, which have no intrinsic tongue posture. The

movement velocities for the individual formants are very similar (cf. the spacing between individual formant frequency points in figure 10), except very shortly after the burst transients (cf. the white and black circles).

The variability shown in figure 10 demonstrates the importance of considering VISC *in context*, regardless of the fact that isolated vowels display similar moment-to-moment variation. The AVP for [di:] only starts to become less variable through time around 70 ms. The spectro-temporal variations in the formant centre frequency values for VISC are different in [di:] and [bi:]. This difference depends on the distinctions in the realisations of the plosive onsets, where one is bilabial and the other dental/alveolar.

In acoustic terms, vowel formant centre frequencies change continuously through time, as figure 10 attests. Consequently, the auditory patterning and auditory responses to vowels almost certainly must change with time. Vowels cannot have constant values at every level of representation. Each vowel formant indicator must vary temporally over the course of its phonetic realisation in order for it to be possible for listeners to arrive at a weighted average processing/prototype target. At least at a certain level, scalar representations must be allowed. The final paragraphs of this subsection describe this issue.

Rosner and Pickering (1994, p. 278) affirm that the representation of vowels in auditory space necessitates distributing characteristic production values throughout a vowel's entire duration. Information about such acoustic detail would directly reflect listeners' capability to derive vowel prototypes and/or targets from vowels' phonetic exponents. When assessing vowel production and perception, such conclusions necessitate considering the dynamic properties of vowel articulations, which are highly dependent on temporal

properties. Rosner and Pickering (1994, p. 290) confirm that for listeners to be able to maintain the reference values for different auditory prototypes, a function needs to be introduced which has a particular domain reflecting specific vowel productions. Rosner and Pickering (1994) characterise this function as the 'auditory space path' (or 'ASP' function), which enables speaker-listeners to integrate over the values corresponding to particular vowel paths in auditory vowel space. Vowel prototype value generation depends directly on the values associated with the ASP.

Nearey and Assmann (1986, p. 1299) show that the ca. 30ms regions around the transitions into and out from a vowel comprise the most important spectro-temporal cues to recognition. This is a claim which is particularly important for this research. Rosner and Pickering (1994, p. 298) show that establishing the direction of change constitutes the best possible perceptual exploitation of the acoustic detail associated with VISC. One contrastive property relating to this issue is duration. Long and short vowels exhibit different degrees of VISC (Nearey and Assmann, 1986, p. 1297), which means that longer vowels undergo larger amounts of spectral variation through time than short vowels. This property reflects the more diphthongal qualities of long vowels. Other things being equal, it should in principle be more computationally demanding to recognise long vowels early on from plosive onsets in general compared to short vowels. This claim can be substantiated by the fact that larger deviations from the average weighted representation for a given vowel (i.e. its average formant frequency) contra its actual VISC variability would necessitate more stringent and/or perceptually demanding computations than for vowels with less variable VISC patternings. Since there tends to be much less variation in short vowels than in long ones with respect to how much the resonances deviate from their average weighted values, the required perceptual

computation will take longer to achieve. Therefore, although perception works on changes and alternations in the speech signal, it is important to remember that moment-to-moment variation in VISC reflects changes for the vowel alone rather than between different sounds. Therefore, larger changes from the computed average centre frequencies for a vowel will be reflected in greater variation in VISC. This claim is also supported by Gussenhoven's (2007) findings, according to which low vowels, which are inherently longer, are more difficult to recognise than high vowels, all else being equal.

In summary, vowel formant centre frequencies change through time in two respects: for their movement velocities and especially the direction in which the individual vowel formants move. These two properties may alternate somewhat differently as well depending on the phonetic qualities of contiguous consonantal sounds. After ca. 30ms into the aperiodic phase in a plosive onset, the listener may be able to establish the direction of formant change in an upcoming vowel (Nearey and Assmann, 1986, p. 1299). Rosner and Pickering (1994) show that listeners compute averages of the total amount of variability in VISC, in order to derive representative values (or indicators) for vowel formants. This computation enables recognising different vowel sounds more reliably.

The next two subsections take up two other key issues relating to timing which can be seen to share a close relationship with VISC, which are 'articulatory-phonetic timing', 'information encoding for vowels' (at the phonological level) and 'order effects'.

2.1.2 Articulatory-Phonetic Timing

There are two properties of timing relating most closely to the temporal and articulatory aspects of speech timing: these

properties are “the Minimum Gesture Duration” and “Syllabic Target Alignment” (cf. Xu, 2009).

An articulatory gesture is defined as a unidirectional movement approximating toward a particular articulatory target state. The important thing to consider in this work remains how much of an effect the minimum perceptible duration of an articulatory gesture can have on surface variation (see Xu, 2009, p. 908). According to Klatt (1976, p. 1215), a sound segment is observable only if it is of longer duration than the minimum duration allowed by employing the maximum speed within the context of a given articulation. Unless a sound is of a given and adequate duration when produced at the maximum articulatory velocity of the articulators involved, it may be omitted from an articulatory-perceptual viewpoint (Xu, 2009, p. 910). A sound can only be compressed so much in articulatory terms before it becomes totally perceptually masked or overlaid by surrounding articulations: for example, certain vowels and/or consonants often receive very little stress in phrases such as ‘operatic society’ (= [sˈsaiətɪ]), to the extent that the segment in question is inaudible and/or articulatorily unmeasurable (Laver, 1994, pp. 147-48). Even stressed vowels need to have a given duration in order to be recognised as such in English (see Klatt, 1976, p. 215): an articulatory gesture cannot be compressed beyond this physiological-perceptual standpoint and still be recognised as a stand-alone segment (such as the release of a plosive or as a vowel).

Xu (2009, p. 910) discusses the fact that there are not only long articulatory transitions between sounds, but also at utterance onsets. An example of Mandarin Chinese tones is given. As argued in this subsection, the features observable at the beginning of a tone can signal a common articulatory origin that is in reality implemented before voice onset.

It has also been shown by Janse (2003) that the speed at

which speakers are asked to utter stimuli affects the intelligibility of the output, so that faster rates are less intelligible than normal and slower rates. When this process is implemented through synthetic manipulation on a computer up to three times the maximum rate of a speaker, intelligibility remains fairly high. This finding shows a) that the human speech perception mechanism is less constrained than the articulatory one, especially in terms of processing capability and b) that articulatory rather than perceptual constraints constitute the real 'hindrance' for information encoding.

The phonetic realisation of the syllable can be seen as the time interval during which articulatory target movements are approached and overlaid. Syllable onsets, which can be viewed as the real-time realisations of the acoustic output, serve as time markers: they contain information on the unidirectional movements toward the onset, vowel, coda as well as related suprasegmental properties, since the articulatory planning for all these constituents and parts of the output starts simultaneously, see Xu, 2009, p. 911:

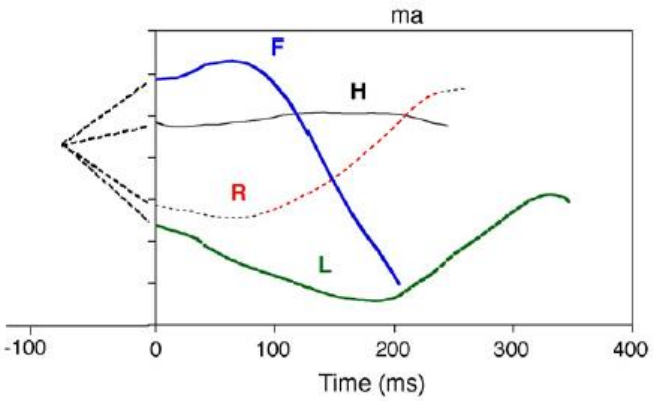


Figure 11: Isolated Mandarin tones with hypothetical silent initial F0 movements (Xu, 2009, p. 911, fig. 3)

The point Xu (2009) is making can be substantiated when assuming that prior to articulating a word or an utterance, speakers' vocal tract configuration will be neutral. Speakers

may adopt a fairly neutral articulatory configuration prior to articulating a monosyllabic utterance. However, since the F0 values for the four tones of Mandarin are distinguished at voicing onset (Xu, 2009, pp. 910-911), it may that the vocal-fold tension adjustment starts prior to the onset of voicing. For example, the beginning portions of the intonation contours together might be seen to point back to a neutral value (see the black lines on the left hand side of figure 11). Although the addition of the lines in figure 10 could be seen as arbitrary, the point Xu (2009) is making relates to the need to have adequately organised recurrent co-onsets of events to serve as time markers in speech. The example through which this point is demonstrated here corresponds to the onset of voicing. What properties remain available for encoding contrasts is assessed in the next subsection.

2.1.3 Functional Timing (Information Encoding)

Since the perceptual responses to different articulatory targets are determined by initiatory, phonatory and articulatory mechanisms, the temporal alignments of syllable junctures and turning points between them cannot be directly controlled for encoding contrasts. Arvaniti and Garding (2007) and Atterer and Ladd (2004) claim that there are cross-language and even cross-dialectal differences in F0 alignment, whereas Kohler (2005) has shown that listeners are sensitive to experimental manipulations of turning point locations. Such effects can just as well be considered as differences in the underlying pitch targets and target assignment for given syllables. For example, the assignment of a particular vowel target to a given syllable will always lead to at least some changes in the FPD of its articulatory alignment with neighbouring syllables (i.e. this is beyond speakers' articulatory control). Such changes could be used to deduce what the underlying target aimed at might be

(Xu, 2009, p. 919).

As speakers cannot control for the temporal alignment of underlying pitch and articulatory targets, only duration remains a controllable space in terms of information encoding. However, this space for controlling pitch targets and articulation remains a considerably large tool for encoding contrasts (Xu, 2009, p. 920). Two good examples of this kind of control are gemination for consonants (e.g. in Finnish and Estonian) and duration for vowels (in a range of languages, such as Finnish, German, Icelandic and Thai). Duration is used to distinguish long and short consonants and vowels in these languages, and is directly available as a means of encoding contrasts.

The important issue in this context is that duration remains the only phonetic property having to do with the actual temporal magnitude to which vowel sounds extend through acoustic space that speakers have notable control over (and which correlates with VISC), in terms of distinguishing similar sounding words with similar syllable structures. Despite duration also reflecting e.g. speaking rate and attention, speakers can distinguish vowels and words containing long and short vowel counterparts to the extent they wish in order to signal contrasts in monosyllabic utterances. Other lexically non-contrastive properties (such as articulatory-phonetic timing and speaking rate) are not entirely under speakers' voluntary control. Articulatory gestures require a minimum duration to be perceptible, while vowel sounds as short as 10ms have been shown to be able to be distinguished in durational terms by listeners (Rosner and Pickering, 1994, p. 294).

In summary, certain durational properties are not discernible in vowels, while the categorisation of vowels into long and short categories remains entirely at speakers' voluntary control. In the final subsection on vowel timing, the we consider potential distinctions related to stimulus ordering,

and whether issues such as randomising the order of presentation might affect vowel recognition and its time course.

2.1.4 Non-Linearity in Vowel Perception: Order Effects and Perceptual Confusions

According to Repp and Crowder (1990, p. 2080), order of presentation in vowel perception experiments with random stimulus ordering can have several effects. Vowels presented in one direction are more often reported to be different from another than vice versa. Cowan and Morse (1986) suggested a vowel neutralisation hypothesis account to explain why such perceptual effects occur. Having been presented, the first vowel in a pair changes its quality in memory toward a more neutral schwa – i.e. listeners judge vowel quality according to certain "reference points" within the vowel space. Repp and Crowder (1990) conducted a set of three experiments using a wide range of vowels confirming Cowan and Morse's hypotheses. However, the hypothesis on the direction of change from, say, /ɪ/ to /iː/ was revised by Repp to suggest that the vowel presented first changes its quality in memory toward the interior range of the vowel space (Repp and Crowder, 1990).

Repp and Crowder (1990, pp. 2080-2081) go on to argue that a substantial portion of vowel discrimination performance can be explained by the contrast effects between the members of stimulus pairs. For example, direction of change influenced the recognition of /iː ɪ/, whereas for /e ɛ/ no such effects were observed. For this reason, /i/ is difficult to discriminate from a subsequent /ɪ/, whereas more robust discrimination is observed when the order is reversed. The recognition function may depend on vowel quality, so that for back vowels a reference value similar to /o/ is used, since

vowels yield somewhat different order functions (Repp and Crowder, 1990, p. 2083). Repp and Crowder (1990, pp. 2084 and 2086) also provide information that the test environment may influence such results. Interstimulus intervals may also affect perception, so that order effects increase with longer intervals.

Though it may be tempting to suggest that randomisation of stimulus order counterbalances for order effects, the particular order that the stimuli do occur in may influence the perceptibility of given vowels so that the perception of one is either enhanced or decreased. On an average such effects may not be substantial in a study of this kind.

To round up this subsection, it is possible to see perceptual confusions between vowels as deriving partly from order effects. The constraints set by the particular ordering and the amount of time taken by listeners between listenings could therefore influence whether, for example, in /ɪ/ in 'pin' is more likely to be confused with /ɛ/ in 'pen' or /a/ in 'pan', and vice versa.

Subsection 2.1 has summarised the general properties underlying the time course of vowel perception in a range of contexts. The most important of these issues are:

a) the moment-to-moment spectral variation associated with VISC
b) what properties remain controllable for speakers in distinguishing vowel sounds
c) order effects associated with different presentation orderings in vowel experiments.

The next subsection takes up the temporal dynamics of vowel perception by looking at the recognition in English monosyllables and other similar utterances (such as lVCVCs,

see West, 1999b).

2.2 Vowel Timing and Aspiration in English CV(V)/Cs

This subsection looks at the perceptual problem of recognising vowels from voiceless aspirated plosive onsets. The literature suggests four approaches, 1) contrast and representation, 2) phonological/structural variation, 3) FPD and coarticulatory direction effects and 4) long-domain coarticulation. This 4-way presentation of the literature allows showing to what extent the results of previous studies are consistent and what secondary topics or questions still remain unanswered in those studies.

The aim of this part of the literature review is to offer a commentary of each theme and the relevant results. Second, the purpose, design, findings and analysis methods are examined and critiqued. Since the methods of interpretation between some of the older (e.g. Cullinan and Tekieli 1979 and Winitz et al, 1972) and more recent studies (Hawkins and Stevens, 1985, Hawkins and Slater, 1994 and Hawkins and Nguyen, 2001) differ, this approach allows relating weaknesses in each of the older studies to similar issues described in the third chapter on the methodology.

2.2.1 The Relationship of Contrast and Representation to Recognition

Tekieli and Cullinan (1979) and Cullinan and Tekieli (1979) take a generative approach to recognition. The approach of both studies to contrast and representation is binary and quantitative. Tekieli and Cullinan (1979) examined vowel recognition timing in CV monosyllables with short vowels. The following stimuli were used: /i ɪ u ʊ æ ɛ ɑ ʌ/ and /b p d t g k ʧ ʤ/, yielding a balanced set of 64 CVs. Segments heard by 18

female listeners over headphones in three one-hour experimental sessions consisted of the initial 10 to 150 milliseconds of each stimulus in 10-ms steps. The stimuli in the second paper (Cullinan and Tekieli, 1979), which investigated recognition timing from aspirated plosives, comprised a subset of the 64 original obstruent stimuli with /p t k/ as onsets, yielding a total of 24 stimuli. Although it is not fully obvious from the wording in the text in either study that the gating paradigm was used, the method seems to be identical. All 1080 stimuli were segmented temporally and presented in a random order to listeners (Cullinan and Tekieli, 1979, p. 123-124). The results show that:

- The recognition of CVs involves binary choices between phonemes. The way listeners recognise duration, height and frontness involves distinguishing between long/short, high/low and front/back vowels. Tekieli and Cullinan (1979) suggest that responses tend to comprise lax[8] vowels having similar frontness and height values to those of the original heard stimulus (Tekieli and Cullinan, 1979, p. 117).

- Cues to frontness and height values occur within the first 10ms, whereas the tense-lax feature does not reach threshold until after 30ms. It is argued by (Tekieli and Cullinan, 1979, p. 117) that duration has phonemic value in English (this view is particular to Cullinan and Tekieli's claims).

- Recognition is more reliable from /t/ than from /k/

[8] For practical reasons, the tense-lax distinction of earlier studies is treated as equivalent to [+/- long] in this study.

and /p/. Considering the phonetics behind these types of onset, this is an odd claim. The surface contact area at the lips and tongue dorsum for /k/ and /p/ is more extensive than that for /t/, which also requires more rapid closing-opening movements when using the tongue tip/blade. It should in principle be expected to receive more reliable recognitions from /k/ and /p/ than from an apical/dental plosive like /t/. Tekieli and Cullinan (1979) and Cullinan and Tekieli (1979) do not offer an adequate explanation for the finding, which suggests a rather simple view of contrast. Rather than supporting their findings with solid phonetic considerations, Tekieli and Cullinan (1979) and Cullinan and Tekieli (1979) choose to refer to similar findings by Winitz et al. (1972) in reinforcing their claim. Dissimilar earlier findings on place of articulation of the onset by e.g. LaRiviere et al (1975) are ignored. Table 1 describes Cullinan and Tekieli's main results:

TABLE 4. Recognition threshold durations (in msec) for consonants in CV syllables.

	Consonant								*Means*		
Vowel	/b/	/d/	/g/	/dʒ/	/p/	/t/	/k/	/tʃ/	*Voiced*	*Unvoiced*	*All*
/i/	40	30	30	70	10	20	20	30	42	20	31
/ɪ/	20	30	40	60	10	10	20	20	38	15	26
/u/	40	30	50	50	10	10	10	20	42	12	27
/ʊ/	20	30	30	60	20	10	20	20	35	18	26
/æ/	30	20	20	60	10	10	20	20	32	15	24
/ɛ/	20	20	30	50	30	10	20	20	30	20	25
/ɑ/	20	30	40	60	20	10	20	20	38	18	28
/ʌ/	20	30	30	50	10	10	20	20	32	15	24
Means:											
Tense Vowels	32	28	35	60	12	12	18	22	39	16	28
Lax Vowels	20	28	32	55	18	10	20	20	34	17	25
High Vowels	30	30	38	60	12	12	18	22	39	16	28
Low Vowels	22	25	30	55	18	10	20	20	33	17	25
Front Vowels	28	25	30	60	15	12	20	22	36	18	27
Back Vowels	25	30	28	55	15	10	18	20	37	16	26
All Vowels	26	28	34	58	15	11	19	21	36	17	26

Table 1: Recognition threshold durations (in ms) for consonants in CV syllables. (Tekieli and Cullinan, 1979, p. 111, table 4)

Each vowel phoneme is listed on the top left-hand side of table

1, with the corresponding threshold duration results detailed toward the right of these values on the top part of the x-axis. The bottom left-hand corner gives the binary distinctive feature values for each vowel, with the threshold duration results detailed similarly. The right-hand side of table 1 shows the mean recognition thresholds for entire CVs from both voiced and voiceless obstruents, with the means for all results detailed on the far right. For example, by inspecting the top part of the bottom half of table 1, it can be seen that the recognition thresholds from /p t k/ differ according to vowel quality, so that recognition from /k/ is similar for tense and lax vowels (18 vs. 20ms). However, with /p/ and /t/ different results are received for lax vowels, where /p/ trails /t/ by 8ms (18 vs. 10ms). Identical results to the ones for /p/, /t/ and /k/ are received in terms of how height affects recognition timing. This claim does not apply to frontness, which is listed toward the bottom middle part of table 1.

In both papers by Cullinan and Tekieli, views on representation are based on a phonemic and linear view of phonological processing. For example, a larger magnitude of perceptually significant information on vowel quality is transmitted by frontness than by height (Cullinan and Tekieli, 1979, p. 129): the larger spacing covered within the vowel space by frontness is seen as transmitting more acoustic-perceptual information than height qualitatively. Frontness has less influence over the coarticulation between sounds and in particular vowel timing and perception than height (see e.g. Gussenhoven, 2007 and Harrington et al, 1999). For example, a 250 Hz change in F1 from [ɪ] to [ɛ] may involve a greater distinction compared to an equivalent change in F2 between e.g. [iː] and [uː], regardless of potential fronting of /uː/ in

most English varieties spoken in England. Despite there being less leeway for F1 to move in a vowel, more significant changes are correlated with height than with frontness (e.g. the degree to which VISC varies in high vs. low vowels). The lesser leeway for F1 to move translates to a proportionally greater perceptual distinction with a given alternation than for an equivalent change in F2. Cullinan and Tekieli's conclusions in this instance place too much emphasis on quantitative measures, where more qualitative ones are needed.

The same issue applies to the tense-lax distinction, since Cullinan and Tekieli (1979) argue that it has considerably less influence over the amount of vocalic information transmitted than either frontness or height. Since the qualities of long contra short vowels do vary somewhat in frontness and height due to the centralisation typical of short vowels (cf. e.g. Van Bergem, 1993) Cullinan and Tekieli (1979) suggest that the amount of information provided on vowels is correlated by their positioning within the vowel space. Since it is not clear from Cullinan & Tekieli's (1979) study whether the vowels studied were classified as long or short, this claim on duration seems premature. Considering the spectro-temporal distinctions between long and short vowels with respect to VISC, the claim is hard to defend. It also seems very odd to claim that duration has 'phonemic value' *in English*, as real words vowels were not studied in either paper by Cullinan and Tekieli.

In summary, previous studies having a contrast/representation type of approach take a relatively narrow and binary view of recognition. Insufficient attention is paid to the FPD and place of articulation of onsets in terms of recognising vowel quality. No explanations are provided why recognition from /t/ is more reliable than from /k p/. Insufficient mention is made of phonetic and phonological differences between long and short vowels. Rather, it is stated

that responses tend to comprise lax vowel responses for the most part, without offering explanations for the finding. The next subsection describes the second strand on structural variation and recognition in vowel recognition.

2.2.2 Structural Variation and Vowel Recognition

LaRiviere, Winitz and Herriman (1975) and Winitz, Scheib and Reeds (1972) take a structural and descriptive approach. LaRiviere et al (1975) investigated the reliability of recognition for plosives and vowels from CVs minus the vocalic transitions of plosive-vowel CVs (experiment one) and various segments, comprising the aperiodic portion (i.e. plosive burst + aspiration) + the vocalic transition and/or the full vowel (experiment two). The two experiments comprised /p t k/ as onsets + /i a u/, with ten phonetically naive undergraduate students listening to stimuli on headphones. The results indicate i) that the vocalic transition is not a necessary or sufficient perceptual cue for the recognition of plosive onsets and ii) that the aperiodic portion bears the heaviest perceptual load in terms of vowel recognition.

Table 2 shows the proportions of correct answers for vowels from CVs in the 1975 study by LaRiviere et al. The left-hand column indicates each of the nine CVs, whereas the three columns on the right and middle show the proportions of correct responses with transitions of 30m, 50ms and 70ms:

	CV	Original	Segment Aperiodic	Aperiodic + vocalic trans.	Vocalic trans. + vowel	Vocalic trans.
(c)	/pi/	1.00	0.83	1.00	1.00	0.60
	/pa/	1.00	0.88	1.00	0.91	0.96
	/pu/	1.00	0.78	0.93	0.88	0.58
	/ti/	0.98	0.85	1.00	1.00	0.70
	/ta/	0.98	0.73	0.98	0.93	0.88
	/tu/	0.98	0.70	0.26	0.98	0.06
	/ki/	0.93	0.93	0.88	0.98	0.88
	/ka/	0.98	1.00	0.91	0.90	0.88
	/ku/	1.00	0.55	0.98	1.00	0.56

Table 2: Proportion correct (c) vowel identification from each segment (LaRiviere et al, 1975, p. 473, table V)

Table 2 shows the original CV utterances in the leftmost column. The equivalent proportions for each CV stimulus are located in the middle and right-hand side of table 2. For example, in the middle two columns of table 2 can be seen the proportions of correct responses to stimuli comprising the consonant aperiodic portions and aperiodic portion + vocalic transitions. Adding the vocalic transition (see middle column) gives rise to an increase in recognition for /p t k/, with the biggest increase for /p/ (cf. the top rows of the second and third columns). Vowel quality interacts with this aspect, so that recognition of /u/ suffers more overall than that of /i/ and /a/ (see the third, sixth and ninth rows in the second and third columns from the left).

The results for /p/ are not consistent with those of Winitz et al (1972), Cullinan and Tekieli (1979) and Tekieli and Cullinan (1979). This finding shows two things:

i) methodological aspects can affect the results of transmitted

vocalic information from different types of onset, *and*

ii) issues concerning phonological, structural and especially coarticulatory variation must be emphasised in more detail in studies on vowel recognition.

For example, LaRiviere et al (1975) suggest that the recognition of /u/ suffers more than that of /i/ and /a/ for certain onsets. Therefore, an account needs to be given of the implications of increasing the coarticulatory distance between the phonetic properties of the onset and that of the nucleus, which may distort recognition.

The aperiodic and vocalic transition portions are redundant for vowel recognition from /p/ (LaRiviere et al, 1975, p. 474). The aperiodic portion carries the heaviest perceptual load in terms of recognising CV constituents. That is, adding *more of other* information on an upcoming vowel (such as the vocalic transition and/or the beginning of the steady state portion) does not offer the same degree of perceptual advantage to a listener as hearing the aperiodic portion (cf. increments between stimulus options in table 2). The vocalic transition alone is not a necessary or sufficient cue to either plosive or vowel recognition in CVs. The vocalic transition may constitute a more essential cue to perceptual cohesion than to recognition according to LaRiviere et al (1975).

Since the aperiodic portion carries the largest perceptual load in terms of recognition, experiment two offers good evidence for coarticulatory cues of the vowel on the aspiration portion. However, LaRiviere et al's (1975) discussion is not always transparent, and remains descriptive. The results and conclusions presented remain partly unclear.

The purpose of the study by Winitz et al. (1972) was to

investigate the perception of stimulus segments excised from words with initial and final /p t k/, constituting the plosive burst, and burst plus 100 ms of vowel (Winitz et al, 1972, p. 1309). College students with no training in linguistics or phonetics served as subjects in sets of four or fewer: the precise number of listeners is not mentioned. In contrast to the other experiments detailed thus far, the listeners heard the stimuli over loudspeakers in an IAC sound module.

Only the second (VV) condition in each experiment looking at vowel recognition is presented. In contrast to the other studies reviewed thus far in this subsection, Winitz et al. examined the recognition of not just the vowel, but the CVCs that they had been lifted from: examples of sentences include "Toot that horn at your old coot", "Keen eyesight can't be beat" and "Pop the cork over the top" (see Winitz et al, 1972, p. 1310 for the complete list).

		SPOKEN /i/ /p/	/i/ /t/	/i/ /k/	/ɑ/ /p/	/ɑ/ /t/	/ɑ/ /k/	/u/ /p/	/u/ /t/	/u/ /k/
Initial										
/i/	/p/	0.83			0.17			0.01		
	/t/		0.97			0.08			0.27	
	/k/			0.90			0.03			0.05
/ɑ/	/p/	0.12			0.58*			0.17		
	/t/		0.03			0.70			0.08	
	/k/			0.05			0.87			0.20
/u/	/p/	0.05			0.25			0.82		
	/t/		0.00			0.22			0.65	
	/k/			0.05			0.10			0.75
Final										
/i/	/p/	0.53*			0.30			0.09		
	/t/		0.67			0.33			0.30	
	/k/			0.88			0.25			0.27
/ɑ/	/p/	0.33			0.43*			0.13		
	/t/		0.11			0.35			0.13	
	/k/			0.00			0.53*			0.18
/u/	/p/	0.14			0.27			0.78		
	/t/		0.22			0.32			0.57*	
	/k/			0.12			0.22			0.55*

Table 3: Confusion matrix for condition VV, burst only, Expt. II (Winitz et al, 1972, p. 1313, table VIII)

The numbers represent the proportion of correct answers for CVs (top). Proportions on the diagonal not better than chance at $p < 0.05$ are starred.

		SPOKEN								
		/p/	/i/ /t/	/k/	/p/	/ɑ/ /t/	/k/	/p/	/u/ /t/	/k/
					Initial					
/i/	/p/	0.95			0.02			0.01		
	/t/		0.91			0.03			0.38	
	/k/			0.92			0.03			0.04
/ɑ/	/p/	0.02			0.97			0.10		
	/t/		0.04			0.86			0.06	
	/k/			0.05			0.91			0.19
/u/	/p/	0.03			0.01			0.89		
	/t/		0.05			0.11			0.56	
	/k/			0.03			0.06			0.77
					Final					
/i/	/p/	0.89			0.02			0.13		
	/t/		0.95			0.04			0.27	
	/k/			—			0.05			0.17
/ɑ/	/p/	0.01			0.97			0.06		
	/t/		0.01			0.75			0.05	
	/k/			—			0.82			0.09
/u/	/p/	0.10			0.01			0.81		
	/t/		0.04			0.21			0.68	
	/k/			—			0.13			0.74

Table 4: Confusion matrix for condition VV, with 100 ms of adjacent vowel, Expt. II (Winitz et al, 1972, p. 1313, Table X)

Each number represents the proportion of correct answers for CVs (top). Proportions on the diagonal not better than chance at $p < 0.05$ are starred.

In tables 3 and 4 are given the proportions of recognitions in Winitz et al (1972), as well as the proportional distribution of perceptual confusions across each stimulus category. The top halves of tables 3-4 present results on the initial CV portions, with the results for the VC parts presented in the bottom parts of tables 3-4. The vowel-plosive combinations are displayed in the left-hand columns and top rows. By inspecting the numeric values in the six boxes (cf. left, right and middle) in tables 3 and 4, the recognition proportions for different vowels can be seen. For example, the top left-hand side of table 4 shows that recognition of /i/ from /p t k/ gives 95% correct responses, with 0% confusions as /ɑ/ and 5% as /u/. Recognition levels for /i/ are high and partly dependent on plosive place of articulation (see the second and third columns on the right of the top left-hand column of table 4), with the proportions of

incorrect responses being lower than for /ɑ/ and /u/. This finding suggests that the articulatory gestures for /i/ occur jointly with the release of /p t k/. It is suggested by Winitz et al. (1972, p. 1313) that this finding offers strong evidence supporting the established claims of anticipatory coarticulation (see e.g. Öhman, 1966 and Daniloff and Moll, 1968). /u/ is confused most readily with /i/ following the alveolar plosive /t/. It is concluded by Winitz et al (1972) that the high burst for /t/ may be interpreted by the listener as /i/, where F1 and especially its F2 value would be more concomitant with the burst for an alveolar.

By looking in more detail at the recognition proportions presented with the burst portions only (see table 3), it can be discerned that the proportions of perceptual confusions are greater than for the 100ms aperiodic portion CVs. For example, by inspecting the left-hand part of the middle column in the top part of table 3, it can be seen that recognition of /u/ from /p/ gives 58% correct responses, with a 1% and 25% spread for responses to /i/ and /ɑ/, respectively.

The conclusions presented by Winitz et al. (1972) are brief and descriptive, lacking sufficient detail on the reasoning behind the findings (see e.g. p. 1316). Key methodological details, such as how many listeners participated are not mentioned. Winitz et al.'s use of real English CVC words is an apt choice. However, since the word to be recognised was not made unpredictable, little control was exercised over how this particular issue might skew recognition.

Winitz et al (1972) discuss the significance of the fact that the listener needs to adopt different listening strategies when faced with surface phonetic variability. In this sense,

Winitz et al come perhaps closest of all studies on vowel recognition from plosives in characterising the perceptual importance of listening and coarticulatory strategies.

From the viewpoint of this research and the theoretical framework applied in it, LaRiviere et al's (1975) and Winitz et al's (1972) study come perhaps closest to assessing the role of FPD in the recognition of monosyllables. The statements made are often explained in more detail and less attention is paid to the primacy of phonemes in speech perception than in e.g. Cullinan and Tekieli (1979). Two good examples of this distinction are Winitz et al's claim about surface phonetic variability and listening strategies as well as LaRiviere et al's claims about recognition being most reliable from /p/. On the other hand, since word frequency and semantic context were not sufficiently controlled for by Winitz et al (1972), this thesis study remains better motivated in this respect (cf. chapter 3).

2.2.3 FPD and Coarticulatory Direction Effects

The studies by Ostreicher and Sharf (1976) and Waldstein and Baum (1994) approach vowel recognition from the viewpoint of coarticulatory direction effects and FPD. The authors compare the magnitude of anticipatory and carryover coarticulation and their perceptual consequences. The two studies look at different structures: Waldstein and Baum (1994) investigate the perception of CVs and VCs, while Ostreicher and Sharf (1976) also look at VCVs and CVCVs. Ostreicher and Sharf (1976) discuss the perception of conversational speech with respect to mono- and disyllabic utterances.

The purpose of Ostreicher and Sharf's study was to determine i) to compare coarticulatory effects on the perception of consonants and vowels, while ii) ascertaining to what extent anticipatory and carryover coarticulation affect recognition.

The four syllable types of CV, VC, VCV and CVCV were presented binaurally through headphones to a group of 45 listeners, whose task was to recognise the subsequent or preceding sound. The consonantal stimulus choices comprised /p t k b d g f s ʃ v z m n/ while the vowels examined constituted /i ɜ[9] u o ɔ/.

For vocalic features, Ostreicher and Sharf (1976, p. 292) show that the recognition of height was significant above chance at the 0.001 level in 4 out of 6 listening conditions, whereas the recognition of frontness was significant at the same level in all six conditions.

Ostreicher and Sharf (1976, pp. 292-293) claim that proximity in the vowel space positively affects recognition, so that phonetically similar sounds are more reliably recognised. The results showed that 4,382 error responses were closer to the correct answers, whereas 3,115 errors were farther from the correct response. For consonant and vowel recognition, a goodness of fit chi-square analysis showed significant differences at the 0.001 level.

Ostreicher and Sharf (1976, p. 293) show that five out of six comparisons of subtests favoured anticipatory effects with none favouring carryover ones. This result gave an overall finding of 33 instances favouring anticipatory coarticulation and only two having significantly greater values for carryover coarticulation.

In sum, anticipatory effects may be more important for recognition from obstruent consonants than carryover effects. This finding could be explained by the potentially more mechanistic nature of carryover coarticulation: planning for sounds yet to come may require more detailed planning than

[9] The authors use the ð symbol consistently for this vowel in their paper, which may simply be a misprint by the journal editor.

moving away from sounds already realised (see e.g. Whalen, 1990).

Ostreicher and Sharf (1976, p. 297) claim that in terms of the directional effects coarticulatory features undergo in CV, VC and CVCV/VCV-type utterances may be used by listeners in conversational speech. Listeners could anticipate the articulation of an upcoming sound and enhance perceptual speed and efficiency or to aid the recognition of sounds already spoken that have not been accurately perceived (e.g. due to background noise). Coarticulatory cues in speech are supplemental to those occurring in sounds to be recognised. In conversational speech listeners may pay attention to coarticulatory cues' functional value, whether or not their recognition levels are high (Ostreicher and Sharf (1976, p. 298). Such claims can be seen to emphasise the importance of different listening strategies in a similar sense to Winitz et al. (1972).

Waldstein and Baum (1994) investigated the recognition /i u/ from /ʃ/ as well as /t k/ and vice versa. The main purpose of the study was to ascertain to what extent production and perception of such stimuli by speaker-listeners with hearing loss compares with recognition for normal hearing listeners. 10 speakers and 10 listeners participated (5 hearing-impaired and 5 normal hearing), who were presented with stimuli comprising only the initial consonant + the aperiodic consonantal portion: the experiments comprised a 4-way distinction in terms of coarticulatory direction and type of speaker-listener, with 5 hearing-impaired and 5 normal-hearing listeners in each of the 4 participant groups.

Recognition from stimuli produced by speakers with normal hearing was more reliable than from those produced by the hearing-impaired speakers. All vowels were recognised above chance level in the anticipatory condition, except for /i/

following /ʃ/ as produced by the hearing-impaired speaker group. Recognition accuracy for carryover instances trailed that for anticipatory ones (Waldstein and Baum, 1994, p. 952).

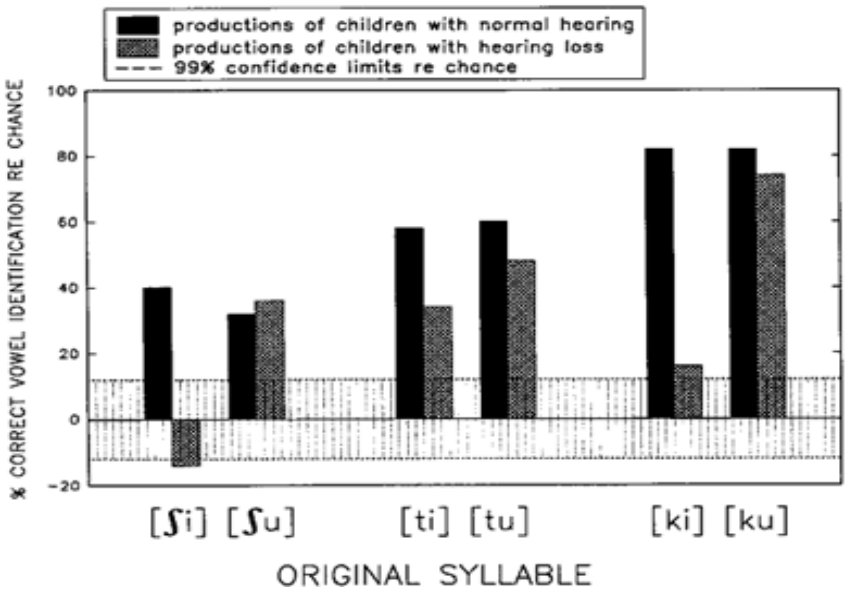


Figure 12: Mean percent correct listener identification of missing vowels excised from CV syllables produced by children with normal hearing and children with hearing loss (Waldstein & Baum, 1994, p. 955, fig 1.)

(Note: Scores were corrected so that a chance score of 50% is represented by 0% in figure four. The shaded area shows the 99% confidence limits for scores expected on the basis of guessing).

For the recognition of CV productions by children with normal hearing, it can be seen in figure 1 that both vowel quality and the place of articulation of the onset affect recognition (see the black bars on the middle and right-hand side of figure 12). In contrast to most other studies reviewed in 2.2 (cf. e.g. Winitz et al, 1972 and LaRiviere et al, 1975), Waldstein and Baum's results show similar recognition levels for /u/ and /i/ from all three types of onset, although recognition from /ʃ/ and /t/ trails recognition from /k/. This finding may reflect the larger contact area for velars than for alveolars and palatals, engendering more reliable acoustic-perceptual cues.

To summarise the two studies by Waldstein and Baum (1994) and Ostreicher and Sharf (1976), sufficient detail on vowel recognition from a wide range of obstruent consonants using a wide range of listener types is provided. Sufficient explanatory detail is provided on the bidirectionality of coarticulation by Ostreicher and Sharf (1976. The two studies together help to motivate this study of English as spoken in England, since insufficient reference is made to contextual effects, such as the potential phonetic co-extensiveness between onsets and codas. The results on vowel confusions by Ostreicher and Sharf (1976) form the basis of the hypothesis on how listeners make selections on vowel response choices as well (cf. subsection 2.5 on hypotheses).

2.2.4 Long-Domain Coarticulation and Airflow in CV(C)s

This subsection will focus mostly on CVNs, as these are most relevant to this research in the three studies reviewed in this subsection, and as they have very different phonetic consequences from stimuli with coda /p t k/. Some references are first made to previous research for stimuli with lateral onsets (see West, 1999b and Hawkins and Nguyen, 2001) in order to establish the reasoning for the inclusion of CVNs in this subsection more strongly (also see Hawkins and Stevens, 1985). There are five individual parts to this subsection, the first two dealing with long-domain coarticulation associated with liquids, and the latter three dealing with the acoustic consequences of long-domain coarticulation and airflow associated with nasal codas. Thus, the studies reviewed in this subsection also partly focus on aspects of speech production (cf. e.g. Cohn, 1990, Stevens, 1998 and Chang et al, 2011).

The studies on the role of long-domain coarticulation and airflow in the perception of coarticulation entertain a more

non-segmental view of recognition than the other studies reviewed thus far.

West's (1999b) perceptual study on the perception of distributed coarticulatory properties of English /r l/ provides reliable evidence that long-domain coarticulatory information about the /l/-/r/ distinction is perceptually available to listeners. For example, listeners were able to distinguish words such as 'berry' and 'belly' in a carrier phrase when both the liquids and parts of the surrounding vowels were replaced by noise (West, 1999b, p. 405).

A factor requiring clarification in this context is the issue of secondary resonances associated with specific sounds and how it may affect their coarticulatory properties and co-ordination with other adjacent and/or non-adjacent sounds. Discussing English /l r/, West (1999b, p. 406) shows that the clear/dark terminology is not restricted to liquids: Kelly and Local (1986, p. 304-5) refer to a description of English /n/ as being "duller" than German /n/, in the sense that the glide into a nasal in German is more rapid and thus perceptually clearer than in English from an impressionistic perspective.

Kelly and Local (1989) offer the suggestion that the clear/dark terminology is best viewed as the reflex of the significant but largely neglected phonetic or phonological phenomenon, which the authors term "resonance". Despite Kelly and Local's use of this term in connection with a radical view of phonology, the view of secondary articulations as reflecting "resonance" is not new. For example, Delattre (1965, p. 13) claims that for apical consonants, tongue shape and point of articulation play a role in creating the auditory impression of a language: apicals contribute to the degree of "frontal resonance". In sum, the term "resonance" may comprise a

significant reflex for the perception of coarticulation.

The claims on long-domain coarticulation and resonance associated with /l r/ suggest that a clearer view of how tongue and lip-movement dynamics may affect the co-ordination and timing of coarticulatory movements should be established. One potential solution to this is to pursue a wider search for other types of secondary resonances in other sounds (see e.g. Kelly and Local's 1986 reference to the 'dullness' of English /n/) and ask whether such properties of other sounds may have significant effects on perception. Kelly and Local's (1986, 1989) studies suggest that these kinds of resonance distinctions may be associated with different phonetic realisations for different sound types. Specific points or places of articulation within the vocal tract may realise different acoustic effects. For example, dorsal consonants and vowels lead to fronter or backer articulations, while apical consonants are associated with different kinds of secondary articulations and fronter or backer articulations.

The analysis of the perception data in West's (1999b) research was performed in terms of the linguistic material completely replaced by noise. A segment consisting of noise that began in a consonant preceding the core portion of the liquid and which ended in the first consonant subsequent to the core portion of the liquid sound was labelled as replacing the sequence VliqV. Noise segments beginning in the consonant preceding the core portion of the liquid sound which end early in the following vowel were labelled Vliq. Those stimuli ending in the middle of the following vowel were labelled as Vliq1/3V, while the ones ending late in the vowel were denoted as Vliq2/3V. Noise segments that ended late through a consonant were denoted as replacing half of the consonant: for plosives these segments replaced the hold phases while leaving at least some of the burst portion audible (West, 1999b, p. 413).

As far as the resonance distinctions in different dialects are concerned, West (1999b, p. 418) notes that RP listeners correctly recognised stimuli for a wider range of noise categories than the Manchester participants. Both groups exhibited a remarkable long-domain effect: correct recognition when the noise obscured (V)rVCV1/2C and lVCV1/2C. The results are consistent with the supposition that RP and Manchester English share similar long-domain resonance distinctions for liquids.

According to Hawkins and Nguyen (2001, p. 1), syllable-onset /l/ in British English has differing phonetic properties depending on coda quality: the lateral is longer and often has a lower F2 frequency before voiced codas. The five experiments conducted by Hawkins and Nguyen (2001) explored the perceptual power of these properties and F0. Using a forced choice procedure, listeners were asked to recognise synthetic word stimuli as 'led' or 'let'. The latter half of each stimulus was replaced by noise. The most reliable cue to recognition was the duration of the lateral sound; the influence of the frequency of F2 mainly depended on keeping vowel quality constant. Listeners learn which cues are most effective: some listeners choose duration rather than spectral properties relatively late in the perceptual procedure. The results support word recognition models with non-segmental lexical representation that is sensitive to systematic variation in FPD.

Hawkins and Nguyen (2001) propose the following:

> "... We hypothesize that even very subtle acoustic-phonetic properties can be salient perceptually as long as they indicate linguistic structure... Such systematic subtle phonetic variation will not necessarily provide strong perceptual information, but, by adding natural

> variation, it will increase the perceptual coherence of the speech, making it easier to understand in adverse conditions ([4][10])".
>
> Hawkins & Nguyen (2001, p. 1)

Hawkins and Nguyen's (2001) claim that phonetic properties of the coda may be temporally co-existent with properties of the onset in English single word CVC syllables is helpful. Hawkins and Nguyen (2001) emphasise the potential articulatory and acoustic influence of non-adjacent sounds forming part of the same lexical item. As a key aside, this type of claim is the kind of example referred to in chapter 1 about transcending the debate on phonemes vs. prosodies/non-segmental structures in favour of a more neutral polysystemic view of phonology and phonological processing.

In summary, Hawkins and Nguyen's (2001) research shows that perceptual cues to coda voicing are distributed across the words 'let' and 'led', not just the rhyme portions. Listeners display sensitivity to whether spectral properties of the onset and nucleus vary systematically and naturally (Hawkins and Nguyen, 2001, p. 4).

Hawkins and Nguyen (2001) hint at the perceptual complexity and sensitivity of the decision-making process: the training data suggest that listeners learn a lot about the phonetic properties of 'let' and 'led' during the listening process, so that they first experienced long laterals as spoken more slowly than short ones. Listeners then gradually started to focus more on duration, which is perceived as a more reliable cue (Hawkins & Nguyen, 2001, p. 4).

In the remainder of this subsection, the issue of nasality in the rime part of CVs and CVNs is discussed. The perceptual and phonetic consequences of coupling to a second resonance chamber are considered. The effects of coda nasality on aspiration in plosive onsets are examined, while presenting

what consequences different articulatory settings may have on such aerodynamic properties relevant to vowel recognition in monosyllabic utterances.

The phonetic properties particular to anticipatory nasalisation in vowels comprise increased bandwidth, lowered amplitude, nasal coupling and introduction of zeroes[10], changes in vowel quality, spectral balance and higher-frequency components. According to Hawkins and Stevens (1985, p. 1560), the main articulatory characteristic behind the production of nasalised vowels comprises the introduction of an acoustic coupling between the oral and nasal cavities at a location ca. halfway along the vocal tract (stretching from the glottis to the lips). This coupling has various acoustic effects, which include 1) shifting the natural frequencies of the vocal tract compared to the equivalent formant frequencies for a corresponding non-nasal vowel and 2) the addition of nasal pole-zero pairings to the vocal-tract transfer function. Hawkins and Stevens (1985) show that of these two effects, the main and most consistent effects on the spectrum of a vowel tend to be at low frequencies in the vicinity of F1. The shift in the F1 frequency can be explained by the gradual increase in the cross-sectional area of the velopharyngeal opening. The coupling to the nasal cavities tends to lead to the introduction of an extra pole-zero pair near F1.

Now is a good time to discuss how coda nasality may affect the phonetic properties of the aperiodic phase of English voiceless plosive onsets. The main two pieces of research dealing with this issue are Cohn's (1990) PhD thesis on "Phonetic and phonological rules of nasalization" (for English, Sundanese and French) and the aerodynamic study of nasality in Taiwanese and French by Chang et al (2011).

According to Cohn (1990, p. 152), oral airflow is very

[10] Spectral areas with little or no energy (Stevens, 1998, p. 198) introduced into the vowel filter function

high during aspiration and could magnify the effect of slight nasalisation in vowels. For example, in two productions of the word ‘pen’ reviewed by Cohn the second instance of the word has extensive nasalisation during the aperiodic phase. The onset of nasalisation coincides with the onset of aspiration, so that nasal airflow is of near-identical magnitude with the oral airflow.

Since Cohn (1990, p. 154) shows that aspiration is never nasalised in word forms like ‘pet’ and ‘ped’, the nasalisation *cannot* be seen as spontaneous. Rather it is triggered by the presence of a following non-adjacent nasal consonant. This finding confirms historical researches of the close connection shared by nasality and aspiration (see e.g. Ohala, 1975 and Matisoff, 1975). Cohn (1990) suggests a phonological rule for this phenomenon, according to which a [+ nasal] specification may spread back to a sound specified as [+ spread glottis].

According to Chang et al’s (2011) study on the phonological patterning and phonetic implementation of nasality in French and Taiwanese (2011, p. 436), various contextual influences, including onset and coda quality as well as manner may have significant influence over the phonetic implementation of nasalisation and the magnitude of nasal airflow in CVN monosyllables. For example, Chang et al show that there are systematically different effects of nasal anticipatory coarticulation induced by the phonetic exponency of the onset. Voiced plosives do not appear in nasal contexts in Taiwanese. In onset position, aspirated stops and fricatives have more nasal coarticulation. Coda /n/ includes a smaller amount of anticipatory vowel nasalisation in both languages compared to /m/ and /ŋ/: this result is important theoretically for this research, since only *alveolar* nasal codas were used (cf. chapter 3).

It is recognised that the results detailed by Chang et al (2011) are not peculiar to English. However, it has been shown in this subsection that the phonetic implementation of nasality may be dependent on the presence/absence of nasal vowels, so that presence of a phonemic opposition for nasal vs. oral vowels induces greater levels of resistance against anticipatory nasal coarticulation in CVNs. For this reason, these kinds of effects might be more widespread in English than in languages with a nasal-oral vowel opposition. The perceptibility of a vowel may reduce even further in the context of alveolar nasals, because it is easier to maintain a non-coronal closure in codas (Chang et al, 2011, p. 439). Table 5 details Chang et al's results:

		Taiwanese		French	
Coda					
	[m]	[n]	[ŋ]	[m]	[n]
% of nasalised volume	31	25	33	44	38
% of nasal time	40	34	41	40	30

Table 5: The effects of codas on the degrees of nasalization in the Taiwanese and French CVN contexts (Chang et al., 2011, p. 438, table 5)

The left-hand side of table 5 lists the two categories examined by Chang et al (percentage of nasalised volume in the middle and percentage of nasal time in the bottom corner of table 3). In the middle right-hand side of table 5 can be seen the various values associated with different places for Taiwanese (middle) and French (far right). For example, French /m/ has 44% of nasalised airflow volume and 40% of nasal time, whereas /n/ trails /m/ by 6 and 10 per cent in these respects, respectively (44 and 38% vs. 40 and 30%).

To round up this subsection, the phonetic consequences of speakers adopting particular articulatory settings and/or constellations for phonological processing from CVNs is discussed and described.

Carignan et al (2011, p. 668) suggest that speaker-listeners may compensate for the high degree of F1 centralisation in the context of high vowels by raising the tongue body in order to counteract the perceived nasality during the articulatory settings and constellations adopted during such vowel articulations. Carignan et al (2011, p. 669) refer to work by Wright (1975, 1986) who showed that listeners may also misperceive vowel height with low vowels, so that nasalised [ã] was perceived as higher than oral [a].

There are two more phenomena relating to nasality and vowel height in CVNs that still need to be addressed in this subsection. These phenomena relate a) to the location of the first nasal pole and b) airflow impedance. MacMillan et al (1999, p. 2913) pose the question whether the interaction between perceived vowel height and nasality results from the interaction results from a sensory process, decision mechanism:

> "A configuration derived by a multidimensional scaling analysis revealed a perceptual interaction that was stronger for stimuli in which the nasal pole/zero complex was below rather than above the oral pole, and that was present before both nasal and oral consonants... Judgments of nasalization depended on *F*1 as well as on nasalization, whereas judgments of height depended primarily on *F*1, and on nasalization more when the nasal complex was below than above the oral pole. This pattern was interpreted as a decision–rule interaction that is distinct from the interaction in basic sensitivity".
>
> MacMillan et al. (1999, p. 2913)

In summary, MacMillan et al (1999) show that

i) that the kind of FPD associated with anticipatory nasality may have a strong bearing on both vowel recognition and perceptual intelligibility

ii) individual differences relating to the acoustic location of the nasal pole/zero complex may have an important bearing on perceived nasality and perceived vowel height

For airflow impedance, the text will refer to Stevens (1998). Figures 13-16 describe the positions and constellations of the pharynx and oral cavities and the velopharyngeal port in /æ ɑ e o u ʊ i ɪ/ and the French nasal vowel /ɑ̃/:

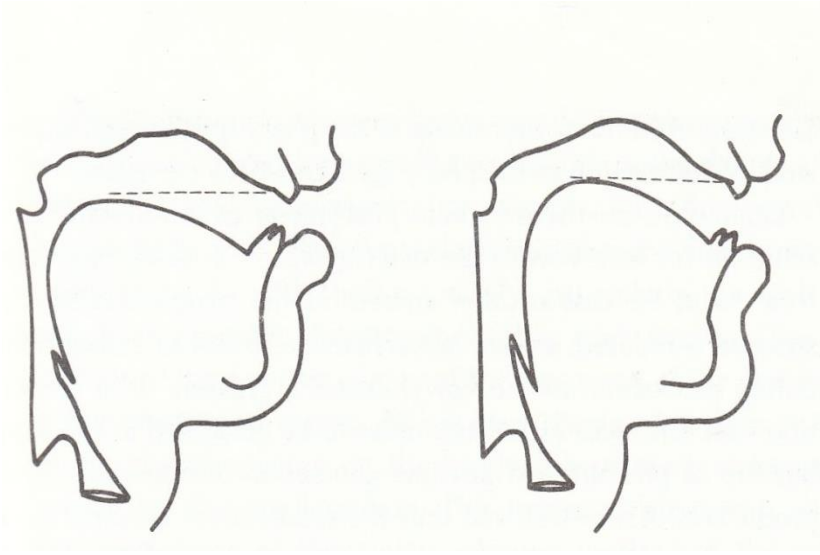


Figure 13: Mid-sagittal vocal tract configurations for the low vowels /ɑ/ (left) /æ/ (right) (Stevens, 1998, p. 269, fig 6.6)

In figure 13 we can see the mid-sagittal vocal-tract configurations for two low vowels, /ɑ/ and /æ/. The bottom

part of each picture half in figures 13-16 shows the glottis and outer part of the thyroid, whereas the top halves show the tongue and lips as well as front teeth and mouth openings. Figures 14-16 for /e o u ʊ i ɪ ɑ̃/ are organised similarly as figure 13.

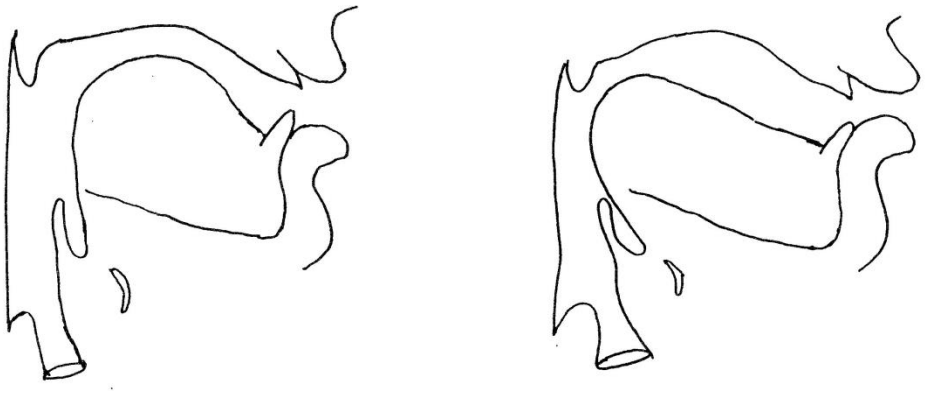


Figure 14: Mid-sagittal vocal tract configurations for the mid vowels /e/ (left) and /o/ (right) (Stevens, 1998, p. 271, fig. 6.7)

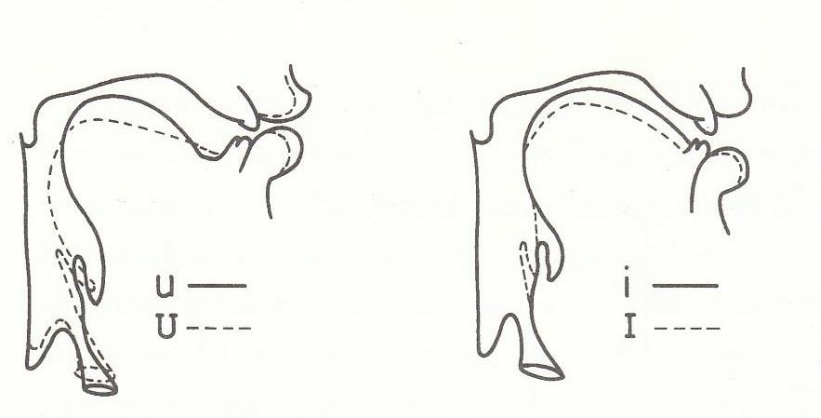


Figure 15: Mid-sagittal vocal tract configurations for the high vowels /u ʊ/ (left) and /i ɪ/ (right) (Stevens, 1998, p. 295, fig. 6.23)

The dotted lines describe the articulatory settings/constellations for the lax vowels.

Figure 16: Mid-sagittal vocal tract configurations for nasalised /ɑ̃/
Stevens (1998, p. 305, fig. 6.29)

Are the types of articulatory constellations described in figure 16 likely to affect the production and introduction of zeroes and resonances depending on vowel quality? Although it has been established earlier in this subsection that vowel height has an effect on the capacity of speakers to nasalise vowels, the broader articulatory constellations involved must also be considered, not solely the role of F1. For example, Stevens (1998, pp. 306-312) shows that the vocal tract vs. nasal cavity airflow transfer function and the size of the velopharyngeal opening in relation to the constrictions adopted in the vocal tract may affect the phonetic quality of the acoustic output. Figure 17 gives a schematic of the airflow and volume transfer function for nasalised vowels:

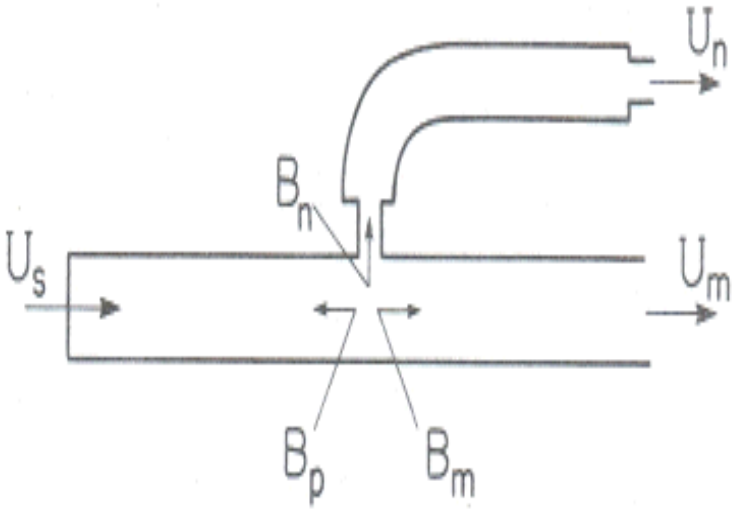


Figure 17: A schematisation of the shapes of the vocal and nasal tracts for a nasalised vowel (Stevens, 1998, p. 305, fig. 6.30).

The volume velocities Us, Un, and Um at the glottis, nostrils and mouth are shown, as well as the acoustic susceptances at the coupling point looking into the pharynx (Bp), the nasal cavity (Bn) and the mouth cavity (Bm).

For the purposes of analysing vowels that are specifically nasal, the vocal tract can be modelled as a system of resonators (Stevens, 1998, p. 305). There are two distinctive outputs to this resonator system, the volume velocity U_m at the mouth (see the lower right-hand part of figure 17) as well as the volume velocity Un at the nose (see the top right-hand part of figure 17). The sound pressure at a distance can be seen to comprise the combined resulting output $U_m + U_n$. The airflow and volume transfer function itself $(U_m + U_n)/U_s$, on the other hand, reflect the sum of the individual transfer functions U_m/U_s and U_n/U_s. All these functions have different zeroes but the same poles.

What are some of the acoustic consequences of this type of coupling in the context of the kinds of articulatory constellations and settings described in this subsection? There may be systematic differences in the sizes and shapes of the pharyngeal and mouth cavities as well as the opening of the velopharyngeal port. For high and low front vowels, a

comparatively larger portion of the aspiration may be absorbed subsequent to release, since a significant part of the air has to travel through a less uniformly shaped vocal tract, resulting in larger increases or decreases in airflow impedance during a CVN. The shaping of the overall articulation may have significant effects on the phonetic exponency and FPD of the aperiodic phase (see Cohn, 1990).

Having fully covered all the more phonetic literature behind vowel recognition and vowel timing, the final review exemplifies the phonological treatment of vowel recognition.

2.3 The Phonological Treatment of Vowel Recognition

Next, we will describe how the two main strands of literature reviewed in the previous two subsections are best given a phonological treatment. Subsection 2.3.1 takes up the phonological units and devices relevant to the general modelling of CV(V)/Cs and their temporal organisation in this context.

2.3.1 A Formal Model for Reconciling Inconsistent Findings on Vowel Recognition Timing: Units and Devices Available

In this subsection a detailed account is given of what aspects of phonological structure and phonological constraints are relevant to the treatment of VISC and (to a lesser extent) rime nasality. Additional and clarifying comments and descriptions are offered wherever these two properties affect the perception of temporal properties in phonological processing. For example, there is a need to adequately conjoin the 30ms locus point for vowel recognition (cf. e.g. Nearey and Assmann, 1986) in the aperiodic phase with the phonological and especially the temporal processing of vowel quality.

The descriptions given in this subsection are largely based on Coleman's (1998) monograph on phonological representations, along with additional descriptions and conclusions given by Hawkins (2003), as well as Simpson (2005) and Sprigg (2005). We begin by exemplifying syllable structure, while then moving onto feature spreading, domain of contrast, co-extensiveness of phonetic interpretation and, last, functional differences occurring at distinctive places in structure.

Coleman (1998, p. 279) shows that the grammar can be seen not to employ the category of 'vowel' at all, since nuclei are best analysed as three different types of object, short vowels, unchecked nuclei (i.e. onset + rime without coda) and vocalic rimes in English. As the phonetic exponents of vowels spread over entire syllables (Coleman, 1998, pp. 224-225), this treatment of vowels is justified. Since vocalic features are shared across higher elements of structure, and contrasts are expressed at different levels in this thesis, domain of contrast and its relationship with feature sharing are two key questions in vowel recognition.

Whether or not listeners are asked to distinguish between long and short vowels, they will need to take the kind of lexical detail associated with English CV(V)/C syllables into account in processing, since the phonetic exponents of these two types of object differ with respect to VISC. Since the phonetic instantiation for a vowel may in principle affect the precise location of the locus point for inferring formant trajectories, phonetic exponency may remain a perceptually significant variable.

Coleman (1998, p. 285) shows that features in English phonology are in many ways morphophonological and relational, while having no intrinsic phonetic interpretation. A good example of this property is voicing, in the sense that not all categories made with vocal-fold vibration are necessarily

classed as [+ voice] and vice versa. For example, voiced plosives in English are often phonetically voiceless (Docherty, 1992, pp. 115-116), while stressed intervocalic /t/ in words/phrases such as ‘better’ and ‘get a...’ may often be realised as [t̬] in many Tyneside accents (e.g. Watt and Milroy, 1999, p. 29): this generalisation is true for most other varieties spoken in England. Phonetic exponency is viewed as a secondary issue in terms of phonological processing and representation. Rather, distribution and functional oppositions determine the properties of feature spreading and domain of contrast.

Listeners make attempts to deduce how feature sharing and different domains of contrast are specified within phonological representations, in order to be able to work out what they are hearing and what is likely to follow a given constituent in parametric terms. Figures 18-19 give examples of feature sharing, temporal overlap of constituents and co-extensiveness of phonetic interpretation in an English monosyllable:

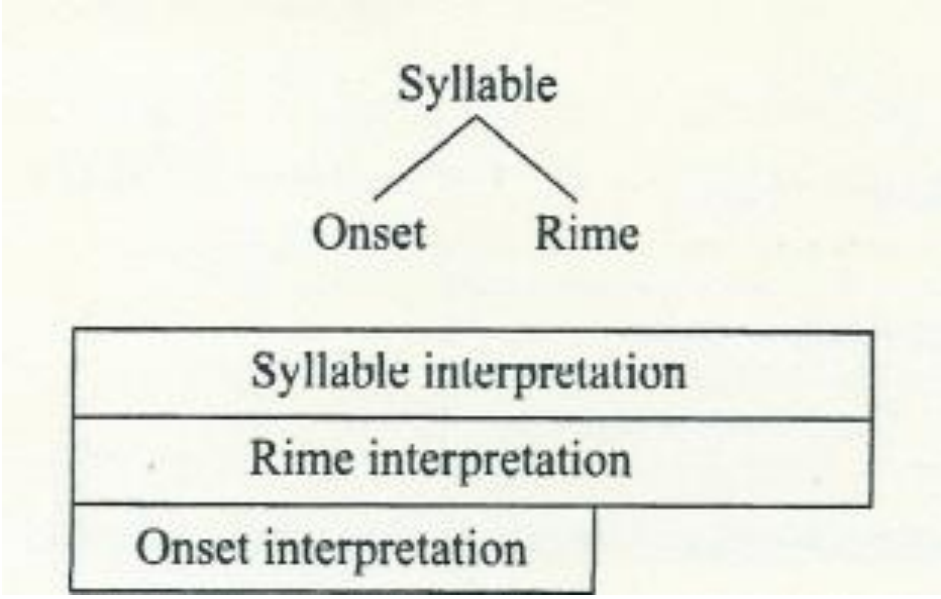


Figure 18: Temporal interpretation of syllable, rime and onset constituents
(From Coleman, 1998, p. 224, fig. 5.26)

Figure 18 shows how the interpretation of the exponents of the rime and syllable can be seen to co-extend over an entire CV(V)/C syllable. Since the exponents of the onset in large part overlap those of the syllable and rime, listeners may often be able to narrow down the quality of the upcoming rime portion to a fairly small number of contrasts or candidate categories. For example, having heard 'I think you say t' (with a [t^h] realisation containing only part of the aperiodic phase at the end of the utterance in /t/), listeners may be able to deduce that the upcoming vowel is /ɛ/ (as e.g. in the word 'ten'). Given enough phonetic information on the rime portion, it may also be possible to work out properties of the coda portion. Figure 19 illustrates this perceptual problem:

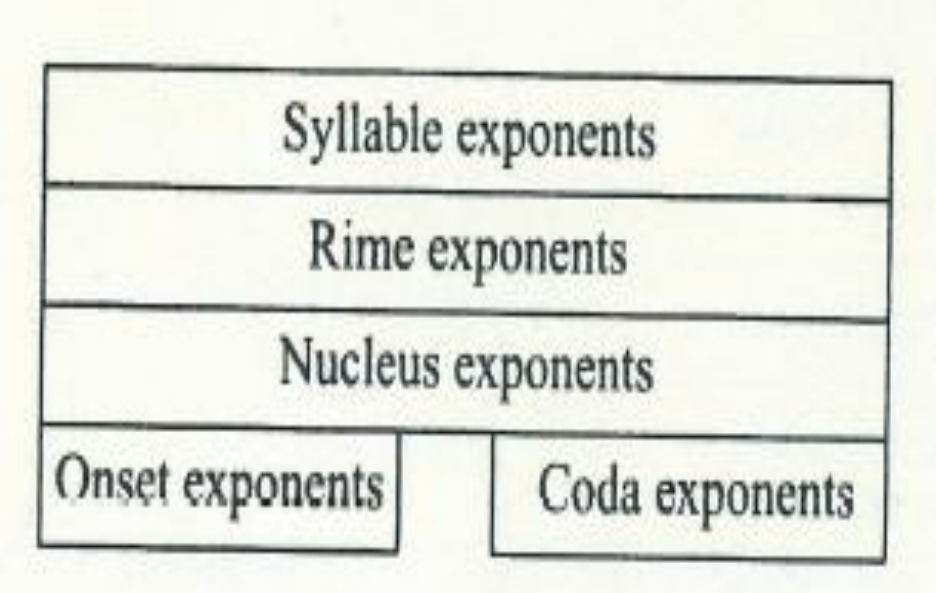


Figure 19: Temporal interpretation of syllable constituents (From Coleman, 1998, p. 224, fig. 5.27)

Since properties of the coda overlap those of the nucleus and those of the onset are temporally coextensive with the beginning part of the nucleus portion, listeners may be able to narrow down the phonetic quality of the coda, given enough cues to, for example, manner or airflow properties. For

example, Hawkins and Nguyen (2001) show that listeners may be able to distinguish ‘let’ and ‘led’ reliably during the realisation of the initial lateral. The phonetic exponents of the two types of onset /l/ differ significantly due to coda voicing.

When hearing exponents for structures and constituents of this kind, listeners need to take into account the varying distributions and in particular the functional oppositions of both consonants and vowels, as well as their structural positions. For instance, even though nasality may spread into the onset from codas (see Cohn, 1990 and Chang et al, 2011), it cannot be determined that nasality is a syllable-level feature, since its opposition is neutralised in onsets (Coleman, 1998, pp. 285 and 294). For example, words like ‘nan’ and ‘knee’ contra ‘ban-bad’ affirm the validity of this claim. In ‘nan’ and ‘knee’, the closure for the consonant is classified as [+ nasal] rather than being attested at the level of the onset: hence nasality is not contrastive in onsets. Feature spreading is not all that determines the correct specifications for contrasts, rather domain of contrast does.

As far as co-extensiveness of phonetic interpretation is concerned, this phenomenon can be seen in two different ways, the first of which is temporal interpretation (see Coleman, 1998, p. 216), which relates to phonetic exponency. On the other hand, there is parametric interpretation, which can be defined as:

> “a *relation* between phonological categories (feature structures) at places in structure (i.e. nodes in the syllable tree) and sets of *parameter sections*. A parameter section is a sequence of ordered pairs, each of which represents the value of that parameter a particular (salient) time”.
>
> Coleman (1998, p. 229)

Next, the temporal interpretation of phonological structures will be exemplified. According to Coleman (1998, p. 216), the distinction between ‘head’ and ‘non-head’ constituents is central to the model of phonetic interpretation. The temporal interpretation can be viewed in two ways: a) the general principles by which the phonetic exponents of phonological units are governed by a specific type of ordering with respect to each other in time and b) the parametric interpretation of consonant-vowel transitions. The former of these two aspects of temporal interpretation relates to the concatenation, co-catenation (i.e. vertical arrangement) and ordering of different pieces of structure, whereas the latter relates to the ways in which vowels are overlaid onto consonants in terms of their timing.

Coleman (1998, p. 225) asserts that it is good to ask whether the fact that the syllable, rime and its nucleus are coextensive in time makes them separate objects at all. This comment by Coleman is justified, since the phonetic implementation of each constituent is for all intents and purposes simultaneous. For theoretical purposes, the syllable rime and nucleus need demarcating: as neither the rime nor the nucleus is optional, it is possible to refer to each as ‘heads’ of pieces of structure, whereas ‘coda’ and ‘onset’ can be referred to as margins, which are equitable with non-heads.

Armed with this distinction, it is possible to make some generalisations about temporal interpretation (Coleman, 1998, p. 225). First, “the temporal domain of the head of a constituent is coextensive with the temporal domain of the whole constituent”. Second, “the temporal domain of the modifier of a constituent begins or ends at the same point at the temporal domain of the whole constituent, but is shorter” – i.e. for example, onsets and codas and terminal nodes. Last, the second point leads us to the conclusion that “the temporal domain of

the modifier constituents overlap the temporal domains of their head sisters" (Coleman, 1998, p. 225) – i.e. coda exponents may be temporally coextensive with parts of the onset. The principle is: whatever level of structure is being considered, the head is interpreted before its modifiers.

Coleman (1998, p. 229) offers the example of labiality, whose domain of phonological representation must be determined in relation to the functional opposition role it plays, even if its phonetic exponents can be observed throughout the entire temporal extent of a syllable. It must be shown whether the representation is located at or lower than the syllable node. However, the positioning of the latter could still mean that the phonetic exponents of labiality are coextensive with those of the syllable node (Coleman, 1998, pp. 229-230). An explicit specification of the phonetic interpretation of given pieces of structure through specific temporal constraints is needed.

Since nasality in the rime affects the spectral details of the rime, as well as those co-varying with vowel length and even phonetic properties of the aperiodic phase, a detailed account of its distribution and exponency is required. This claim can be explained by the complex articulations and the advanced contrastive/phonological planning that speaker-listeners need to carry out in adequately implementing nasality. As Temple (2009, pp. 152-153) confirms, nasality in English as a phonetic property is highly non-segmental, and is not often strictly co-temporal with all the other properties of the sound to which it "belongs".

This claim by Temple lends credibility to the proposition that the planning that is performed in e.g. CVNs could in principle be thought of as phonological. For example, as place for coda nasals may play a significant role in terms of nasal airflow and volume in various languages (cf. e.g. Chang et al, 2011), it is entirely possible that the vowel quality is more readily available in words forms like 'come' and 'king' than

from CVN monosyllables, since bilabial and velar nasals distort the spectro-temporal properties of initial plosives less than alveolar ones (cf. Chang et al, 2011, p. 437). Whilst nasality is a feature of the rime, the phonetic implementation of all sounds in an English monosyllable and in particular the coda may affect the phonetic properties of the onset. While such properties are not strictly speaking polysystemic, rather relating to articulation, they do suggest a broader view of polysystemicity in CVCs than in most previous studies.

2.4 An Evaluation of the Methods of Earlier Studies

In the papers on contrast and representational aspects of perception and phonological/structural variation, the authors of previous studies on vowel recognition from plosives have virtually nothing to say about how more subtle aspects of perception and complex perceptual mapping between constituents of different sizes may affect vowel recognition (see e.g. Cohn, 1990, Hawkins & Nguyen, 2001 and Goffman et al, 2008 for alternative views). Instead, the commentaries provided are descriptive and mostly relate to the acoustic aspects of recognition, including voice onset time (VOT) and formant movements. A more useful approach would take a wider perspective of recognition and emphasise the creation of linguistic and lexical meaning as well as coarticulation as a dynamic phenomenon.

A partial exception to the segmental-phonemic trend is Ostreicher and Sharf's (1976) study on recognition from obstruents. This paper offers a significant theoretical contribution to studies on the perception of coarticulation, since the authors recognise that isolation and canonical word forms taken out of context can only tell us so much about the true nature of speech perception. For example, Ostreicher and Sharf

recognise that recognition from obstruent-type consonants may have a role to play in the perception of conversational speech, a claim that is rather atypical of a 1970s paper. Winitz et al (1972) recognise the importance of surface phonetic variability, a position consistent with current research.

Stimulus choice in most studies reviewed is limited to /i a u/ in CV nonsense syllables with short vowels, other near-maximally contrastive vowels, or a large set of eight vowel phonemes (cf. Cullinan & Tekieli, 1979 and Tekieli & Cullinan, 1979). Only Winitz et al (1972) has investigated real words. Offering participants a choice between only two or three vowels may not give a thorough picture of recognition. This factor is particularly obvious when maximally contrastive vowels are exclusively used, since listeners should be more capable of distinguishing such stimuli in the first place. At the opposite end of the spectrum, giving listeners too many choices representing all vowel qualities may lead to confusion and other problems, since they are being asked to choose given categories from quite a large sample of materials (making the perceptual problem more complex and increasing the cognitive load).

A more balanced stimulus set using forced choice with four or five vowel choices may tap better into more subtle aspects of perception. For example, if listeners are asked to distinguish stimuli of the type “tin, tan, ton, ten”, this method may give a better picture of the processing of differences in vowel height and/or duration, because smaller acoustic differences are involved than looking at, only /t k/ + /i u/ (e.g. Waldstein & Baum, 1994) or /p t k/ + /i u ɑ/ (e.g. LaRiviere et al, 1975 and Winitz et al, 1972).

Perhaps the most important criticism appertaining to stimulus choice in previous studies is that it is often restricted

to /i a u/ (cf. e.g. LaRiviere et al, 1975, Winitz et al, 1972 and Waldstein & Baum, 1994) and yet claims about the tense-lax distinction are made. It is not in fact obvious that both long and short vowels were studied at all in any previous study (although Winitz et al used *one* short vowel against several long ones). any claims about the long-short distinction are at best tentative in previous research on the timing of vowel recognition from plosives. The key deficiency related to this issue is the focus on nonsense syllables in most previous research (with the single exception of Winitz et al, 1972). Although the use of nonsense stimuli allows for more optimal control than the use of real words (such as 'peel', 'keep' and 'tot' investigated by Winitz et al, 1972, see p. 1310), any claims made about phonological contrast and representation in such studies can be seen as tentative at best, since both long and short vowels in different syllable structures in real words were not studied in previous research. It is an open question whether results on lab speech and in particular on artificial data can be transferred to real words. For example, some of the studies reviewed in 2.2. include a set of CVs with "prolonged"[11] vowels of up to 150ms in duration (Tekieli Cullinan &, 1979, p. 104), which means that the sustained voiced portions of the vowels in such stimuli will never be sufficiently long enough in duration to be interpreted as long vowels (which are normally at least 200-250ms in duration, Hillenbrand et al, 1995, p. 3103). In fact, if we compare up to 150ms duration "vowels" with short vowels as produced by ca. 140 American English speakers of various ages by Hillenbrand et al (1995, p. 3099 and 3103), even short vowels normally have durations of ca. 180ms or longer.

For these two reasons concerning vowel duration, findings from such stimuli by Tekieli and Cullinan (1979) and

[11] It can be seen as peculiar that Tekieli and Cullinan (1979) and Cullinan and Tekieli (1979) represent "prolonged" vowels using short vowel symbols.

Cullinan & Tekieli (1979) may not give a representative picture of vowel recognition *for English*, since

a) no solid frame of reference (e.g. a word or a morpheme) is provided for processing, i.e. the forms provided have no lexical meaning.

b) claims are made by Tekieli and Cullinan (1979) about duration having "phonemic value" (p. 117) in English, yet none of the vowels in the produced stimuli is of sufficiently long duration to be fully representative of English.

To summarise, using *only* nonsense words as stimuli is problematic, since some of the resulting stimuli may have psychological reality, whilst others are meaningless. Using the kind of methodology applied in most previous research on vowel recognition from plosives (e.g. Cullinan and Tekieli, 1979 and Waldstein and Baum, 1994) is not as feasible as the one used in this study: the only way to achieve complete control over this issue would be to use synthesised plosive + short vowel stimuli in CV syllables only. Such an approach would not enable looking at how distinctions in phonetic exponency relating to vowel length (such as VISC) affect vowel recognition timing (= the most important secondary research question in this work). The resulting stimuli would comprise neither real meaningful words nor real speech, however good the synthesis. Such a methodology would also restrict the number of available stimuli, since to make the stimuli at all meaningful, only [pʰ tʰ kʰ] + [ɪ ɛ a ʌ ʊ ɒ][12] could be used as stimuli for the varieties investigated in this research, giving 3 * 6 (= 18) stimuli only. That amount

[12] These are the six short vowels that exist in English varieties as spoken in England. Studying other short vowels would be less representative of vowel processing in English.

comprises ca. 1/4 of the amount in the current experiment (60), which would probably limit the conclusions that can be drawn.

In their defence, word frequency is not an issue in these studies. However, since the stimuli examined are for all intents and purposes meaningless (that is, they are not words, syllables or morphemes), it is questionable to what extent such findings reflect vowel recognition timing *in English*. Since it has been established thus far that the magnitude of variability in VISC correlates with vowel length (with long vowels requiring more demanding perceptual computations), the results of some previous studies can be seen to only tell a small part of the story behind vowel recognition, even if it is assumed that processing of "prolonged" 150ms vowels in CVs does represent real English words sufficiently.

Since it has been shown in research on vowel timing that the exponents of long vowels differ in terms of articulatory-perceptual timing, some of the claims with respect to this issue can be seen as premature or based on the wrong type of experimental stimuli. A better alternative will be to look at various syllable structures with differing sounds and exponents, *for both long and short vowels*. If there is a constraint of using only syllables with short vowels (e.g. LaRiviere et al, 1975) or vowels with very short durations (cf. e.g. Tekieli & Cullinan, 1979), the conclusions we can draw with respect to phonological processing become fairly limited.

2.5 Secondary Research Questions

Several more research topics arise from the previous discussion. In terms of both perception and production, the following questions are very relevant to the study of vowel recognition from aspirated plosives, comprising further gaps in coarticulatory theory:

i) To what extent does English syllable structure and its relationship with VISC and phonetic exponency influence vowel recognition? For example, to what extent does the presence/absence of a coda affect the time course of vowel recognition? Although some of the previous studies on vowel recognition do study different structures (e.g. Cullinan & Tekieli, 1979 investigate CVs whilst Winitz et al only study CVCs), no mention is made in them about how phonological and syllable structure might affect the time course of vowel recognition in English. Therefore, considering the previous research on vowel timing and VISC, this is an obvious secondary research question to ask.

ii) How are perceptual confusions for different vowel types best explained, and to what extent does acoustic similarity between response choices affect recognition? Ostreicher and Sharf (1976) investigate this topic for North American English. However, since the theoretical framework in this study and the stimuli used are quite different from those investigated by Ostreicher and Sharf (both in terms of structure and the number of speakers recorded), previous answers to this question may not fully apply to CV(V)/Cs as produced in English varieties spoken in England.

iii) To what extent do differences in coarticulatory strategies between varieties spoken in England

affect vowel recognition and its timing aspects (see e.g. Wells, 1982, West, 1999b, Hawkins and Slater, 1994 and Kelly and Local, 1986)? This topic is not addressed in previous studies on the perception of coarticulation and vowel recognition, because previous research places too much emphasis on the perceptual targets listeners aim for in processing and uses nonsense syllables as stimuli (therefore, the question is not theoretically as relevant or as obvious). The fact that short vowel systems in southern and northern varieties of English (as spoken in England) have a different number of contrastive categories (5 in northern and 6 in southern accents) compared to North American varieties also makes this question relevant. This issue may have consequences for the ways in which speakers phonetically co-ordinate the articulation of different vowel sounds in CV(V)/Cs and how they are perceptually interpreted.

iv) Do coarticulatory direction effects (see e.g. Ostreicher & Sharf, 1976 and Modarresi et al, 2004) having to do with the bidirectionality of coarticulation contribute significantly to vowel recognition? For example, does phonetic influence related to anticipatory vs. carryover coarticulation concurrently affect the timing of vowel recognition in English? If so, how would such results extend previous findings (in particular those of Ostreicher and Sharf, 1976)? For example, though Ostreicher and Sharf study anticipatory and carryover coarticulation in

vowel recognition, almost no mention is made on how they together might reinforce recognition.

v) How can the findings of previous studies looking at non-segmental coarticulatory phenomena and long-domain resonances (cf. e.g. Hawkins and Nguyen, 2001, 2004 and Hawkins and Stevens, 1985) in single word utterances be reconciled with similar segmental studies on vowel recognition from plosive-vowel CVs (e.g. LaRiviere et al, 1975 and Winitz et al, 1972)? Despite liquids not having been included in this study, this issue is very relevant to this research, since in the studies referred to in this thesis (e.g. Hawkins and Nguyen, 2001, 2004), the same kind of FPD at the same place in structure was looked at, exhibiting a similar polysystemic non-segmental distinction. We *cannot* dismiss such findings solely because almost only liquids have been studied from this perspective in previous research. The discussion thus far shows that spreading of features in monosyllables has similar phonetic consequences across a range of stimuli and structures (cf. e.g. Goffman et al, 2008 and Coleman, 1998). The ways in which phonological contrasts are spread in monosyllables should be more significant in vowel recognition and coarticulation than what types of phonetic exponents they comprise.

vi) What would a phonological model capable of accounting for long-domain coarticulation in

> CV(V)/C utterances look like and what does it need to do in order to achieve that? This research question arises jointly from the primary research question and the other secondary topics, as a range of phenomena relevant to vowel recognition have not been investigated or modelled in previous vowel recognition studies.

Next, the hypotheses relevant to the primary and secondary research questions are described.

2.6 Hypotheses

We will now present the main hypotheses arising out of the main research question presented in chapter 1 as well as the literature reviews presented in 2.1-2.3. From the perspective of formulating hypotheses for the secondary research topics, the three areas reviewed in 2.1-2.3 (vowel timing, vowel recognition from plosives and its phonological modelling) are to be considered jointly, since only a synthesis of their results will allow offering a detailed account of how recognition timing works in practice. Hypothesis a) relates to the primary research question, whilst hypotheses b-g) detail hypotheses relevant to the secondary research questions:

a) It is hypothesised that the temporal point at which listeners will achieve correct and reliable vowel recognition from real word aspirated plosive-V(V)/Cs is 30ms subsequent to release (Nearey and Assmann, 1986, Rosner and Pickering, 1994).

b) Acoustic similarity is the most important perceptual criterion in vowel recognition in the way in which listeners select response choices (Ostreicher and Sharf, 1976). When being

uncertain about a given vowel response option, listeners will tend to select a response choice that will be phonetically quite similar to the underlying excised vowel.

c) Carryover and anticipatory coarticulation both have a significant effect on recognition: the phonetic implementation of the onset and the coda in the same syllable may together affect the time course and reliability of vowel recognition (also see Coleman, 1998 and Hawkins & Nguyen, 2001).

d) For onsets, it is hypothesised that alveolar onsets give rise to significantly fewer correct vowel responses than either bilabials or velars (with no significant difference between the latter two), because alveolars have more complex articulation with more demanding co-ordination between the passive and active articulators compared to bilabials and velars. For example, alveolars have a higher amplitude burst at high frequencies (Stevens, 1998, p. 364), as well as a descending F2 transition, rather than a rising one.

e) For vowel length, short vowels are hypothesised to be recognised earlier and more easily than long vowels, because the articulation underlying long vowels engenders a more complex perceptual computation with respect to VISC (cf. Nearey and Assmann, 1986).

f) For vowel height, it is hypothesised that high and especially high front vowels are recognised more reliably than mid and low ones, respectively, because high vowels are easier to recognise early since their less variable VISC patterning engenders a more easily implemented perceptual computation.

g) Increasing presence of nasality during the aperiodic phase (see Cohn, 1990) will distort listener ability to recognise the

vowel correctly, because the presence of additional nasal resonances in CVN monosyllables and potential spectral dampening resulting from nasality will obscure the underlying formant pattern.

Having outlined the hypotheses for all the research questions in this study, we will make some general comments on each of them. The rest of this subsection considers each hypothesis and where they come from.

What is the relationship of hypothesis a) to previous findings? Since some of the results of the previous studies on vowel recognition (2.2) differ from the results of more general studies on vowel timing (see e.g. Cullinan and Tekieli, 1979 and Winitz et al, 1972 contra Rosner and Pickering, 1994 and Nearey and Assmann, 1986), we need to be able to adequately reconcile their results: the previous literature on vowel recognition suggests reliable recognition to be quite possible, and in many cases very likely, from the burst portion alone (see e.g. Winitz et al) and/or very shortly (ca. 10-30) after release (see e.g. Cullinan & Tekieli, 1979). The literature on vowel timing and VISC suggests that vowel formant trajectories start approximating more rapidly towards their final steady-state trajectories at ca. 30ms subsequent to release, sometime *after* the burst. Listeners might therefore not be able to reliably recognise vowels until that 30ms point has been reached.

As the main strands of literature on general properties of vowel timing on the one hand and vowel recognition timing from plosive onsets on the other are in conflict, their findings must be reconciled. The fact that stimulus durations and timing intervals between individual gates differ in previous studies on vowel recognition (cf. e.g. Winitz et al, 1972 and LaRiviere et al contra Cullinan and Tekieli, 1979) makes it difficult to draw consistent hypotheses from them alone. A more general source to explain where the main hypothesis comes from is required.

Hypothesis b) is mainly based on the claims made by

Ostreicher and Sharf (1976), according to which listeners more often choose vowel options that are more similar to the real response choice than ones that are phonetically distant.

Hypothesis c) is based on the claims made by Ostreicher and Sharf (1976) on the significance of both anticipatory and carryover coarticulation in vowel recognition from plosives and on the studies reviewed in 2.2.4 (e.g. Cohn, 1990 and Hawkins and Stevens, 1985), according to which vowel nasalisation in CV(N)s may be delay or distort listener ability to recognise the vowel early on. The findings of e.g. Heid and Hawkins (2000), Hawkins and Slater's (1994) and Hawkins and Nguyen's (2004) on the availability of cues to coda voicing in onset laterals (e.g. 'led' vs. 'let') and coda fricatives (in e.g. 'boozy/doory') are relevant to assessing the significance of coarticulatory direction effects. Cues to both the onset and the vowel may relate to the way in which the phonological feature specification of the coda influences the phonetic exponency of CVCs specifically. In sum, on at least one level of interaction between different systems in language, the temporal processing of vowels is best explained through a polysystemic non-segmental analysis.

Hypothesis d) is based on general coarticulatory dynamics and direction effects (see e.g. Ostreicher and Sharf, 1976, Stevens, 1998 and Modarresi et al, 2004). Some of the previous studies on vowel recognition from plosives are not fully consistent as far as how onset place of articulation can affect vowel recognition (cf. e.g. LaRiviere et al, 1975 contra Cullinan & Tekieli, 1979 and Tekieli & Cullinan, 1979). It is not possible to formulate a clear hypothesis from them.

Hypothesis e) is in agreement with the findings of Rosner and Pickering (1994) and Nearey and Assmann (1986). Based on the results of previous studies on vowel timing and VISC, short vowels should be more reliably recognised than long ones in English varieties spoken in England, as also

suggested by Cullinan and Tekieli (1979) for North American varieties.

As for hypothesis d), hypothesis f) does not allow us to draw a fully consistent hypothesis vis-à-vis the previous literature. Only Cullinan & Tekieli (1979) make any claims on height and frontness. It seems premature to draw any conclusions for such vowel features based on a single study. In the formulation of hypothesis e), this study relies on claims made in studies on general coarticulatory dynamics, such as Gussenhoven's (2007) claim on low vowels taking on average longer to recognise and produce than high ones, and on VISC studies. Since low vowels are more variable in VISC than high ones in terms of the variability of formant centre frequencies from their average values (cf. Rosner and Pickering, 1994), frontness should have less coarticulatory influence on vowel recognition than height. It is more difficult to make straightforward predictions on frontness, since so few of the relevant studies make any claims on it as a feature.

As far as hypothesis g) is concerned, Hawkins and Stevens (1985), Cohn (1990) and Chang et al's (2011) studies show that there may be significant nasality present during the aperiodic phase in CVNs and that nasalised vowels are on average harder to recognise than oral ones in CVs, the resulting hypothesis g) is warranted. It is not possible to state how hypothesis g) relates to the previous literature, since the influence of coda exponency on vowel recognition was not studied in previous research concerning recognition from plosives.

Having detailed the research questions and the main hypotheses underlying the perception experiment and main research question on which this study is based, the methodology is described next.

3. Methodology

3.1. Overview

In this chapter, the vowel perception experiment is described and justified. First, key aspects of the experimental design are justified and described (3.2 and 3.3). Next, the various aspects of the sample and materials are detailed (3.4-3.5). The last section contains a review of the analysis methods (3.6).

3.2 Experimental Design and Rationale

A gating task experiment was used to assess the research questions. In the gating paradigm, participants are asked to deduce the quality or properties of subsequent linguistic constituents based on what they have heard so far (see e.g. Grosjean, 1996 for a review). For instance, given the example of "I think it's s-' ", listeners would probably expect to 'hear' a word whose initial sounds are /s/ + a vowel (e.g. 'sitting', 'sad' or 'Sara'). A similar experiment assessing semantic priming would be one where the listener hears 'I think you s-', containing only part of the [s] segment. This final segment would probably suggest that a verb beginning with the sound /s/ follows 'you' (a noun, proper noun or an adjective could not follow 'you'). The next section explains the reasons why the gating paradigm was used in this research.

3.2.1 Justifications for Choosing the Gating Paradigm

There are several different variants of the gating paradigm within linguistic research, some looking at semantic or syntactic factors (see e.g. Tyler & Wessels, 1985) and stimulus

properties, whereas others have assessed phonetic and phonological properties of subsequent words, sounds or phrases (see e.g. and Winitz et al, 1972 and Ranbom & Connine, 2007).

We will also show in this section in what ways the primary and secondary research questions help in driving the thesis forward and interpreting the results, and in particular how that applies to the methods of this research.

Since the gating paradigm delivers stimuli of varying duration, it makes it possible to assess both the temporal and phonetic properties for several time slots (e.g. containing 10, 20, 30 and 40ms of aspiration) and any accompanying FPD directly. For example, let us assume a sentential utterance such as "I think you c*an*' " is gated so that resulting stimuli contain various durations (e.g. 20, 40 and 60ms) of the phonetic exponents of the onset in 'can' (= /k/). The FPD contained at the end portions of such an utterance may give listeners cues to following sounds due to the overlaying of distinctive sounds in the real-time phonetic output (see e.g. Marslen-Wilson and Tyler, 1980, Grosjean, 1996, Shockey, 2003 and Coleman, 1998). The gating paradigm is a good method to use in the context of tests assessing the way temporal properties of speech sounds are perceived. The design and the way the listeners are stratified (see 3.3.2) enables assessing how different types of phonetic detail allow listeners to process vowel information and update their perceptions through time.

Grosjean (1996, p. 601) shows that the gating paradigm is easy to use and running participants can be done using little additional equipment, although stimulus preparation may take some time (if not automated). The gating paradigm is probably the most practical and also the theoretically most applicable tool available for a vowel recognition experiment, unlike for example, eye-tracking (e.g. Duchowski, 2007), brain-imaging (e.g. Shulman et al, 2004) and similar paradigms, which would

require additional equipment and resources. Such experimental procedures would complicate the interpretation of the results and do not allow for as good theoretical coverage of perception timing as the gating paradigm (cf. Grosjean, 1996). Eye-tracking and similar procedures reflect qualitatively different and either narrower or more general aspects of perception (e.g. visual attention and/or psychophysical aspects of perception), which are not that relevant in this study. Such paradigms are also to some degree invasive, in that they necessitate applying equipment onto participants. Using such experiments would make it harder to generalise results to online processing as the listening situation would be more unnatural and potentially cause discomfort for participants.

The gating paradigm allows exercising precise control over what and how much acoustic–phonetic information is presented to listeners, since different types of stimuli with different durations can be prepared, and in different linguistic contexts (e.g. assessing semantic, syntactic or phonetic priming). Gating can indicate the required amount of acoustic–phonetic information to identify a stimulus quite precisely. The gating paradigm allows us to investigate several different kinds of dependent variables, which makes it applicable in many areas of linguistic research (see e.g. Grosjean, 1996). The gating paradigm is an ideal choice from the viewpoint of asking several related research questions, such as on vowel recognition timing and syllable structure.

The gating paradigm can be considered a powerful experimental tool, especially if it is possible to show that the stimulus candidates proposed reflect what goes on in the mind during listening (Grosjean, 1996). Since listeners must be capable of comparing heard auditory information with stored representations (e.g. Clopper & Pisoni, 2004 and Moore, 2008), it seems very likely that the gating paradigm reflects at least the most basic properties of on-line processing, rather than only

post-lexical processes (e.g. Caramazza, 1997). This claim is also in agreement with the fact that speakers structure the phonetic detail of their utterances to include specific combinations of linguistic properties in context (e.g. Foulkes and Docherty, 2006, pp. 415 and 432).

As far as the general design of my experiment is concerned, the use of the gating paradigm makes it possible to answer the main research question in detail. In sum, the research questions asked in this thesis on the timing of vowel recognition and secondary phenomena associated with it help to drive the rest of the thesis forward, as follows:

a) The primary research question is driven forward by the implementation of the gating process, since a narrow 10ms window for temporal incrementation has been chosen. The fact that this choice is consistent with previous perception studies (e.g. Cullinan and Tekieli, 1979) as well as the kind of phonetic variation that typifies VISC (see e.g. Rosner and Pickering, 1994 and Nearey and Assmann, 1986) are two other key issues that shape this choice. On the one hand, the temporal magnitude of the interval between the chosen time slots conforms to choices made in previous research, which also reflects the type of temporal variation that typifies spectral changes in VISC.

b) The FPD associated with the temporal incrementation of phonetic information and especially its theoretical precision may allow to quite precisely estimate how different types of coarticulatory strategies and accompanying FPD can affect recognition. This issue is particularly relevant for answering the secondary research question on vowel confusions.

c) The stimulus structures chosen and especially their phonological and phonetic shapes make it possible to assess whether long-domain coarticulation may affect recognition, since a) codas chosen have different types of articulations (voiceless plosives vs. voiced alveolar nasals) and airflow and b) since the logic behind the polysystemic approach emphasises the availability of cues to different sounds from the type of subtle FPD associated with different structures (rather than vice versa, as in FPA): "in perception, a reasonable hypothesis is that if the sounds in two utterances differ, then one or more things in their structures differ" (Hawkins, 2010a, p. 485). One key implication of this claim by Hawkins with respect to this thesis is that each constituent in a CV(V)/C can have some effect on the acoustics and perception of its other constituents.

Having demonstrated and exemplified the experiment applied in this research and the justifications for its application, the various aspects of the experimental design will be described

3.3 Participants

3.3.1 Speakers

Two young male and two young female speakers were recorded for the experiment. The division of speakers was performed according to the criteria of accent and gender so that each speaker had a different sociolinguistic background to every other speaker. All the four speakers (two males and two females) recruited were between 19-25 years of age. The northern male speaker is from Burnley, Lancashire, whereas the northern female is from Lincolnshire. The two southern speakers are from Bristol (the male speaker) and Maidstone,

Kent (the female speaker). Choosing two male speakers and two female speakers (one of each with five and another with six short stressed vowels) may give a better representation of vowel perception timing than having a random set of speakers, or just having one speaker (as in most previous studies on the timing of vowel recognition in English). Since the whole of the English speaking population in England is being targeted in linguistic terms, it was deemed necessary to have male and female speakers from several different parts of the country.

3.3.2 Listeners

The listeners were 24 18-25 year old native speakers of English brought up in England. There was an equal number of southern and northern listeners (southern ones with six short vowels and northern ones with five). The listeners participating in the experiment were stratified as follows:

Participants listening to a southern speaker:
Three male and three female southern listeners hearing a southern male or female speaker
Three male and three female northern listeners hearing a southern male or female speaker

Participants listening to a northern speaker:
Three male and three female southern listeners hearing a northern male or female speaker
Three male and three female northern listeners hearing a northern male or female speaker

All of the 4 listener groups had at least 2 southern or northern speakers, with two of the groups having 3 southern vs. 3 northern listeners, and 2 vs. 4 in the other groups.

At the initial stage, one of the aims of the experiment

was to look at the potential role of accent, in particular with respect to social and regional differences in vowel realisation, e.g. concerning /u:/-fronting and the number of short vowels in a speaker's vowel system (/ɪ a ɛ ɒ ʊ/ for northern speakers and /ɪ a ɛ ɒ ʊ ʌ/ for southern ones). The results of an earlier version of the experiment proved inconclusive in this respect. All additional questions concerning vowel exponency in different regional accents were put aside, including that for /ʊ ʌ/ in northern and southern accents. This secondary topic did not constitute part of the revised experiment. The revised stratification makes it possible to combine all 24 listeners into one data set, treating the four different mini-experiments as one large set.

24 listeners were deemed sufficient for statistical analysis, since previous studies on vowel recognition from plosives have used 10-20 listeners. All these criteria for stratification are based on a) being able to compute reliable statistical tests on the data and b) the results being as representative of the English population as possible. They are not e.g. meant to represent different sociolinguistic attributes, since such properties are not particularly relevant for the primary research question in this work. Next, the main approach used in participant recruitment will be briefly described.

3.3.3 Method of Recruiting Participants

The listeners were approached via e-mail asking for their assistance in a perception experiment. Once listeners had confirmed their interest, they were asked to read an e-mail before doing the test itself (see appendix) and comply with its instructions (see appendix).

3.4 Materials

This subsection will describe the methods and choices having to do with the stimulus materials.

3.4.1 Stimuli and Stimulus Structure

The stimuli comprised 60 monosyllabic CVV, CV + {-p, t, k/} and CV + /n/ British English minimal pair lexemes. Table 6 shows each stimulus category (left-to-right). The left-hand column details each vowel quality used where the top horizontal rows describe onset and coda quality. The stimuli with nasal coda (red cells on the bottom right) and CVV stimuli (green cells on the bottom left) are detailed in the bottom part of table 6. Common CV(V)/C words were preferred, however two rare words (‘cuck’ (=cuckold) and ‘cun’ (= a Chinese measure of length), were also included, in order to have a complete set of 60 stimuli (the inclusion of these words will be discussed shortly).

	O	**C** [p]	**O**		**C** [t]	**O**		**C** [k]	
V	[p^h]	[t^h]	[k^h]	[p^h]	[t^h]	[k^h]	[p^h]	[t^h]	[k^h]
[ɪ]	pip	tip	kip	pit	tit	kit	pick	tick	kick
[a]	pap	tap	cap	pat	tat	cat	pack	tack	cack
[ʌ ʊ]	pup	tup	cup	putt	tut	cut	puck	tuck	cuck
[ɒ]	pop	top	cop	pot	tot	cot	pock	tock	cock
								C [n]	
[i:]	pea	tea	key			[ɪ]	pin	tin	kin
[u:]	poo	two	coo			[a]	pan	tan	can
[ɔ:]	paw	tore	core			[ʌ ʊ]	pun	ton	cun
[ɑ]	par	tar	car			[ɛ]	pen	ten	ken

Table 6: Word stimuli used in the gating experiment

High frequency sounds were used, because they give a more representative picture of recognition than using infrequent sounds (e.g. coda /ʒ v f/ or the vowel /ɜ:/, see Fry, 1947). Plosives and nasal consonants coarticulate more extensively with vowels than, for example, fricatives and glides (e.g. Hardcastle & Hewlett, 1999). This aspect influenced the choice of stimuli in the sense that plosives and vowels are much more frequent than other kinds of sounds (Moore, 2008) and most languages have more than one of each of this type of sound. Last, monophthongs were chosen rather than diphthongs as

monophthongs are more frequent and have less complex formant movements than diphthongs (e.g. Nearey and Assmann, 1986 and Rosner and Pickering, 1994). Interpreting results on diphthongs would have been more complicated, especially in terms of assessing the perceptual significance of FPD and the direction of spectro-temporal changes in VISC. The reason for choosing the particular monophthongs [i ɪ ɛ a u: ɔ: ʌ ʊ ɑ:] is based on choices made in previous studies, such as the ones by Tekieli and Cullinan (1979) and Cullinan and Tekieli (1979). Having a set of eight vowels allows maximal coverage of vowels according to frontness, height and rounding. Phoneme frequency was also taken into account, which necessitated avoiding using /ɜ:/. Since this vowel is comparatively rare in English (see e.g. Fry, 1947 and Cruttenden, 2014, p. 156), it was deemed not to represent categorisation as strongly as other vowels. On the other hand, it is acknowledged here that frequency (whether for vowels, consonants or lexemes) can depend on the level of analysis in an experiment. For example, /ɜ:/ might be more common in certain kinds of words than others and it is acoustically very similar to /ə/ in English (which is the most frequent vowel, see Fry, 1947). However, since /ə/ does not occur in stressed syllables (e.g. Cruttenden, 2014) and /ɜ:/ does not occur in many common word forms such as determiners, prepositions and other function words, it was not included in this study.

The carrier phrase “I think you say... X (= X representing each stimulus) was used in order to control for word frequency effects (see e.g. Grosjean, 1980) and so as to make the word stimulus unpredictable from a linguistic-phonetic viewpoint (or otherwise).

Since this study examines issues concerning

phonological processing in CV(V)/Cs, there is a need to ensure that the only cue to the identity of each monosyllable was based on phonetic detail rather than, for example, semantic, syntactic or pragmatic cues.

With regards to the choice of words forms, although ‘cun’ and ‘cuck’ are rare and for many speakers are probably nonsense words, excluding their use in the experiment might have introduced other complications: for example, for certain stimuli participants would have been faced with only three options rather than four, which might have had distorting effects on perception and would have ruined the consistency of the 4-way forced choice method. Such a choice would thus have served to sacrifice the consistency of the experiment based on only 2 words out of 60 (= ca. 3.5% of the stimulus set). It is acknowledged that having even just two rare words or words of different classes (cf. e.g. the verb form ‘tore’) in an experiment like this one is not ideal, since frequency balance could be a confusing variable. However, when working with real words such compromises are typically made. For example, in an experiment looking at lexical access/representation, Munson (2007, p. 209) includes high frequency words such as ‘pot’, ‘top’, ‘put’ and ‘get’ but also ‘pep’, ‘fad’, ‘nape’ and ‘dab’, which are much less frequent than e.g. ‘put’ and ‘top’,. Munson’s study contains 4 relatively infrequent word forms in 80 experimental stimuli (= 5% of all words), where the gating experiment in this study includes 2 rare/nonsense words in 60 stimuli (about 3.5%). Hawkins & Slater (1994, p. 48) include nonsense word forms such as ‘boozy and doory’ in a study on C-to-V and V-to-V coarticulation, which are meaningless. This claim is not true for ‘cun’ and ‘cuck, albeit that they are rare word forms. The stimulus choices in this study conform relatively well with previous research in this respect. This research is not the only recent study to have investigated this kind of a topic using rare words, despite the fact it is known

that word frequency may be an issue in recognition. From the perspective of recent research into vowels, certain limitations in this respect may thus have to be accepted. For the types of research questions asked in this kind of a study, word forms which may be rare or even meaningless may need to be included if sufficiently comprehensive generalisations about e.g. phonological processing are to be drawn.

What matters most in this study is recognition of vowels as parts of real meaningful words, not just whole words which happen to contain a given vowel. The primary research question on vowel recognition is aimed at testing e.g. /p + V/ in "puck/pun/putt/pup" (etc.) rather than the recognition of the word frame from which the vowel is lifted. The words can be seen to provide a meaningful frame of reference for recognising the plosive and especially the vowel rather than the word forms themselves. In sum, the stimulus choices are optimal for the research questions asked.

3.5 Implementing the Design and Experiment

3.5.1 Recordings

A set of four recordings was made in a recording room under quiet conditions at the Department of Language and Linguistic Science at the University of York.

During the four recordings, the experimenter sat opposite each speaker while pronouncing the utterances listed in a random order on two A4 sheets (see appendices): the random ordering of the stimuli is based on avoiding the types of order effects discussed in 2.1.4 as far as possible. The speakers sat comfortably in a chair at a table while reading the stimulus utterances from the A4 sheets during the recording. Each speaker took ca. 30 seconds to practise uttering the stimuli before commencing the recording, while setting up the

level of audio (a technician sat behind a wall facing each participant controlling the audio recording process).

The speakers were sitting still ca. 15-20cm in from of the microphone once the recording had been started. The distance from the microphone was not precisely controlled for, so as not to constrain speakers' articulatory freedom. Had the speakers needed to sit completely tight at a fixed distance from the microphone, it could have interfered with the very research questions, since stimulus production would not have been as natural as possible. Since the intensity levels of plosive bursts vary for bilabials contra alveolars and velars (with bilabials having more intensive burst portions), it becomes exceedingly hard to control for every aspect in this type of an experimental study.

Where errors, hesitations or other disruptions occurred (such as coughs, commotion or mispronunciations), the stimuli were re-recorded after having read out the list of 60 randomly ordered words, arranged as follows (see the appendices for the complete list):

'I think you say' + the beginning portion of each word stimulus (cf. table 6)

The speakers were asked to read a set of instructions before each recording was initiated (see appendices). Each recording session lasted ca. three to seven minutes, depending on the number of mistakes that occurred. Speakers were free to take as much time they wished in articulating each word stimulus, in order to make as natural sounding stimuli as possible.

3.5.2 Stimulus Segmentation

The stimuli were gated at the nearest zero crossings 10, 20, 30 and 40ms into the aperiodic phase of the onset using Praat.

Zero-Crossing refers to a point in the speech wave at which it crosses the abscissa. Figure 20 displays the word ‘paw’ produced by the southern female speaker and (both a spectrogram and waveform are included):

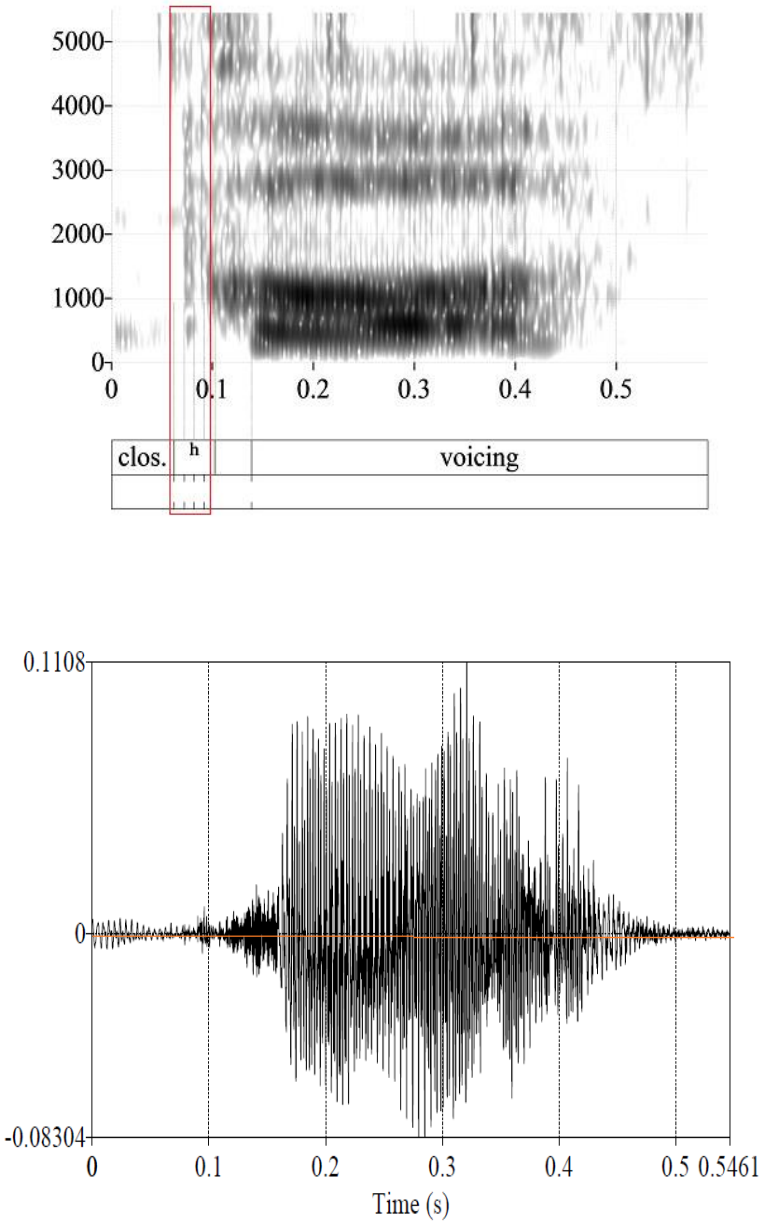


Figure 20: Abscissa in the word ‘paw’ (southern female speaker)

The stimuli were gated at the nearest points in the sound wave where the aperiodic speech signal crossed the area depicted by the orange line in the bottom part of figure 20 (containing a

waveform of the word ‘paw’. The top part of figure 20 illustrates the gate interval points in the same word (cf. the part of the point tier at the interval marked ‘h’).

This method allowed controlling for any click-like effects that might have arisen due to amplitude fluctuations arising from cut-offs at different points within the sound wave. The main reason for segmenting the stimuli at zero crossings and having 10ms gate intervals is based on preventing artefacts coming into the recording. Johnson (2011, p. 49) affirms that in the analogue-to-digital conversion of speech sound waves, we need to sample the speech signal often enough so that we capture all the spectral information we wish to study. For example, when investigating periodicity in a sine wave that repeats its cycle 100 times per second (100 Hz), we need at least two samples per cycle in order to capture its periodicity. However, since the amplitude and the phasing of a speech sound may vary independently of the periodicity (which reflects the vibration of the vocal folds), we would need a lot more samples to determine these properties, since they may vary more randomly through spectro-temporal space (Johnson, 2011). A consequence of this acoustic property relating to amplitude and phasing is that when we gate a stimulus at the point of the abscissa we can control for amplitude fluctuations arising from cut-offs at distinctive points within a sound wave.

Having 5ms or 20ms intervals would not have been ideal from a perceptual viewpoint, since 5ms would probably have given listeners very subtle distinctions to draw in spectro temporal terms, with the opposite applying to 20ms gates. Such choices would have made for either a too long or too short experiment, potentially making any results less reliable, since listener fatigue or lack of motivation might have affected the quality of findings negatively.

Where there were multiple releases, as typically is the case with velar plosives (Stevens, 1998, p. 330) the first release

was chosen as the point at which the stimulus would be excised, because this temporal point is perceptually significant. Otherwise, listeners would have heard multiple releases and differing transitions (leading to changes in resonance fluctuations). Since these kinds of changes are the kinds of alterations considered as changes in FPD in this thesis, the reason for choosing the first release as the point of excision is theoretically significant.

A random ordering for the stimuli was generated in Excel 2007 using the RAND command. The motivation behind this choice was to control for confounding effects having to do with episodic recognition memory, including recency effects.

The stimuli were ordered manually into an online survey in *SurveyGizmo,* where a set of four experiments was set up, one for each speaker type. These four experiments can be treated as one complete experiment on vowel recognition.

The ordering of the stimuli was carefully checked, in order to make sure that each and every interval/stimulus had been included *once* in each of the 24 individual experiments listeners took part in. Since it was possible to make exact copies of previously set-up experiments in *SurveyGizmo,* only the sound files rather than the questions themselves needed to be uploaded again: this aspect facilitated the analysis and verification processes (see 3.6).

The number of stimuli for all speakers was the same in all 24 individual tests heard by listeners. There were 240 word stimuli, 48 of which were CVVs and 192 CVCs. The 192 CVCs each contained 48 CVp/CVt/CVk and CVN stimuli.

3.5.3 Stimulus Presentation and Procedure

This subsection describes the general process each listener went through in responding to the stimuli. Having answered a range of questions about their socio-economic background (including

factors such as age and gender), the experiment itself began: listeners heard each stimulus at once when opening a new page, and were asked to make a choice between four words by responding to the stimuli as 4-way forced choice between (e.g.) ‘key-coo-core-car’ and ‘pit-pat-pot-putt’ (cf. each column in blue, red and green in table 6). Although the stimulus choices as a whole are similar to those in Cullinan and Tekieli (1979), listeners had only 4 (rather than 8) options to choose from for each stimulus. Each word stimulus was embedded at the end of the carrier phrase "I think you say… X" (‘X’ represents each CV(V)/C word form).

Listeners were allowed to listen several times (if needed) to each stimulus by pressing the play icon in each sound file. As it was not possible to control for external variables (such as background noise), it was necessary to ensure that each sound file could be repeated if necessary: the motivation for multiple hearings (if needed) outweighs any counterarguments since it must be ensured that listeners are able to clearly hear each stimulus at least once, even though headphones were used. For example, considering that the listeners might have been faced with situations where they keep sensing background noises beyond their own control at semi-regular/regular intervals (e.g. banging and/or reverberation), this choice can be deemed necessary. Although this slightly complicates the interpretation of the results, it is impossible to control for everything in an external online experiment, rendering multiple listenings of stimuli necessary.

Regardless of potential recency effects associated with multiple listenings to experimental speech stimuli, it had to be ensured that listeners would 1) not be able to return to previous answers after having selected a particular option and 2) that they would be as likely as possible to hear all the stimuli adequately. This choice makes the experiment more feasible and worthwhile to participate in, while increasing listeners’

motivation. It is recognised that this methodology is not ideal. Nevertheless, considering the theoretical and especially the methodological constraints behind on-line experiments, the choices are justified. Having assessed and described the features of the gating experiment, the analysis methods applied to it are described in the next subsection.

3.6 Analysis Method

3.6.1 General Phonetic Aspects of Plosives and the Aperiodic Phase

The three characteristic phases of plosives are described in this subsection before giving more details about aspiration and the transitions into vowel sounds during the aperiodic phase of voiceless plosives in the following section.

Plosive consonants normally consist of three distinct articulatory phases: these are normally known as the closing phase, hold phase and release phase, respectively. During the closing phase, the different articulators (e.g. the tongue and alveolar ridge) come together and begin creating a momentary occlusion in the mouth (or at the lips for bilabial plosives). Somewhat later, air is compressed behind the oral constriction during the hold phase. The duration of the hold phase can be somewhat longer than the closing phase, mainly depending on place of articulation (with bilabials having longer closures than velars and alveolars, Stevens, 1998, p. 346 and Port and Dalby, 1982, p. 143-145). Lastly, during the aperiodic phase the articulators forming the airtight closure come apart and the compressed air is released with plosion (Collins and Mees, 2003, p. 79). The release phase can last up to about 120-150ms (see e.g. Docherty, 1992, p. 113-114). Now is a good time to discuss the general phonetic aspects of the aperiodic phase in more detail. Much of the discussion is focused on aspiration

rather than the release burst and/or potential accompanying frication.

Stevens (1998, p. 324) shows that one of the main consequences of the release in a plosive is to produce in some frequency regions an abrupt increase in amplitude and to generate appropriate distinctive spectral changes. This type of acoustic signature largely depends on the articulatory structure that is used to form the constriction.

The most common sources of sound during the aperiodic phase are the release burst (which may be followed by a period of frication) and aspiration. The release burst is produced when increased air pressure is abruptly released. Air streams out of the mouth at a rapid velocity, producing a brief pressure impulse. The burst noise marks the moment of release. The release phase can be voiced and/or aspirated (Johnson, 1997, p. 131). English does not have voiced aspirated plosives unlike many languages of the Indian subcontinent (e.g. Sindhi and Gujarati).

The two sound sources in aspirated plosives must be distinguished (Johnson, 1997, p. 131-132). The burst noise is produced at the consonantal structure, whereas aspiration is produced at the glottis (so that voiceless aspiration has the arytenoid cartilages quite widely separated). The vocal tract filter for these sound sources is distinctive. The burst is usually very short in duration (see e.g. Stevens, 1998), whereas the aspiration can last up to ca. 150ms. For a few milliseconds following closure, the constriction is too narrow to allow the amount of airflow needed to produce aspiration. Immediately following release conditions are set right for the burst (i.e. a high pressure build-up combined with a very narrow opening) but not for aspiration. Subsequent to this period, conditions are set right for aspiration, when sufficient airflow can be generated due to the somewhat more open constriction.

3.6.2 Plosive-Vowel Transitions in Unaspirated Plosive-Vowel CVs and in Aperiodic Noise

In this section, it is demonstrated how plosive-vowel transitions work in English: the purpose of this part is to substantiate the measurements of acoustic formant data presented in the following chapter. The main aim is to show that since all the data measurements were taken during the early part of aperiodic friction at the transition point between a plosive and a vowel, they are very highly unlikely to exhibit formant centre frequency values that approximate "normal" formant characteristics (see e.g. figures 4-6). This acoustic-perceptual distinction may become much more marked in aspirated plosives than in unaspirated ones, where formant transitions will approach their final trajectories very rapidly (e.g. Stevens, 1998, p. 340-375). Very early on subsequent to plosive release, the approach of F1/F2/F3 towards their steady-state values will be more gradual in aspirated plosives than in unaspirated ones. Although there is virtually no comparable research for this on English, it is convenient to illustrate the differences in e.g. the word 'pea' produced with an aspirated plosive as against a CV with an unaspirated one.

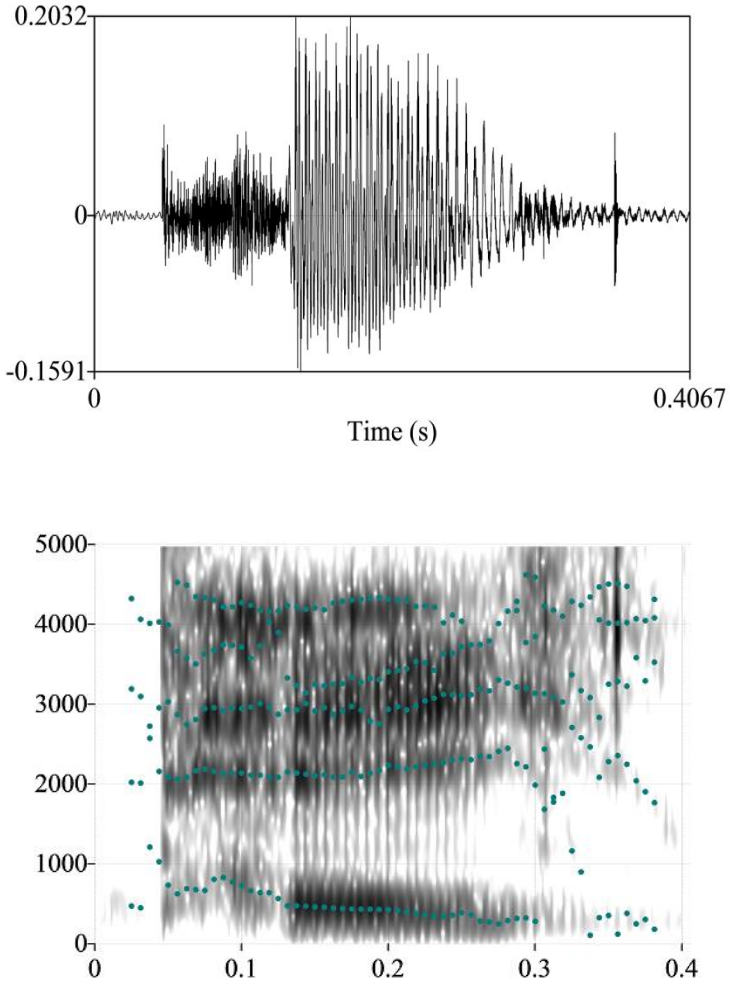


Figure 21: Changes in formant frequencies in the word ‘pea’ as produced by a northern male speaker (waveform and spectrogram)

The measurement methods and analysis options applied to the spectrograms in figure 21 are the same as for those in figures 8-9, along with data from the same northern male speaker.

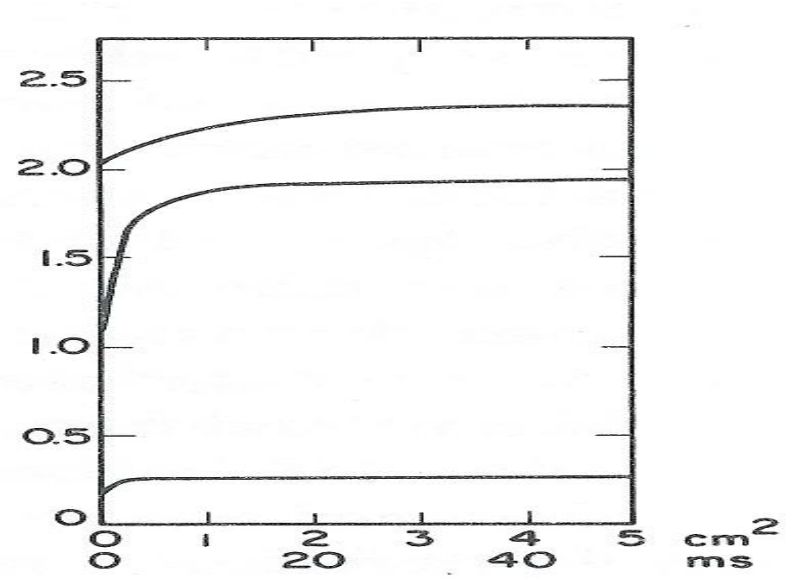


Figure 22: Simulated calculations of changes in formant frequencies in [pi] (adapted from Stevens, 1998, p. 341, Figure 7.14)

Although the formant frequency changes for [pi] in figure 22 are based on simulated tube model calculations by Stevens (1998), they will closely approximate real values for English unaspirated bilabial plosives (as in e.g. 'spit' or 'spar'). Formant traces from natural speech will never look quite like those in figure 22, since formant tracking is an approximation of formants' true values. The important point about this distinction between tube model predictions of formant movements and measuring techniques is that the way we estimate formant values using conventional techniques gives imperfect estimates of their formant centre frequencies. Despite this contradiction, Stevens' predictions are reliable and accurate.

When observing the realisations of the formant transition from the plosive to the vowel in [pi] in figure 22, the transitions are seen to reach their trajectories very quickly after plosive release, i.e. within ca. 20ms for both F2 and F3 between ca. 0 and 20ms (cf. the left hand side of the x-axis). For the aspirated variant in 'pea' in figure 21, the transitions are not

fully complete until ca. 110 ms subsequent to the burst (cf. figure 21 at ca. 0 contra 0.11 seconds, respectively). The data in figures 21-22 give validity to Stevens' (1998, p. 464-465) claim about transitions from a plosive to a vowel in aspirated plosives to be quite similar (but somewhat temporally distinct) to those in unaspirated ones. This distinction is explored and substantiated further in the next subsection.

3.6.3 On the Phonetic Properties of Aspiration and Accompanying Formant Transitions into Vowels

Aspiration is characterised by friction generated at a random source at the glottis. The most typical acoustic constituents of aspiration are F2, F3 and F4. According to Fant (1973, p. 113), the aspiration in bilabials is usually less prominent, due to the fact that the cross-sectional area of the constriction for closures formed at the lips has a lower noise generating efficiency. This phonetic property means that aspiration is likely to be of shorter duration in English [p^h] than in [t^h] and [k^h]. F1 may not be well defined during aspiration due to the acoustic losses associated with the damping of F1 (Stevens, 1998, p. 171). The relatively large (ca. 0.2cm²) cross-sectional area of the glottal opening is sufficient to cause such losses during aspiration. Constrictions in the vocal tract will lower F1, since the jaw position is more closed and constrictions will inhibit the formation of a clear resonance peak for F1 (e.g. Stevens, 1998).

As briefly noted in the previous subsection, virtually all previous research on this topic is on unaspirated plosives, which will allow measuring spectral peaks most easily (see e.g. Clements and Osu, 2002, p. 338). This property depends on the fact that whereas in unaspirated plosives the change from plosion to voiced periodic open approximation is swift with vocal fold vibration for the vowel being switched on almost

instantaneously, in aspirated plosives the vowel resonances have to pass through a period of voiceless aperiodic friction. Since aspirated plosives were studied in this research, it is important to establish more clearly how similar formant transitions during the aperiodic phase will be to those in unaspirated plosives. The following point by Stevens (1998, p. 464-65) is important here:

"The spectrum of the aspiration noise following the frication burst contains peaks corresponding to F2 and higher formants, which show place-dependent transitions in frequency similar to those described in chapter 7 for unaspirated stops. Just prior to the onset of glottal vibration following an aspirated stop, the formant transitions are almost completed..."

Stevens (1998, p. 464-465)

The key analytical issue here is that despite being similar and occupying a similar frequency range, the formant transitions in aspirated plosives will reach their target values at the point where voicing for the upcoming vowel is switched on more slowly (Stevens, 1998, p. 465). In aspirated plosives, the formant transitions reach their steady-state trajectories somewhat later and more gradually than in unaspirated plosives: this point helps to correctly interpret the formant data measurements displayed in chapter 4 (subsection 4.2). However, it is also important to observe that Stevens (1998: 464-65) makes no mention of F1 for aspirated plosives, since it is variably present and our estimates of the first formant are at best approximations, as mentioned in the previous subsection.

In summary, the values received at any given point subsequent to release (and prior to glottal vibration) in an aspirated plosive may be more deviant from steady-state formant values than for unaspirated plosives. Figure 23

illustrates this type of pattern for an unaspirated alveolar plosive-vowel transition:

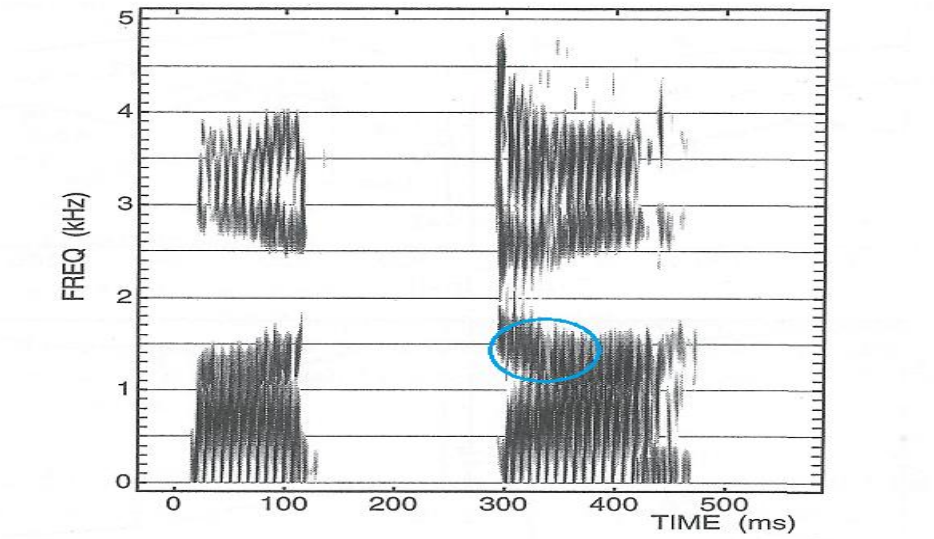


Figure 23: F2 transition in the disyllable [ɑtɑ] subsequent to the plosive release
adapted from Stevens, 1998, p. 363

When inspecting the second formant transition on the x-axis at ca. 300-350ms and investigating its trajectory, F2 has a value of approximately 1500-1700 Hz (cf. the blue circle in the middle of figure 23), whilst even 50ms subsequent to release, the centre frequency has descended by only ca. 200 Hz. Such a value would make /ɑ/ seem more like a front or central rather than a back vowel. The seemingly odd value of F2 can be explained by the fact that at this point in the formant transitions, the formants are not yet sufficiently close to the vowel steady-state portion in acoustic or temporal terms for the observed values to be considered to represent 'normal' vowel formant centre frequencies. Different places of articulation in plosives give rise to different movement velocities with respect to jaw opening (Stevens, 1998, p. 326). Velars lead to slower transitions into vowels, with increasingly more rapid transitions from alveolars and bilabials.

Certain complications need to be borne in mind. For example, individual speakers' productions will yield slightly different values and data for other vowel contexts will have different centre frequencies (Stevens, 1998, p. 465). It may be difficult to give precise estimates of formant data and values from aspiration, due to the lack of a clear formant structure and periodicity during voiceless aperiodic friction (see e.g. Clements and Osu, 2002, p. 338), and the fact that F1 may not be very evident acoustically during this period (Fant, 1973, p. 130), since constrictions in the vocal tract tend to inhibit F1. The amplitude of F1 will therefore be much lower than normally, for example (Stevens, 1998). Having described the properties of formant transitions in plosive-vowel CVs during aspiration and taken note of difficulties in offering precise estimates from them, it is time to detail how the measurements performed.

3.6.4 Spectro-Temporal Analysis of Production Timing

All figures containing waveforms and spectrograms presented in this thesis were taken in Praat. A Gaussian analysis window with a length of 5 ms and a 2ms time step was applied. For the southern female speaker, it was found that a maximum formant frequency of 5500 Hz with five poles gave a better match with the standard Formant (burg) method than applying 5000 Hz as the maximum with five poles (which was used with the other three speakers' stimuli). For certain stimuli produced by the two male speakers, it was found that 4500Hz with four poles gave better estimates of F1 and F2 in particular than applying 5000Hz as the maximum value with five poles.

The standard formant estimation method was applied in analysing the resonance peaks. Formant (burg) may provide a better linear estimation of the formant values than the (hack > keep all) method, which preserves all poles whatever their

values (Boersma and Wennink, 2010). The formant traces visible in green in the spectrograms in this thesis were drawn using the 'speckle' function (with a dynamic range of 40dB).

In order to link the FPD of production in the CV(V)/Cs monosyllabic word forms studied in this research with the perception results detailed in the latter half of chapter 3, a spectro-temporal analysis of each voiceless plosive's aperiodic phase was performed. A Praat script designed by Daniel McCloy (University of Washington) was used to list the centre frequency values of F1, F2 and F3 at 10, 20, 30 and 40ms subsequent to plosive release. The script semi-automates measuring formants from sound files with labelled textgrids.
The chosen script cycles through a given directory of textgrids and finds the associated sound files. The files were opened one at a time which then displayed a table of formant values for the intervals delineated in each textgrid file at the pre-specified time points: the script used prompts the user to either (1) accept the formant measurements, (2) adjust the formant settings and recalculate, or (3) mark the interval as unmeasurable. After this had been done, the process was repeated for the next interval or file.

An interval labelled as 'v' was inserted in each textgrid accompanying each word stimulus sound file to accompany the analysis for each textgrid. Since 4 intervals had been chosen to each be separated by 10ms, the values assigned to the script were '25% - 50% - 75% and end', each denoting the four gate intervals being examined by the script. This specification of the time points at which formant values were to be taken resulted in accepting the formant measurements or adjusting the maximum frequency and recalculating, giving estimates that better matched the spectrogram being displayed on the screen for each word stimulus' 40ms aperiodic phase interval. There were no clearly unmeasurable instances in the produced stimuli, however.

Figures 24-25 illustrate the way in which the measurements were taken using Daniel McCloy's Praat script (cf. the web link https://github.com/drammock/praat-semiauto/blob/master/SemiAutoFormantExtractor.praat):

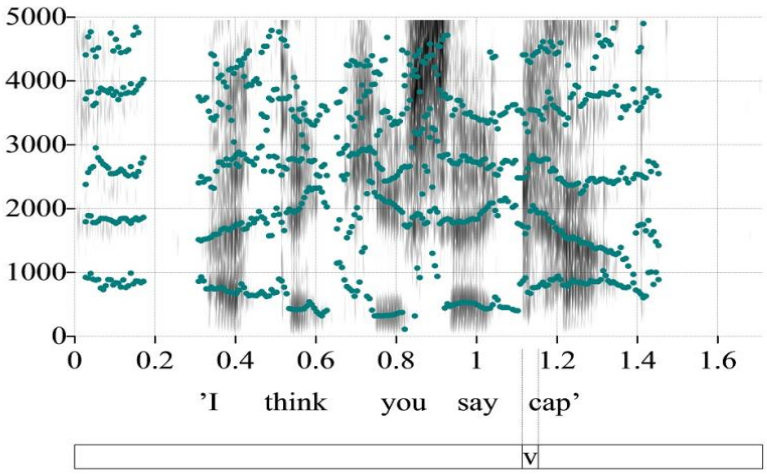
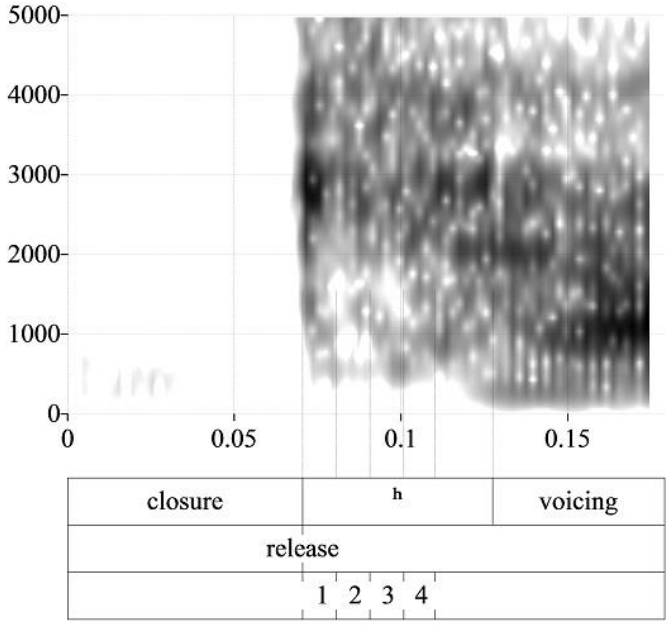


Figure 24: A spectrogram of "I think you say cap"

The interval marked 'v' at the bottom of figure 24 corresponds to the interval marked 'h' towards the lower middle part of figure 25. The numbers '1 2 3 4' displayed at the beginning of the h interval at the bottom of figure 25 denote the 10, 20, 30 and 40ms gate intervals.



[14]

Figure 25: A spectrogram of the '[k^ha]' portion in "I think you say cap".

Figure 24 shows a spectrogram of the utterance 'I think you say cap' produced by the southern female speaker. The 'v' segment around 1.1 seconds in figure 24 and the one marked with 'h' at ca. 0.09-0.11 seconds in figure 25 display the borders of the interval along which each of the four gate 10ms interval measurements were taken. That is, the 'v' segment displayed in figures 24-25 comprised 40ms of talk. As is evidenced by inspecting the formant tracks displayed in green across the spectrogram in figure 25, the centre frequency values for F2 and higher formants during the 'h' interval may be much higher than during the steady-state interval of the vocalic portion (around ca. 0.28-0.4seconds). In summary, during a C-to-V transition occupying the intervening space between a period of plosion and vocalic resonance, the estimated formants during aperiodic noise will be quite different from those ca. 100-200

[14] 'clos' and 'cl.ph' in figure 25 describe the plosive closures and closing phases, respectively.

ms later on, when the upcoming vowel sound has reached its steady state portion. Despite the fact that the formant estimates in the spectrograms provided thus far may not always perfectly align with the actual formant centre frequencies, these are some of the best estimates an automated analysis can offer in this instance.

3.6.5 Statistical Methods and External Analyses of Perception Timing

This subsection describes the statistical analysis methods and external analysis, which were conducted using a combination of Praat, Excel 2010 and SPSS Statistics 20. Since the segmentation process has already been described in 3.5.3, the descriptions focus on the way the responses were analysed in Excel. Listed.xlsx reports of the open questions and stimulus responses were saved from *SurveyGizmo* onto a PC as xlsx files. The analysis of the stimulus data conducted in Praat is also described: this method was necessary for spectral analysis with respect to VISC.

First, since *SurveyGizmo* by default only allows horizontal listing of responses, it was first necessary to transpose the answers into specific Excel files via a special function in Excel. This method of listing answers vertically facilitated the comparisons of given answers vs. correct responses. This verification was performed by listing the given answers in the left-most columns and the correct ones in the next column to the right and then slotting the function =exact (A2, B2), or using corresponding values for each cell in the third column from the left. This method ensured that no correct answers were listed as incorrect and vice versa.

Second, the responses were filtered using the sort + filter function in Excel so that only a given type of answer (e.g. for vowel or onset types) was being examined at any one time.

This function hides any cells that have not been ticked while selecting response variables. Since this same method was used for the analysis of all variables, it helps to ensure the accuracy of the results reported in the following chapter.

It was also necessary to analyse just over half (ca. 52%) of the stimuli manually, i.e. all the incorrect answers given in the entire experiment (= ca. 3075 of 5904 stimuli). However, the =exact (... , ...) function made it very straightforward to differentiate correct from incorrect answers, ensuring the consistency of the analysis. For example, when it was noted that a given participant had given the answer 'tea' for 'two' at 20ms in a particular stimulus instance, the answer was marked as /u:/ in the right-most column of each main file, signalling that it constituted an incorrect answer for /i:/ at that particular gate. This process was repeated using the sort + filter function for each of the ca. 3075 incorrect answers over a period of ca. two weeks.

A fully automated computer based analysis of the results would be ideal for an experiment of this kind. Such a programming method was not feasible in this research. Having fully described the methodological aspects of this research, we will detail the results in the next chapter.

4. Results

4.1 Overview

In this chapter the results of the experiment described in the previous chapter are detailed and assessed statistically. As suggested earlier, the second theme on 'Contrast and Representation' is not applied to this chapter, as it is quite closely related to 'Phonetic Exponency and Constituent Structure', and in order to simplify the presentation. Thus, four main aspects of the timing of production and perception of English monophthongs are described. The first aspect relates to the primary research question whilst the latter three topics deal with the secondary research questions:

1) 'Temporal Dynamics and VISC' (4.2.1 and 4.3.1) describes the significance of temporal variation.
2) 'FPD and Coarticulatory Direction Effects' (4.2.2 and 4.3.2) describes bidirectional coarticulatory effects.
3) 'Long-Domain Coarticulation and Airflow' (4.2.3 and 4.3.3) assesses long-domain coarticulation and airflow for plosive and nasal codas.
4) 'Phonetic Exponency and Constituent Structure' (4.2.3 and 4.3.3) assesses production effects associated with [+ nasal] sounds.

Subsection 4.4 focuses on perceptual confusions for different vowel sounds, as this part comprises the most important secondary research topic in this research. 4.5 teases apart the results described in 4.2.3 by detailing perception results for different CVNs (e.g. for the vowels in 'tin-ten-tan-ton'). 4.6 details aspects of recognition according to lexical frequency.

All the production results presented in this chapter represent *averages* of male vs. female participants' stimulus productions. The idea behind this thinking is to give as representative a picture of vowel production timing during the aperiodic phase by speakers with five contra six short English vowels as possible, for both male and female speakers. Whilst male and female formant values are quantitatively different (i.e. their absolute values differ) due to anatomical differences (in e.g. vocal tract length and the size of the vocal folds), qualitatively speaking the distinctions are much the same. That is, the relationships between formant centre frequencies remain very similar (see e.g. Kent and Read, 2002). In sum, male and female results are distinguished in this thesis in terms of production in particular. However, it may be possible to draw the same perceptual conclusions from both stimulus types, provided that the underlying trends observed in male vs. female vowel resonance productions are not significantly different. The final subsection (4.7) provides a summary of the results, preparing the way for a) showing how the results extend previous findings in the literature and b) the phonological processing model described in chapter five.

4.2. Vowel Timing and Aspiration in CV(V)/C Production

The line charts in figures 26-45 have formant centre frequencies listed on the y-axes with the individual gate intervals (= 10, 20, 30 and 40ms) on the x-axes. Each of the line charts in figures 26-45 describes the temporal evolution of the production of vocalic information *during the early part of the aperiodic phase*. F2 and F3 are described for low, mid and high vowels as the formant centre frequencies vary across the two male and female speakers' vowels. The choices of scales for F2 and F3 in this chapter are based on typical formant values vowel formants tend to comprise during the early part of

the aperiodic phase (as described in 3.6.2-3.6.3). For example, the minima and maxima for F2 tend to straddle 1400-2100 Hz during the aperiodic phase respectively, for stimuli as produced by male speakers (see e.g. figure 26). For female speakers (see e.g. figure 27), the equivalent numbers are ca. 1600-2300, respectively. Each pair of figures (e.g. figures 26 and 27) compares the results for F2 and F3 across the four gate intervals in both male and female speaker realisations of /i: ɪ a ɛ ʌ ʊ u: ɔ: ɑ: ɒ/. F1 is not included in the analysis, since F1 may be difficult to measure reliably during the aperiodic phase.

The y-axes showing the formant centre frequencies in figures 26-45 have differing minima/maxima across vowels with different height features, since vowel formant centre frequencies may differ significantly due to the exponency of height. A similar conclusions applies to male and female speakers' CV(V)/Cs, in that female formant centre frequencies are ca. 15% higher than male ones (cf. Kent and Read, 2002).

4.2.1 Temporal Dynamics and VISC – the Evolution of Spectral Information in Production

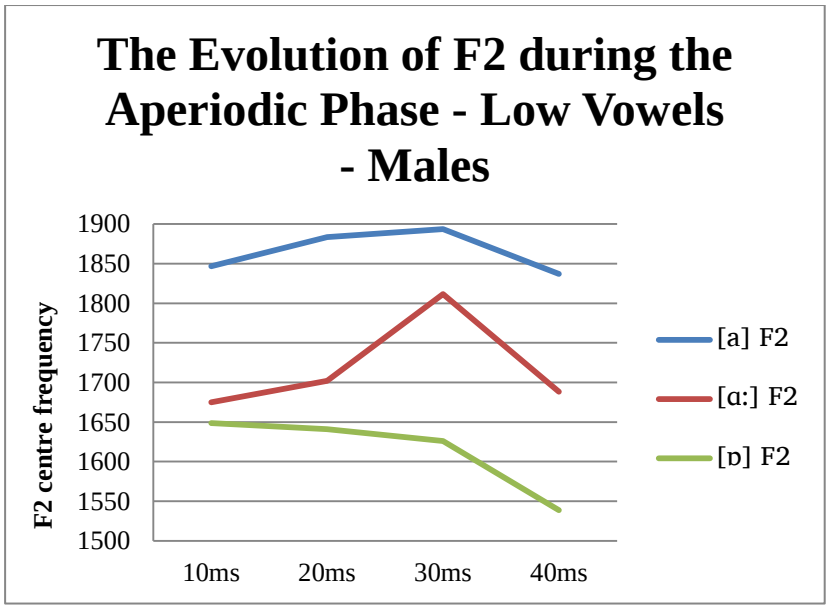


Figure 26: The temporal evolution of F2 in male low vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

By examining the values for male [a ɑ: ɒ] in figure 26, it can be seen that there is not a great amount of spectro-temporal variation in F2 between the individual gates for the low vowels. Their formant trajectories remain comparatively level across time, although an overall descent is evidenced between 30ms and 40m (especially for [ɑ:]).

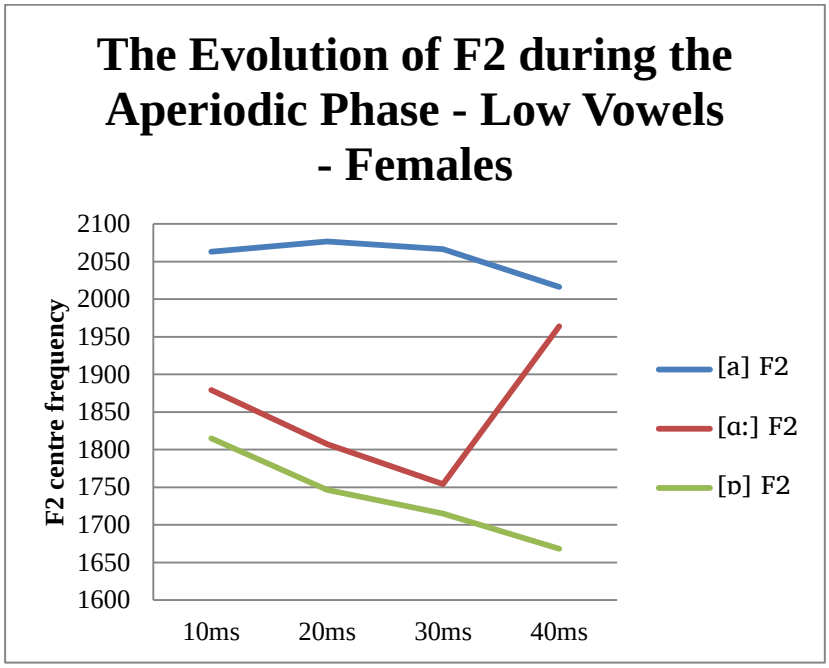


Figure 27: The temporal evolution of F2 in female low vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

For female F2 in [a ɑ: ɒ] (cf. figure 27), it can be seen that more descending trajectories are evidenced for all three vowels compared to the equivalent male values (cf. figure 26), except for [ɑ:] between 30 and 40ms, where a sharp rise of ca. 200Hz is evidenced. In summary, the overall qualitative

distinctions between [a ɑ: ɒ] are relatively similar for both male and female F2 in low vowels.

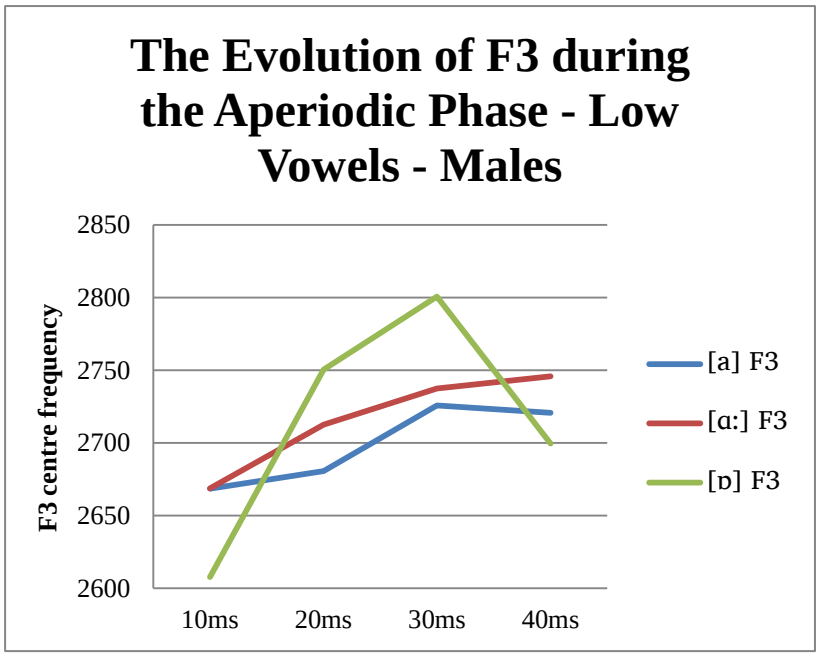


Figure 28: The temporal evolution of F3 in male speaker low vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

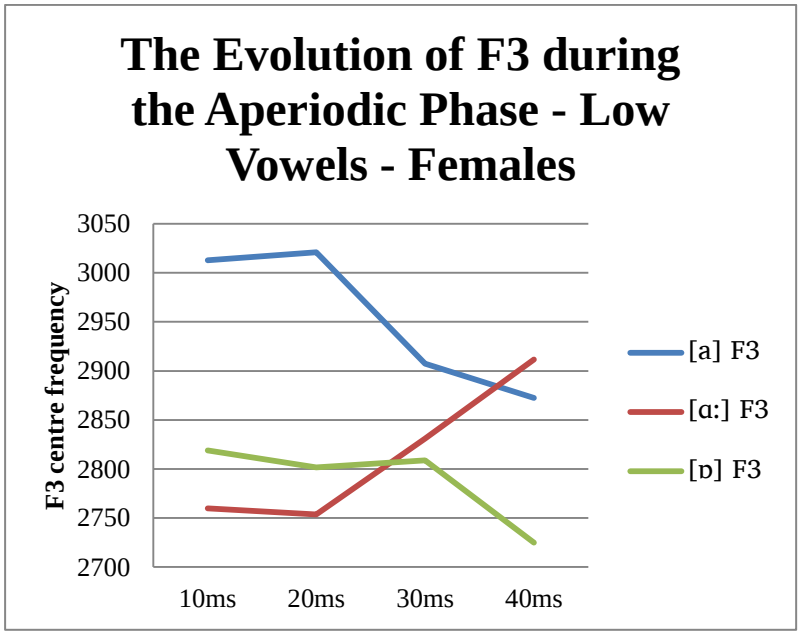


Figure 29: The temporal evolution of F3 in female speaker low vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

For F3, somewhat larger distinctions can be observed in male productions (cf. figure 28) contra female speaker ones (cf. figure 29) than for F2. For example, there are some differences between male and female productions of [ɑ: ɒ] in terms of their trajectories between individual gates. However, the most noticeable difference between F2 and F3 in low vowels relates to the more strongly descending trajectory of F3 in female [a] (cf. figure 29), which descends ca. 150 Hz from ca. 3025 Hz between gates 1-4. The male one rises only 50Hz from ca. 2675Hz (cf. figure 28).

To summarise the differences between male and female productions of low vowels in terms of their overall timing (as well as between individual gates), there are certain significant temporal differences for individual vowels. The main differences are the F3 transitions in [a] (cf. figures 28-29) and distinctions in the F2 transitions for [ɑ:] (cf. figures 26-27). The transitions for the [ɒ] vowel are similar. Most of the transitions are similar in low vowels for male and female speakers, in particular from a qualitative perspective (= the female values are higher in frequency), with 9 out of 12 vowels' transitions being similar from this viewpoint.

The most important point with respect to male and female F2-F3 for low vowels is that both formants have higher values than normally observed during the vowel steady state (cf. subsection 3.6.2-3.6.3). The second important issue is that the formants' centre frequencies remain spectro-temporally variable across time during the first 40ms of the aperiodic phase. For example, male F2 starts at around 1.600-1.700 Hz subsequent to the plosive burst for [ɒ ɑ:] (cf. figure 26) and

finishes at a much higher value for [ɑː] than for [ɒ], and with differing values in between at the 20 and 30ms gate intervals. Contrary to what we might expect from observations of typical F3 centre frequencies, the values for male low vowels rise throughout the first 40ms for [a ɑː ɒ] (with one exception for [ɑː], cf. figure 28).

In summary, F2 and F3 centre frequencies fairly constantly approach their steady state values in [a ɒ ɑː]. An exception to this tendency is female [ɑː], in which F2 and F3 rise between 30 and 40ms rather than descend.

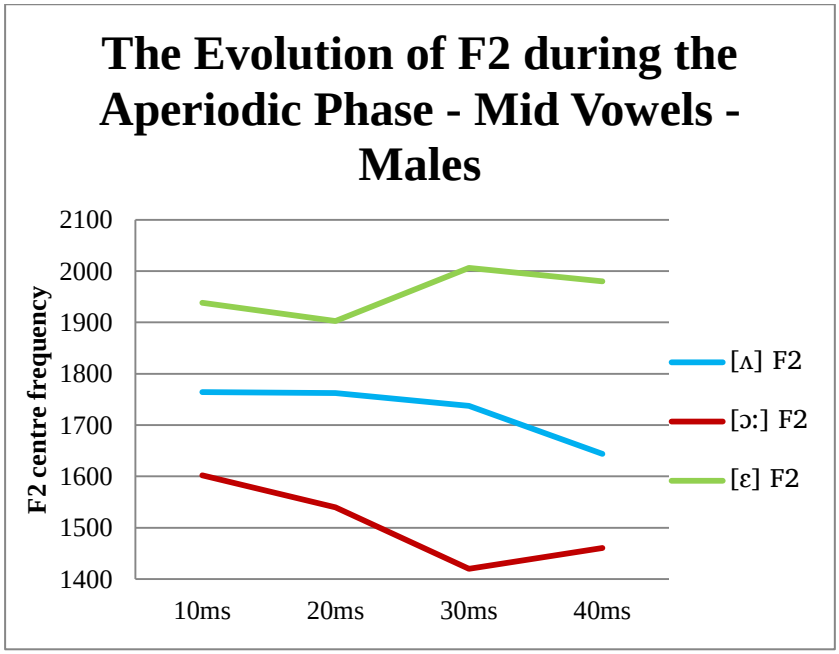


Figure 30: The temporal evolution of F2 in male speaker mid vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

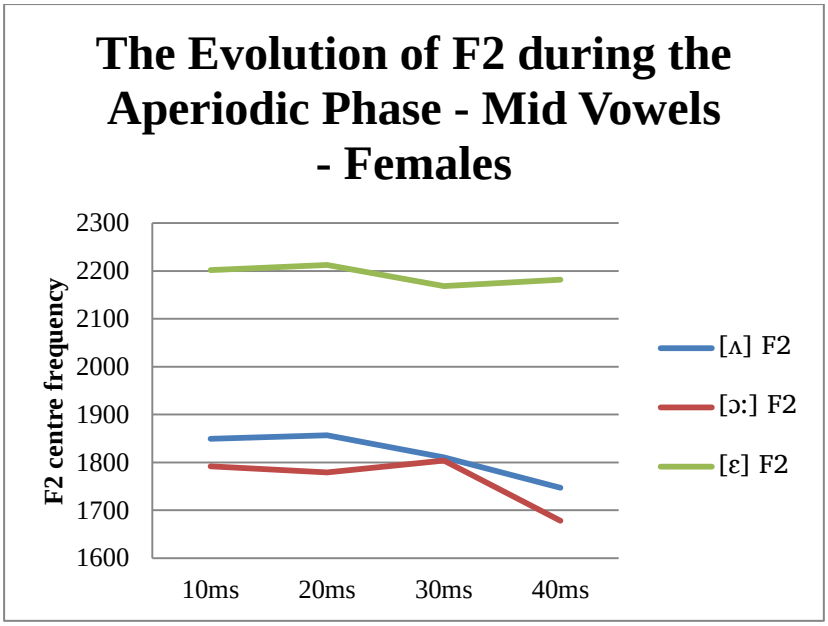


Figure 31: The temporal evolution of F2 in female speaker mid vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

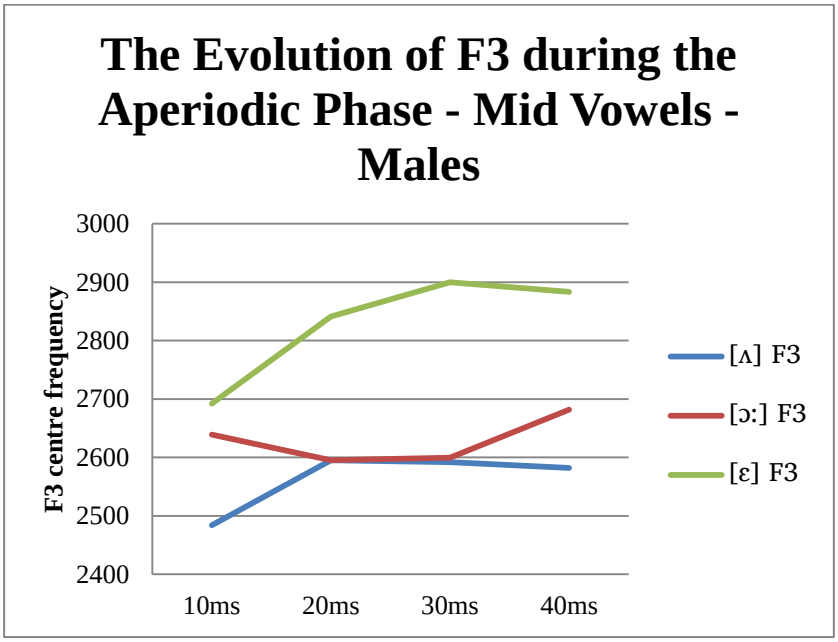


Figure 32: The temporal evolution of F3 in male speaker mid vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

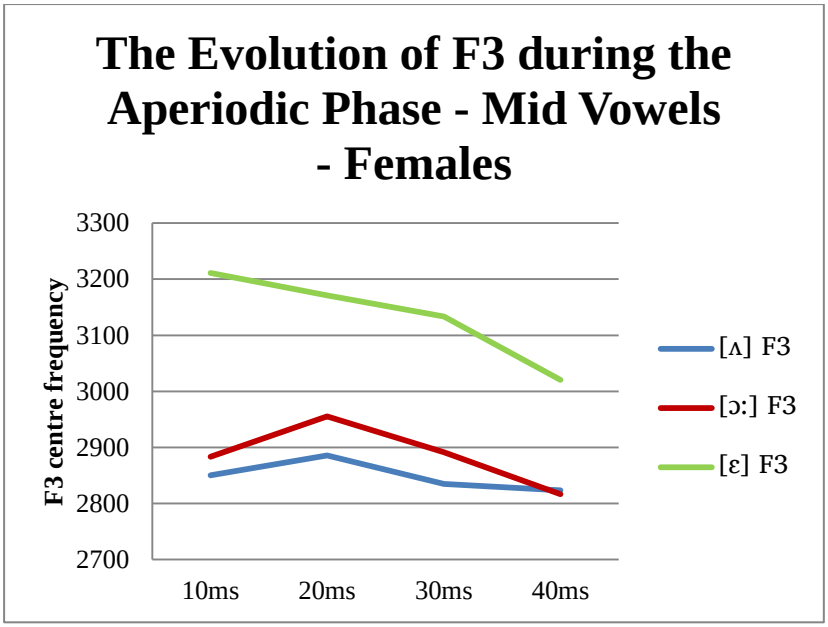


Figure 33: The temporal evolution of F3 in female speaker mid vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

When we compare the observed values for F2 in mid vowels as produced by males (cf. figure 30) and female speakers (see figure 31), we can see that none of the trajectories observed for [ɔ: ʌ ɛ] vary very much in frequency across time. The only potentially significant difference observed for F2 is for [ɔ:], where a small rise observed for female speakers between the 30 and 40ms intervals translates into a small descent in the male speaker equivalents. The separation between male [ʌ] and [ɔ:] is greater than between the equivalent female productions, contrary to what we might expect from theoretical predictions of formant centre frequencies in male vs. female speakers (this issue could e.g. reflect accent differences). Similarly as for the low vowels, larger differences are evidenced for F3 than for F2 in [ɔ: ʌ ɛ] (cf. figures 32-33). This observation is especially true for [ɔ: ɛ] between the 30 and 40ms gates. In the female productions, descents in F3 centre frequencies are observed for

both [ɔː ɛ] (cf. figure 33), which is not true for either [ɔː ɛ] in male productions (cf. figure 32).

In sum, the F2 transitions for mid vowels are similar for male and female speakers. The transitions in F3 are more spectro-temporally variable across time, with larger differences in [ɔː] and [ɛ]. The most important point to be taken for mid vowels is that the trajectories of [ɔː ʌ ɛ] approach their steady state values relatively gradually, with the exception of F2 and F3 in male productions of [ɔː] subsequent to the 30ms gate (cf. figures 30 and 32). Not only do female and male speakers display similar transitions overall, but overall the mid vowels [ɔː ʌ ɛ] (cf. figures 30-33) have somewhat less spectro-temporally variable transitions compared to [a ɑː ɒ] (cf. figures 26-29). Mid vowels approach their steady state values more gradually than low vowels.

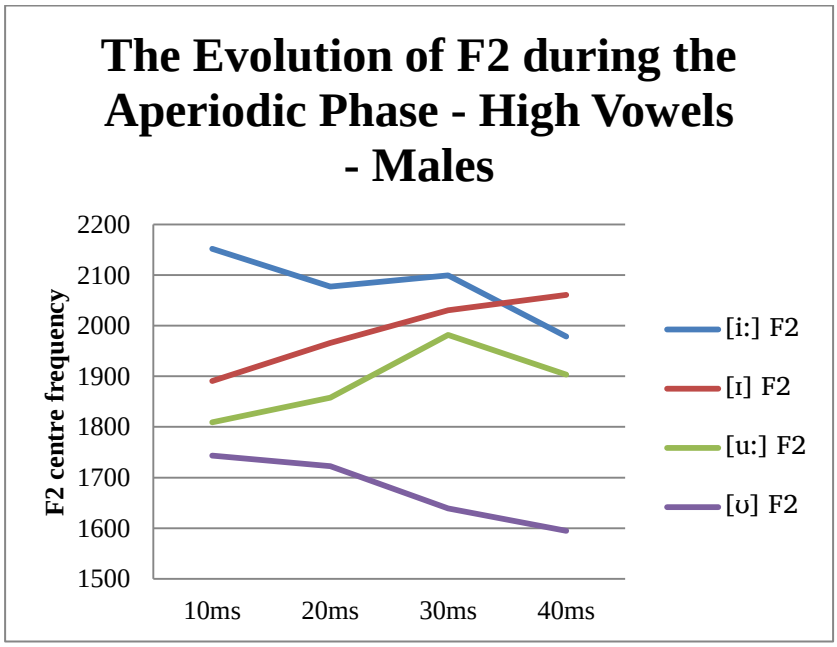


Figure 34: The temporal evolution of F2 in male speaker high vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

For F2 in male high vowels, the individual formants are characterised by spectro-temporally quite dynamic trajectories. The trajectories of [ɪ uː] are more similar spectro-temporally than those of [iː ʊ], in that [ɪ uː] occupy a similar frequency range and both have rising trajectories on the whole (cf. figure 34). [iː ʊ] both have descending transitions, with F2 for [iː] having a much higher centre frequency compared to [ʊ]. Only male [ɪ] has a rising F2 trajectory (cf. figure 34).

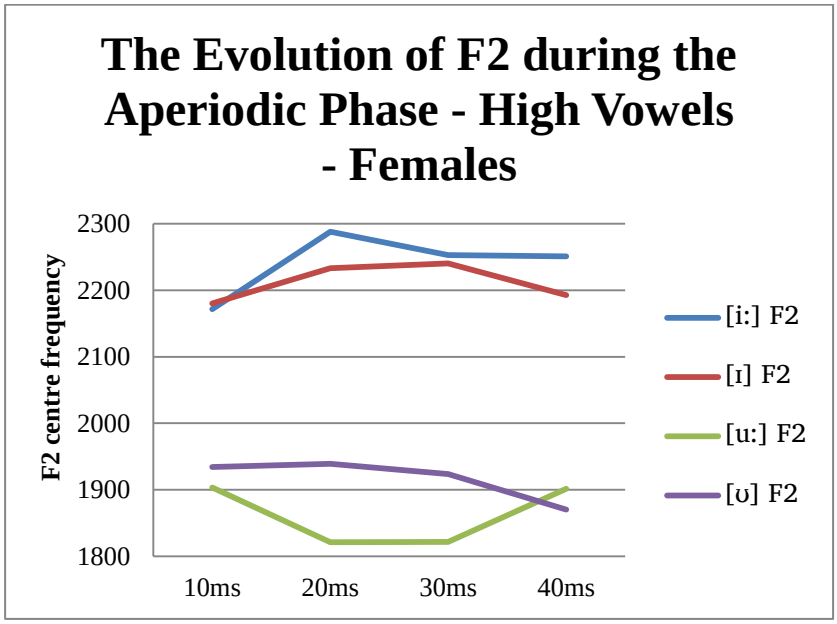


Figure 35: The temporal evolution of F2 in female speaker high vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

The F2 values in female productions have more level trajectories (cf. figure 35), and there is less dynamic temporal variation in female [iː ɪ uː ʊ] than in male high vowels (cf. figure 34). The temporal trajectory of [uː] is relatively level in the female productions (cf. figure 35), whereas in male

productions of [u:] F2 ascends ca. 100 Hz between gates 1 and 4 (cf. figure 34).

The formant transitions are more similar from a qualitative viewpoint, in that both male and female front vowels [i: ɪ] have higher centre frequencies than their back counterparts [u: ʊ] (cf. figures 34 and figure 35).

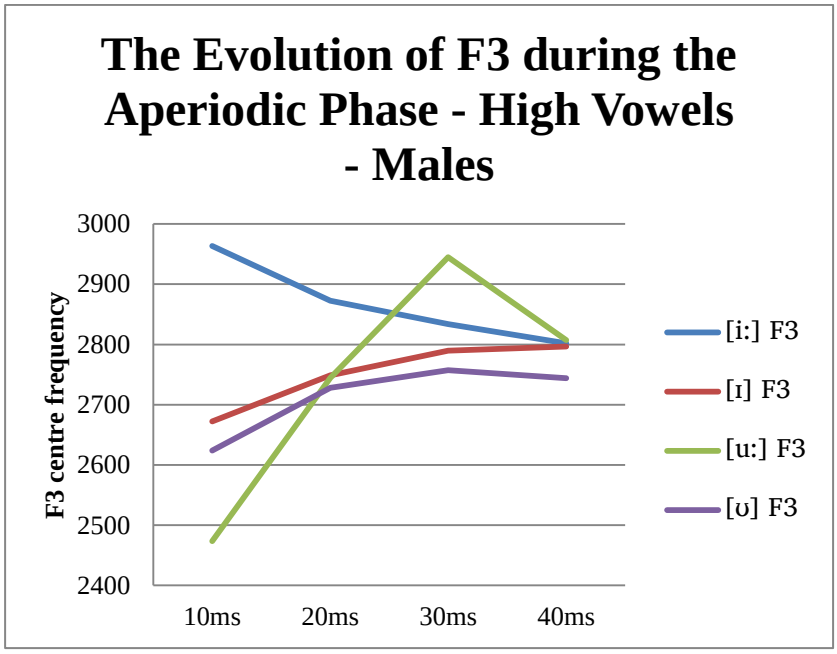


Figure 36: The temporal evolution of F3 in male speaker high vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

Overall, when examining the values for [i: ɪ u: ʊ] F3 in male productions, we can see that the formant trajectories in the high vowels have very different starting points (cf. 10ms gate in figure 36), but near-identical ending points around 2750-2800Hz (cf. the 40ms gate in figure 36). The trajectories of [i: ɪ ʊ] are similar between 20 and 40ms, with the formant centre frequencies straddling 2700-2850Hz (cf. the much larger differences at 10ms). For [u:], a large rise of ca. 500Hz (2450-

2950Hz) between gates 1 and 3 translates into a ca. 150Hz fall between gates 3 and 4 (from ca. 2950Hz to 2800Hz). In this respect, the transitions for F3 in male [u:] (cf. figure 36) are similar to those observed for F2 (cf. figure 34).

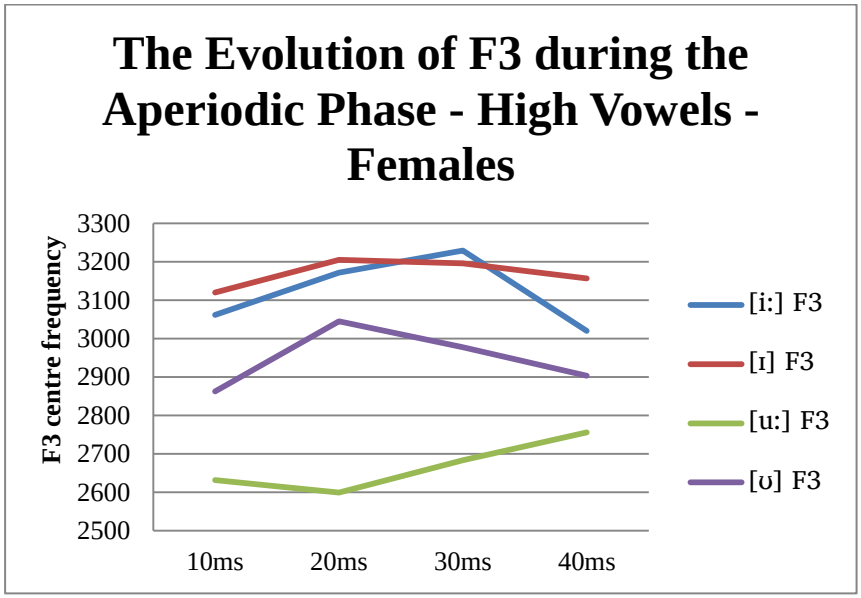


Figure 37: The temporal evolution of F3 in male speaker high vowels between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

The overall trajectories as well as the time course of the individual formant tracks for F3 in female high vowels are similar to those for female F2 (cf. figures 35 and 37). This generalisation is particularly true for the front vowels' [i: ɪ] F2-F3 centre frequencies, which a) do not vary greatly in frequency across time (cf. figures 35 and 37) and b) occupy the frequency range we would expect as derived from Stevens' (1998) tube model predictions in vowel resonances produced during aperiodic friction.

In summary, there are at least some relatively large distinctions between the male and female productions relating to F3. For example, the trajectory of F3 in [u:] is distinctive in the male productions compared to female [u:] (cf. figures 36

and 37), with a much sharper rise in the female variant than in male [u:]. The rises in female F3 for [ɪ ʊ] translate into descending trajectories in their male equivalents. The most important point here is that both the male and female F3s for high vowels converge between 30 and 40ms. In particular between the 20 and 40ms gates, all F3 centre frequencies can be seen to converge towards a more neutral point in the middle of the ca. 2700-3200 Hz region in acoustic-temporal space (cf. figures 36 and 37). The observed trajectories are qualitatively quite similar, despite the differences in spectro-temporal variability between individual gates.

In summary, the individual formant centre frequencies are listed separately for low, mid and high vowels in figures 26-37, since the overall trajectories for F2 and F3 are more clearly lowering towards their steady state targets in high than in either mid or low vowels. The resonances approach their steady state values more rapidly in high vowels. Mid vowels occupy a middle ground in this respect, reflecting their intermediate F2 and F3 values in the vowel space. High and low vowels display qualitatively more rapidly changing trajectories, reflecting their more peripheral positions within the vowel space.

Since the experiment is a repeated measures ANOVA design, there is need to assess whether the proportions of variances between the individual gates for low, mid and high vowels are significantly different from zero. Sphericity is a concept which can be used to assess this statistical property (e.g. Field, 2009). If the distributions of the received values are not significantly skewed and do not have strong peaks or tails, then their values may be shown not to violate the assumption of sphericity. Such distributions have gradually descending slopes when passing from the highest values downwards, and no clusterings of high or low values.

Mauchly's test of sphericity shows significant differences in the proportions of variances between the structural conditions across time for F2 and F3 in 3 of 4 statistical comparisons (i.e. the values for F2 and F3 are not normally distributed for all vowels at each gate interval):

For male F2 sphericity is violated, χ^2 (5) = 15.300 , p < 0.02: at some points in the distribution for F2 between gate intervals, the amount of spectro-temporal variation between low, mid and high vowels is considerable. The variability in formant values between gates has a too skewed distribution to be assessed using a standard parametric test. Since there is no non-parametric equivalent for a repeated measures design with multiple dependent variables, statistical comparisons are not feasible in this instance. For male F3, sphericity is violated, χ^2 (5) = 24.399, p < 0.001.

For female F2 sphericity is violated, χ^2 (5) = 12.888, p < 0.03. Female F3 is normally distributed, χ^2 (5) = 6.856, p = 0.235. The underlying p statistic is not significant, F (3) = 1.761, p = n.s. Post-hoc tests show no differences for F3.

4.2.2 FPD and Coarticulatory Direction Effects

An account of the statistical tests performed on the temporal evolution of F2 and F3 in /p t k/-V-/p t k/ syllables is first given in this subsection. One of the purposes of the first part of this subsection is to show that there are no significant differences with respect to the temporal evolution of vowel production timing for /p t k/-V-/p t k/ monosyllables. The other purpose of this part is to contrast the results for /p t k/-V-/p t k/ syllables with those presented in the second part of this subsection, in which an account is given of how syllable structure influences the phonetic exponency of F2 and F3.

The statistical tests performed on /p t k/-V-/p t k/ syllables affirm that the average values[16] observed for F2 and F3 for female and male speakers are not statistically significant. The following results are observed for /ɪ a ʊ ʌ ɒ/, which are the five vowels in /p t k/-V-/p t k/ syllables in this study. F2 is normally distributed, χ^2 (2) = 5.250 , p = 0.072. F2 is not statistically significant, F(2), p = 0.207 = n.s. F2 for female speakers violates sphericity, χ^2 = (2) = 10.281, p < 0.01. For F3, male values are normally distributed, χ^2 (2) = .477, p = 0.788, whilst the female productions' exponent values violate sphericity, χ^2 (2) = 10.226, p < 0.01. The values for males with respect to F3 are not significant, F(2), p = 0.397 = n.s. In summary, the differences with respect to potential phonetic co-extensiveness between onsets and coda portions in /p t k/-V-

[16] In order not to have a very large number of multiple comparisons here, the average F2-F3 values between each of the four gates were taken rather than comparing values at each gate.

/p t k/ monosyllables do not significantly affect the FPD of vowel production.

Figures 38-41 show the range of spectro-temporal variation associated with F2 and F3 at the four gate intervals in different stimulus structures. The y-axes show the extent of variation in the main formants, while the x-axes show the values at each gate for CVVs (blue lines), CV-/p t k/s (brown lines) and CVNs (green lines). The descriptions for F2 and F3 are separated in figures 38-41 to enable a more in-depth presentation on temporal differences in VISC, as well as to show the differing trajectories for male/female F2 and F3 for different structures (cf. figures 38-39 and 40-41).

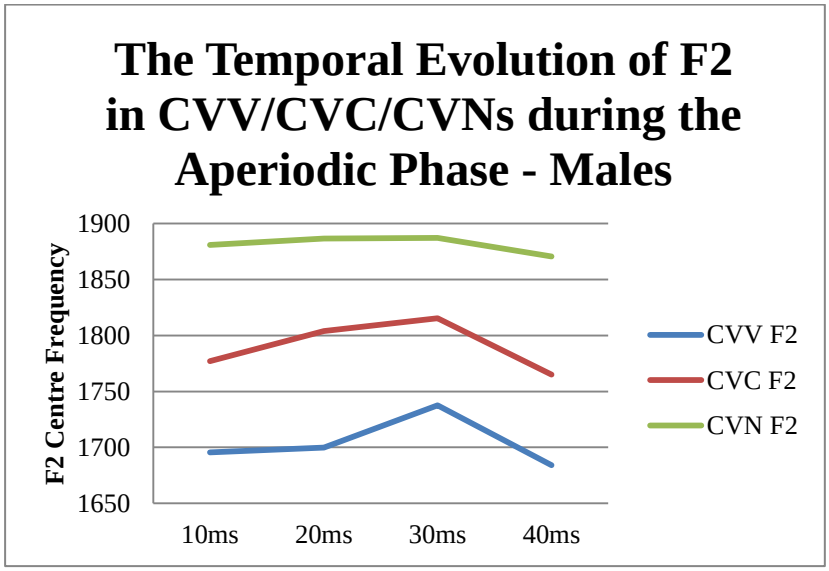


Figure 38: Temporal variation in F2 between CVV/CVC/CVNs at different gates (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

Figure 38 shows that the trajectories for F2 as produced by male speakers start to ascend in frequency between 20ms and 30ms and then descend after 30ms: the values for CVVs and CV-/p t k/ (i.e. CVC) syllables have similar trajectories across as well as between individual gates. Nasality may have an overall raising effect on F2 (cf. the green line in figure 38),

with little spectro-temporal variability in formant centre frequencies in time. We can also see that in contrast to CVNs, the exponents of F2 in CVVs and CVCs approach their target values more rapidly (cf. the middle and right-hand parts of figure 38).

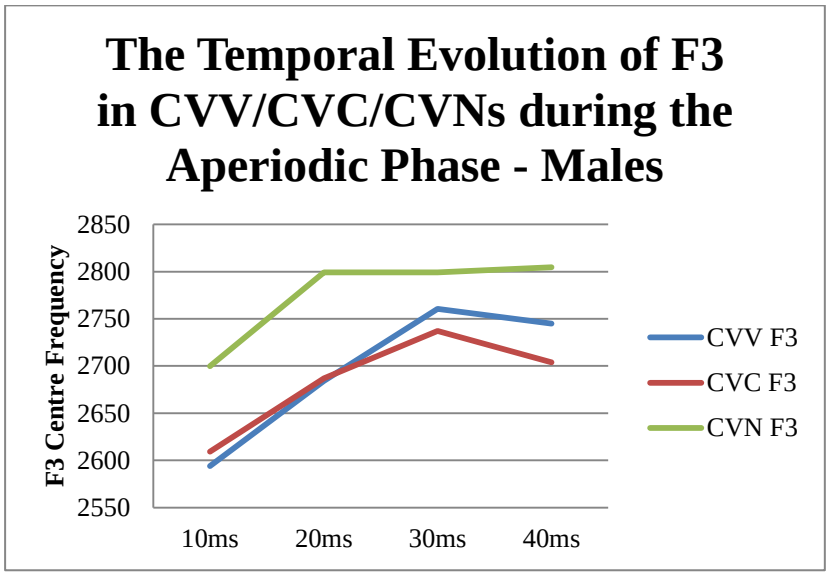


Figure 39: Temporal variation in F3 between CVV/CVC/CVNs at different gates (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

Figure 39 describes the trajectories for F3 in male stimulus structures, whose formant frequencies start to ascend between 10 and 20ms: the values for CVVs and CV-/p t k/ word forms are very similar across the four gates. Nasality has a raising effect on F3, with ca. 125 Hz higher values overall, similarly to female F2 (cf. figure 40 below): the differences between CVCs and CVNs are even larger than between CVVs and CVNs after 20ms for F3 (cf. figure 39), despite their differences in syllable structure and phonetic exponency.

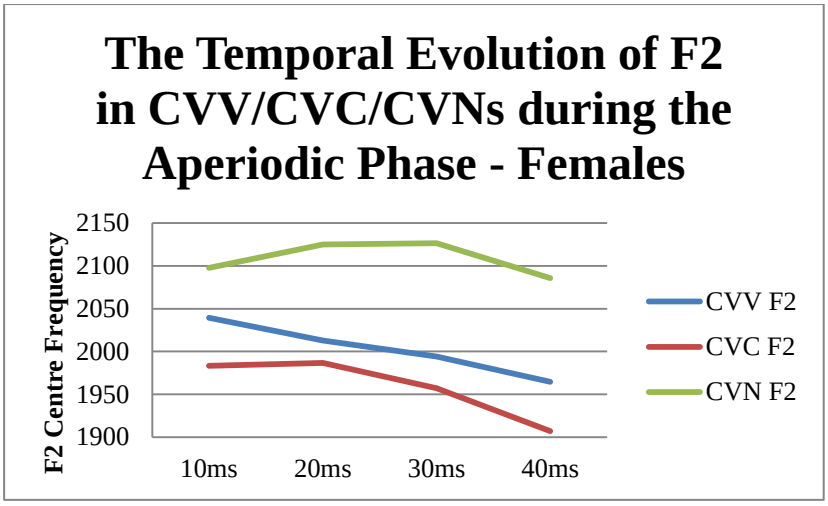


Figure 40: Temporal variation in female F2 between CVV/CVC/CVNs at different gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

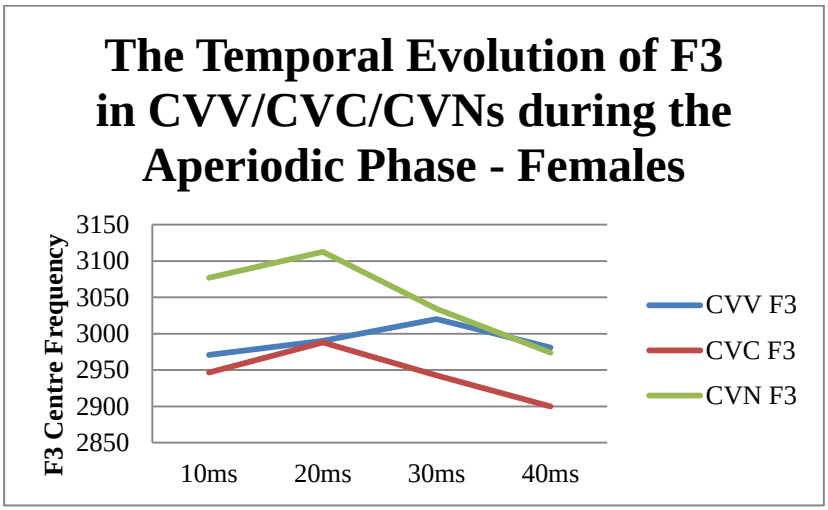


Figure 41: Temporal variation in female F3 between CVV/CVC/CVNs at different gate intervals (onset = /p/, /t/ or /k/, potential coda = /p/, /t/, /k/ or /n/)

For female F2 and F3, the overall trajectories observed in CVCs and CVVs contra CVNs are very similar to those observed for male productions, with one notable exception and one minor distinction: for male F2 centre frequencies, CVCs occupy a near-identical spectral range compared to CVVs (with minor temporal distinctions after 20ms). For females, CVCs occupy a ca. 50-100 Hz lower range. A minor difference for F3 female CVNs is evidenced between gates no 3 and 4, in that the

centre frequency value for CVCs is marginally higher than for CVNs (cf. the blue and green lines in figure 41 at the 40ms gate). However, the temporal evolution for F2-F3 in different syllable structures remains qualitatively similar in male and female productions.

Formant centre frequencies in CVNs always start higher than in other stimulus types, and remain so throughout the early part of the aperiodic phase, with the single exception of female F3 at 40ms: post hoc tests show many of the differences for individual vowels in CVNs to be significant (cf. the next subsection).

To summarise effects associated with CVVs contra CVCs and CVNs, repeated measure ANOVAs display the following differences: male F2 is normally distributed, χ^2 (2) = .450, p = 0.798. F2 is very highly significant, F(2) = 248.508, p < 0.001. Post hoc tests show that all three comparisons are highly significantly different at p < 0.01. F3 for males is normally distributed, χ^2 (2) = 1.111, p = 0.574, and statistically significant, F(2) = 24.3, p < 0.002. Post-hoc tests show that male F3 differs significantly for CVCs contra CVNs, p < 0.02.

For females F2 is normally distributed, χ^2 (2), p = 0.511, and highly significant, F(2) = 65.508, p < 0.001. Post-hoc tests show that all three comparisons are significantly different at p < 0.04. F3 for females is not normally distributed χ^2 (2) = 6.055, p = 0.048).

From these four statistical tests on male and female F2 and F3, we can conclude that the results for different stimulus structures are in line with the descriptions given in figures 38-41 for different vowels. In sum, formant centre frequencies are significantly affected in their FPD, both with respect to vowel quality and the bidirectionality of coarticulation.

4.2.3 Long-Domain Coarticulation and Airflow: Phonetic Exponency and Structure for [+ Nasal] Stimuli

Figures 42-45 display the temporal dynamic properties of five different CVNs with [ɪ ɛ a ʌ ʊ]. F2 and F3 are described for male and female speaker CVNs (cf. figures 42-43and 44-45, respectively). The organisation of figures 42-45 is similar to those in figures 26-43.

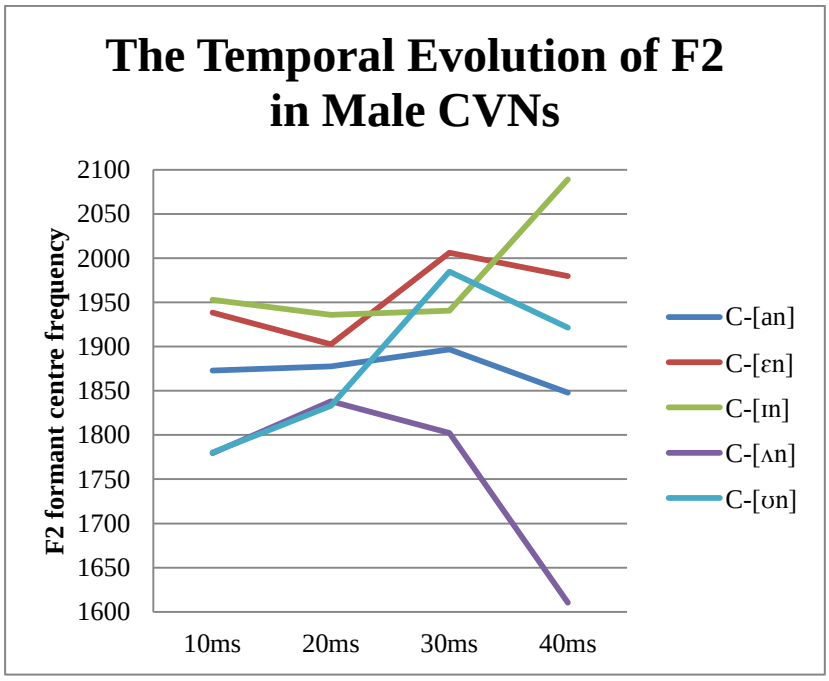


Figure 42: The temporal evolution of F2 between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/)

Figure 42 shows that the formant trajectories in F2 in male CVNs vary quite extensively for [ɪ ɛ a ʌ ʊ] in spectro-temporal terms after ca. 20ms.

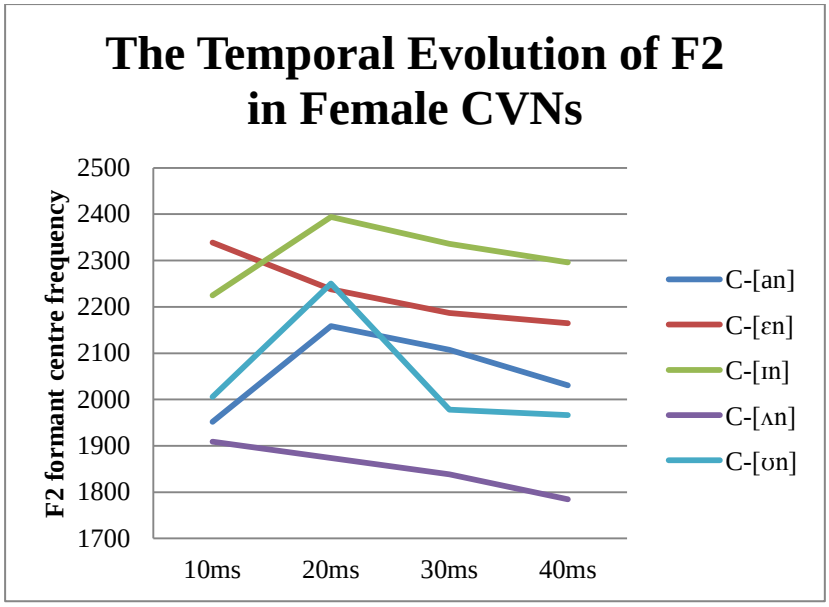


Figure 43: The temporal evolution of F2 in female CVNs between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/)

Figure 43 shows that the spectro-temporal variability observed in female F2 for [ɛ ʌ] is less marked than for female [a ɪ ʊ]:

F2 centre frequencies in [ɛ ʌ] descend slightly throughout, whilst sharp rises are evidenced in [a ɪ ʊ] early on. Qualitatively, the formant trajectories between male and female F2 in CVNs are similar. However, there are some differences in the spectro-temporal variability in F2. For example, male F2 ascends sharply after 20ms (cf. figure 42) in [ʊ], whilst in female F2 descends (cf. figure 43).

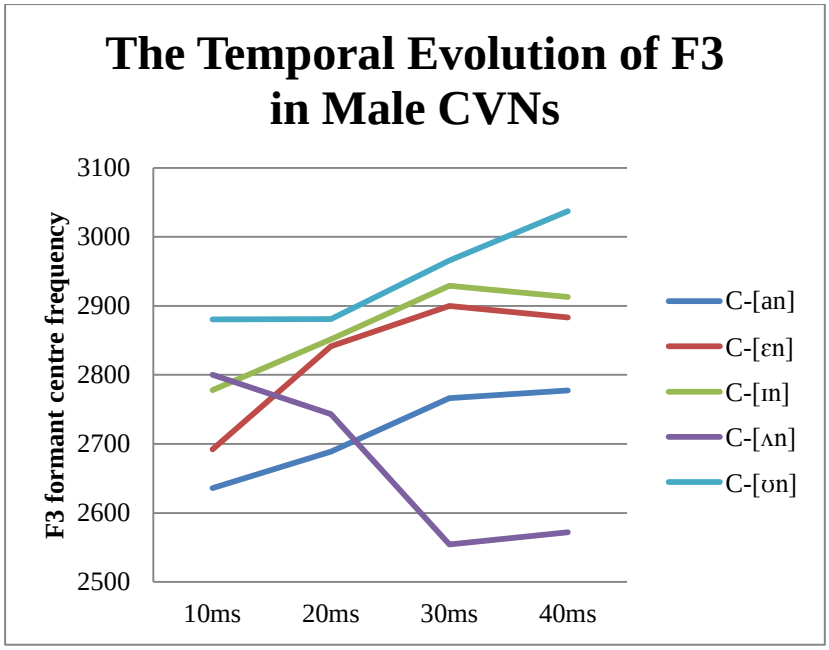


Figure 44: The temporal evolution of F3 in male CVNs between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/)

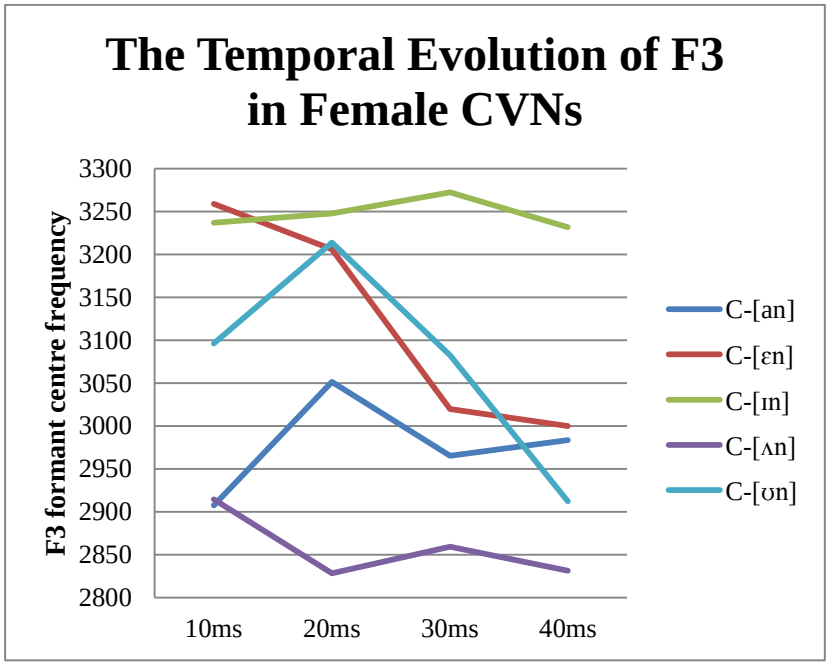


Figure 45: The temporal evolution of F3 in female CVNs between the 10, 20, 30 and 40ms gate intervals (onset = /p/, /t/ or /k/)

For male and female F3, similar conclusions apply as for F2, with the exception of [ʊ]: in all other CVNs for both males and females, vowel formant centre frequencies converge

towards a similar point in spectro-temporal space close to 40ms.

Since the number of stimuli for [CʊN] and [CʌN] is half of that for the front vowel equivalents (i.e. northern and southern speakers' stimuli differ in terms of exponency, cf. table 6), we will only test for differences for [ɪ ɛ a] in this section. A choice was made not to compare against potential differences for [ʌ ʊ].

The production values across time for different CVNs are normally distributed for male F2, χ^2 (2) = 1.205, p = 0.548. F2 is not significant for males, F (2) = 4.713, p = 0.059. Since the p value is close to the alpha level of $p \leq 0.05$, and the F-statistic is relatively large, the exponency of F2 might still be significant perceptually. Post hoc tests show no significant differences for F2 for males. For F3 the production values are normally distributed for males, χ^2 (2) = 3.202, p = 0.202. F3 is very highly significant for males, F (2) = 49.212, $p < 0.001$. Post hoc tests display significant differences between [a] and [ɪ] ($p < 0.002$) and [a] and [ɛ] ($p < 0.04$).

For female F2, the data values for the exponents of F2 violate sphericity, χ^2 (2) = 6.979, p = 0.031. F3 for females if normally distributed, χ^2 (2) =1.820, p = 0.403. The statistical test run for female F3 displays the following significant differences, F(2) = 9.810, $p < 0.02$: post hoc tests show the exponents of [a] / [ɪ] to be significantly different, $p < 0.02$. The differences between [a] / [ɛ] and [ɛ] / [ɪ] are not statistically significant.

4.3 Vowel Timing and Aspiration in CV(V)/C Perception

4.3.1 Temporal Dynamics and VISC: the Evolution of Spectral Information in Vowel Recognition

Figure 46 on the temporal dynamics of recognition is organised as follows: the proportion of correct answers is on the y-axis, with the individual gates displayed on the x-axis. In the middle of figure 46, we can see a trendline across the four gates (in black). Comparing the values across the trendline with the observed values for individual gates shows that the recognition stays below par until past 30ms, with a significant rise in recognition between 30ms and 40ms.

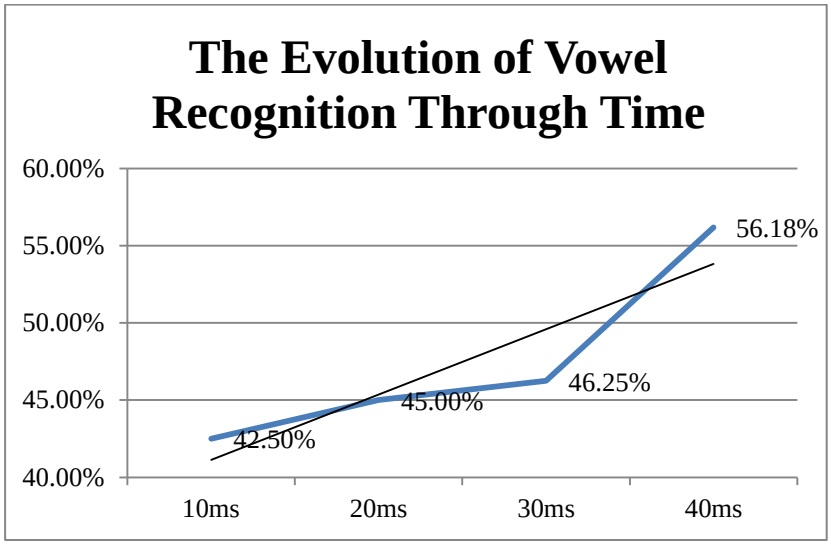


Figure 46: The evolution of vowel recognition through time

The recognition values are normally distributed, χ^2 (5) = 4.512, p = 0.479. The factor of temporal evolution through time is very highly significant, F(3) = 26.293, p < 0.0001. This result shows that as listeners hear more of the vowel, it is recognised more reliably. Post hoc multiple comparisons display highly significant differences between gate number 4 and gates 1-3, as follows: gate number 2 differs from gate interval number 4 at p < 0.0003, whereas gates 1 and 3 differ from gate number 4 at p < 0.001. There is a larger rise in recognition going from 20 to

40ms than from 10ms to 30ms or 40ms. The post hoc tests show that it is not until listeners hear *more than 30ms* of vowel resonance that they can reliably recognise vowel quality from aperiodic friction.

4.3.2 FPD and Coarticulatory Direction Effects

Place of Articulation – Onsets

Figure 47 is organised as follows: the proportion of correctly recognised vowels according to onset place of articulation is shown on the y-axis, with the individual onset types displayed on the x-axis. In the middle of figure 47 can be found the results for the three types of onset (bilabial, alveolar and velar, respectively).

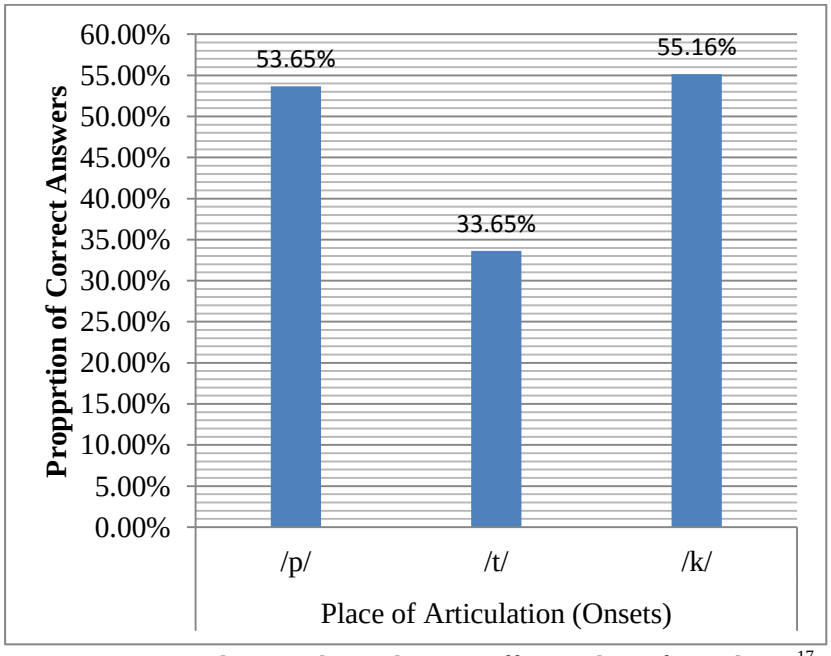


Figure 47: FPD and coarticulatory direction effects - place of articulation[17]

[17] '/ /' for distinctive sounds are used in the remainder of this chapter, since the results in 4.3-4.7 represent perceptually distinct categories of sounds rather than acoustic measurements.

The recognition values are normally distributed, χ^2 (2) = 1.634, p = 0.442. Place of articulation is very highly significant at F(2) = 48.735, p < 0.000001. Post hoc pairwise comparisons display a very highly significant difference between vowel recognition with velar and alveolar as well as bilabial and alveolar onsets (p < 0.000001 for both comparisons). These two results show that listeners find it much harder to recognise vowels from stimuli with alveolar onsets. The stimuli with velar onsets lead to a higher overall level of recognition than bilabial onsets (with a difference of ca. 1.5%, cf. figure 47), a difference which is not significant.

Vowel Quality

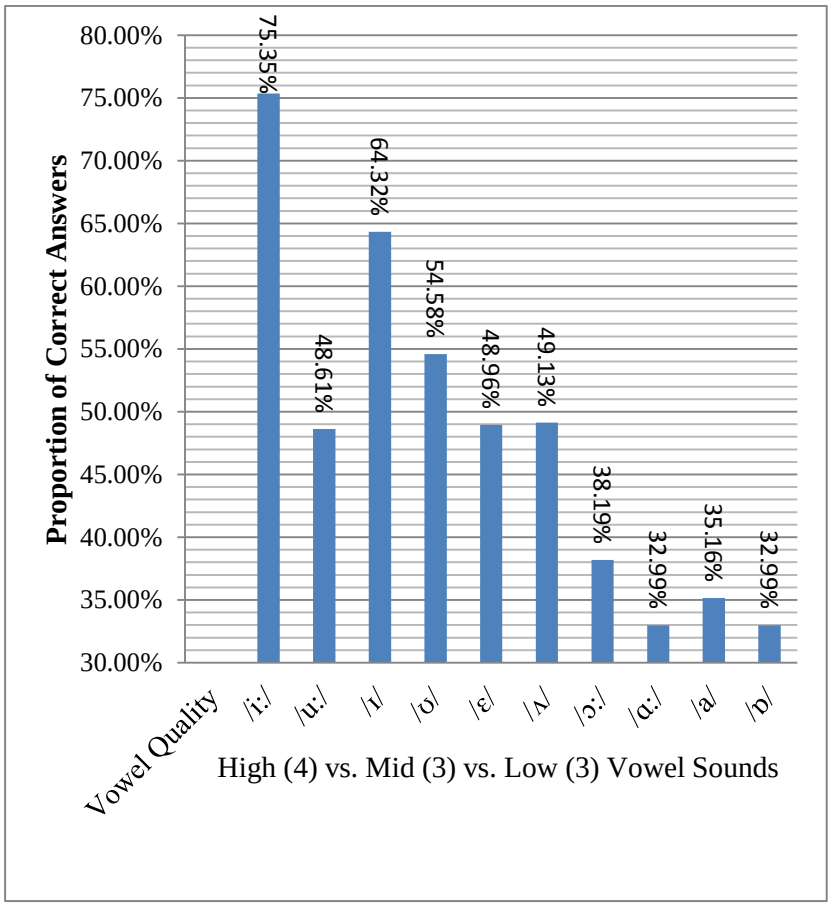


Figure 48: Correctly recognised vowels according to vowel quality

Figure 48 is organised as follows: the proportion of correct answers is on the y-axis, with the individual vowels displayed

on the x-axis. In the middle of figure 48 can be found the recognition results for the 10 vowel categories. The ordering in figure 48 is high to mid to low vowels left-to-right, since high vowels tend to be recognised more reliably and earlier than mid and low vowels.

Since the number of vowel tokens differs across stimulus categories, it was chosen not to compare potential differences related to vowel quality statistically. As explained in chapters 2 and 3, however, having different stimulus numbers for different structures constitutes a necessary sacrifice when considering the primary research question using real word stimuli.

Vowel Recognition Before /p t k/

The recognition values violate sphericity, χ^2 (2) = 10.416, $p < 0.01$. This result confirms that we cannot model recognition statistically before /p t k/ using repeated measures design tests, since there is no non-parametric equivalent test for one-way repeated measures ANOVA.

4.3.3 Long-Domain Coarticulation and Airflow

Figure 49 is organised as follows: the percentage proportions of correct answers for different vowels according to coda type is on the y-axis, with the individual syllable structures available to listeners displayed on the x-axis. In the middle of figure 49 can be found the results for CV-/p t k/ monosyllables and CVNs, with CVVs on the far right.

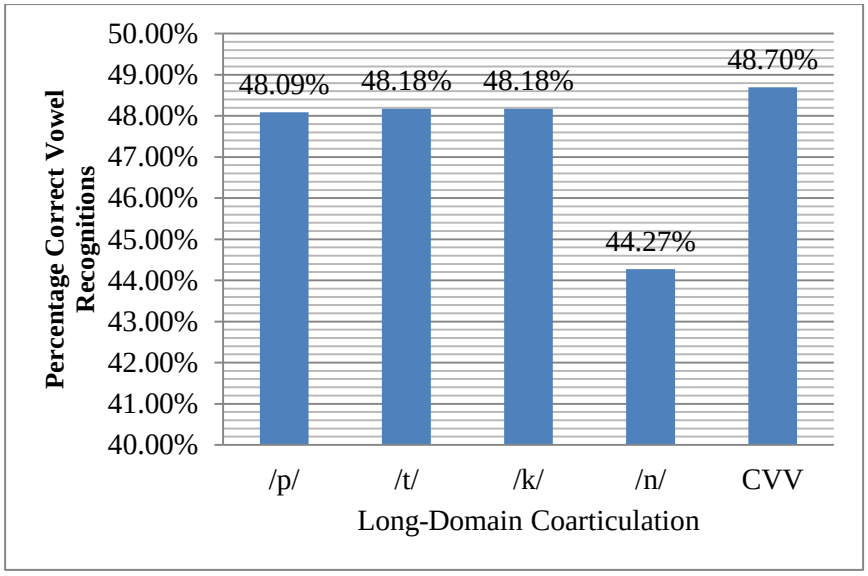


Figure 49: Vowel recognitions across all gates according to long-domain coarticulation

Mauchly's test of sphericity shows that sphericity is violated for long-domain coarticulation, χ^2 (9) = 17.923, p = 0.037. However, there may still be an underlying trend in recognition with respect to long-domain coarticulation, as suggested by the differences in the exponency of CVNs (cf. figures 42-45 in section 4.2.3). The lack of a statistical estimate remains an artefact of the lack of an available test in the context of this study rather than necessarily implying that the trends evidenced are perceptually insignificant.

4.4 Differences in Vowel Recognition Relating to Nasality

In this subsection we will examine effects for vowel types in CVNs. The results for how recognition evolves *through time* with respect to vowel quality and nasality can be found immediately below figure 50 (cf. figures 51-54), this for both CVNs and CVCs. Figure 50, which shows the *overall results* for different CVNs and CVCs, is organised as follows: the percentage proportions of correct responses for each vowel can be found on the x-axis. The blue bars display CVNs, while the

red bars show results for CVC monosyllables. Figures 51-54 display the results across the 10, 20, 30 and 40ms gates, for CVNs (left-hand bars) and CVCs (right-hand bars):

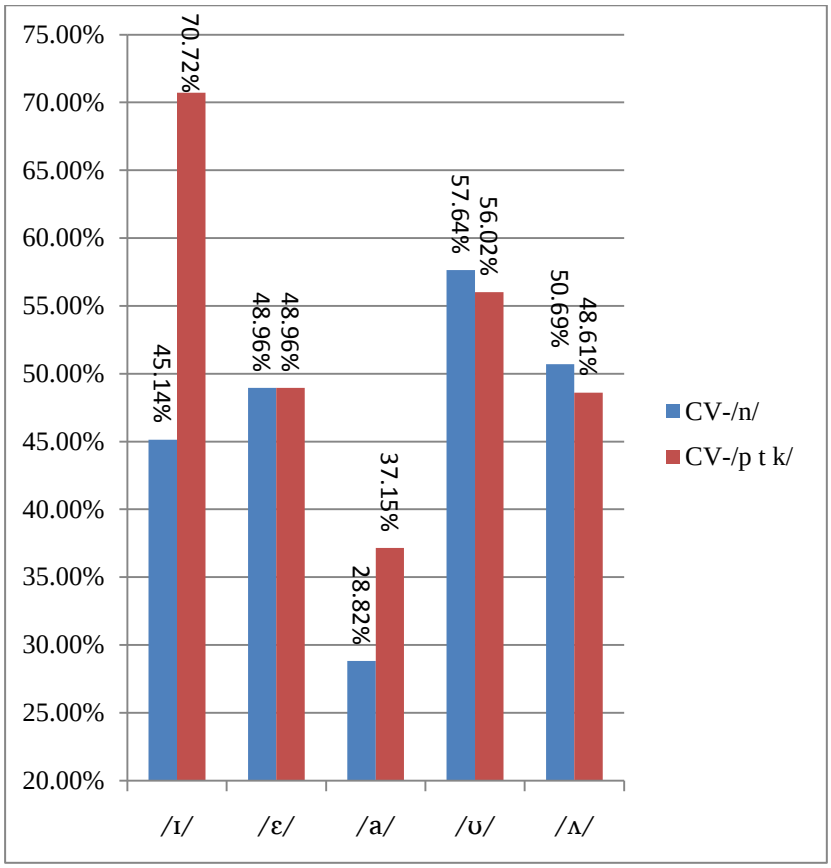


Figure 50: The overall effect of vowel quality on recognition in [+ nasal] and [- nasal] rimes

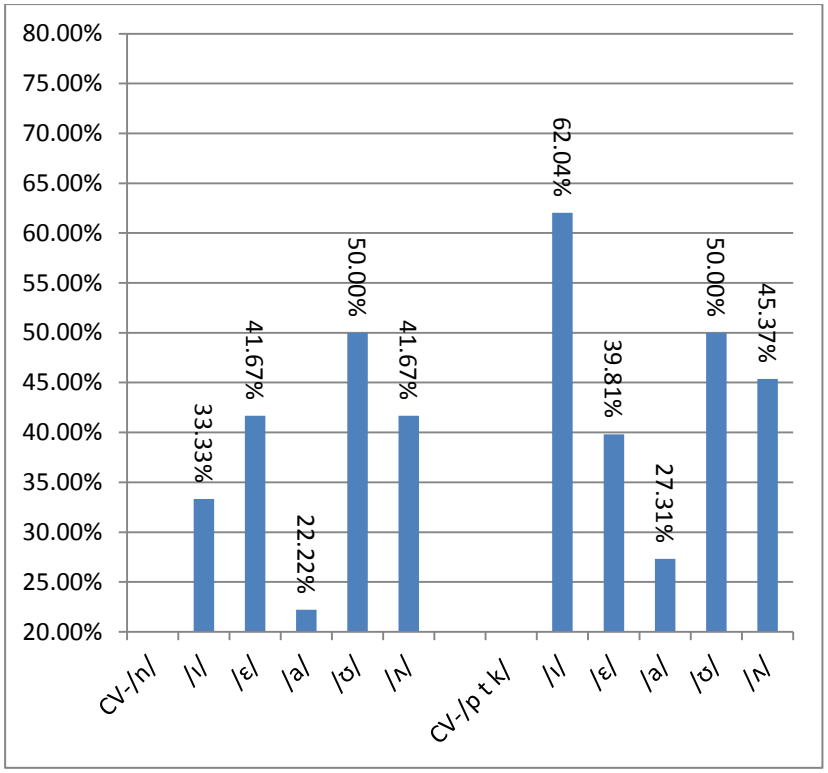


Figure 51: The effect of vowel quality on recognition in [+ nasal] and [- nasal] rimes (10ms gate)

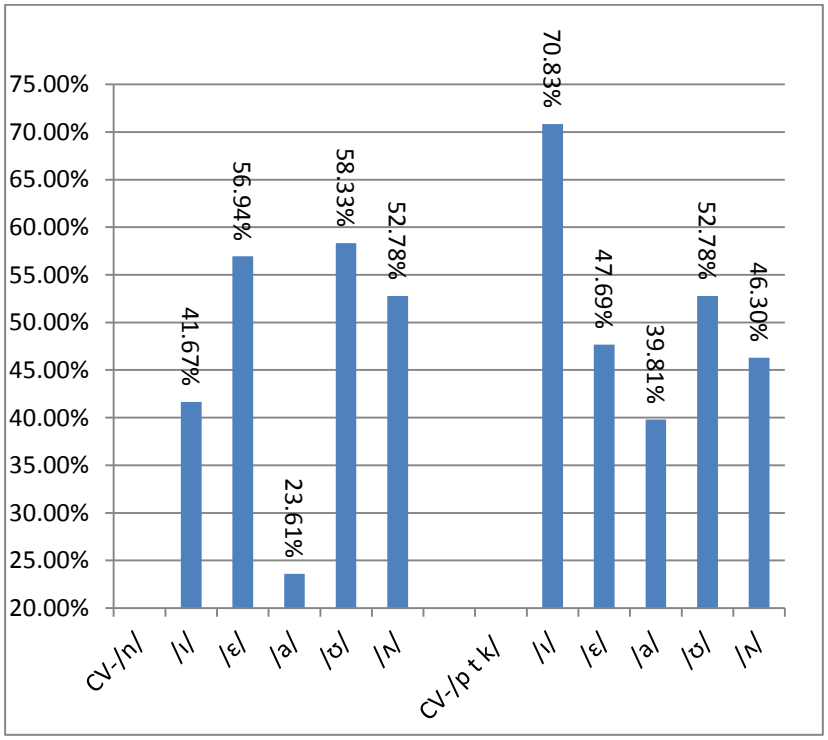


Figure 52: The effect of vowel quality on recognition in [+ nasal] and [- nasal] rimes (20ms gate)

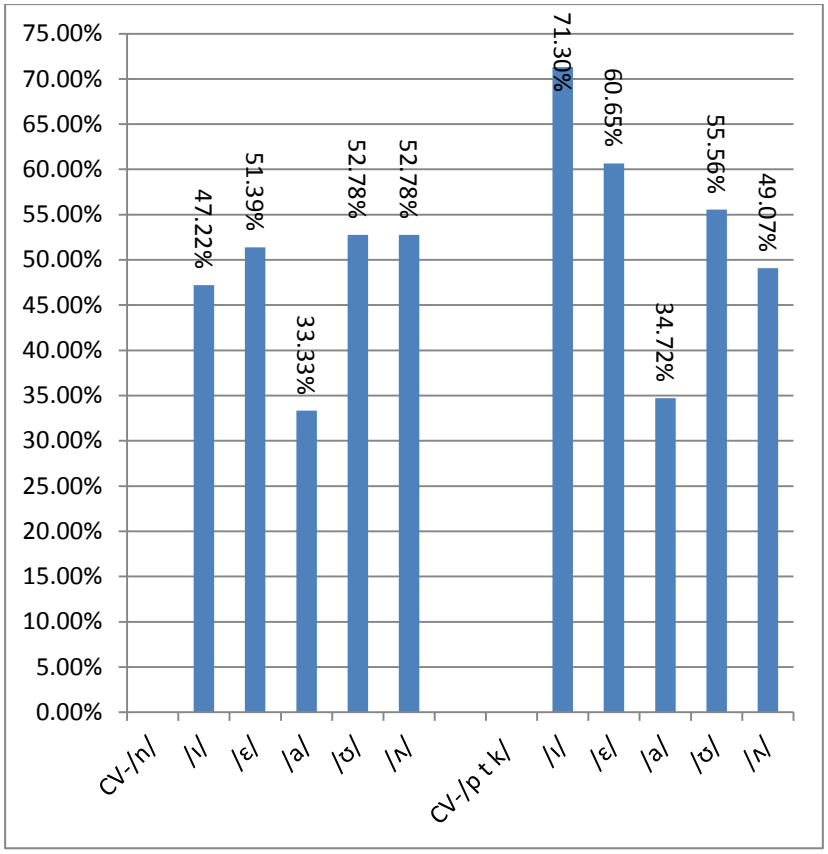


Figure 53: The effect of vowel quality on recognition in [+ nasal] and [- nasal] rimes (30ms gate)

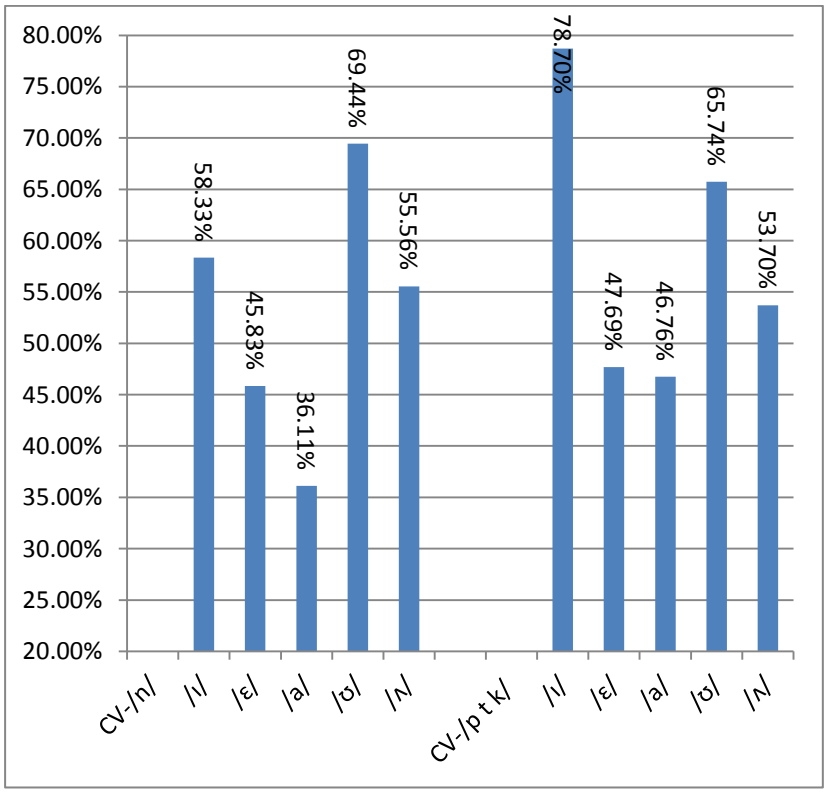


Figure 54: The effect of vowel quality on recognition in [+ nasal] and [- nasal] rimes (40ms gate)

Since /ʌ ʊ/ occurred in complementary stimuli in northern and southern speakers' productions, it was chosen not to compare the results for /ʌ ʊ/ with those for /ɪ ɛ a/. This deficiency does not constitute a theoretical problem, since there is no previous research for nasality's influence on back vowels with respect to height. All the comparisons here relate to /ɪ ɛ a/ (however, also see discussion on /ʌ ʊ/ in 5.4.7).

The results for /ɪ ɛ a/ are normally distributed, χ^2 (2) = 1.895, p = 0.388, Overall, the differences in recognition are significantly different, F(2) = 10.610, p < 0.02. Individual pairwise comparisons show the results for /ɪ/ to differ significantly from those for /a/, p < 0.03. The pairwise comparison for /a/ contra /ɛ/ displays a statistical tendency, p = 0.078.

4.5 Perceptual Confusions and Vowel Length

In this subsection we will look at how perceptual confusions may affect recognition depending on the similarity/dissimilarity of a given vowel relative to the one heard by a listener. Long and short vowels are examined separately, since they occur in different syllable types and have qualitatively different phonetic exponents. The results in tables 7-16 are colour-coded with boxes in green indicating reliable recognitions (ca. 60% or more). Relatively reliable recognitions are shown in yellowish green boxes (showing values between ca. 35 and 60%). Pinkish-orange and pinkish-yellow boxes indicate the most common vowel confusors (between ca 35% and 15%), with red boxes denoting the least common ones (= values below ca. 15%). Boxes marked with # indicate word forms that were not

studied. Tables 7-16 are organised with given answers in columns and heard stimuli in rows (i.e. the values top-down constitute the individual confusion matrices). Tables 7 and 9-12 show long vowels, whereas tables 8 and 13-16 display short vowels.

4.5.1 Overall Values Across Time

Average vowel confusion proportions in the recognition of /iː uː ɑː ɔː/ at all gates

Vowel Quality	/iː/	/uː/	/ɑː/	/ɔː/
/iː/	75.35%	28.13%	18.06%	5.90%
/uː/	15.28%	48.61%	28.47%	38.54%
/ɑː/	6.60%	10.76%	32.99%	17.36%
/ɔː/	2.78%	12.50%	20.49%	38.19%

Table 7: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /ʊ ʌ ɪ ɛ a ɒ/ at all gates

Vowel Quality	/ʊ/	/ʌ/	/ɪ/	/ɛ/	/a/	/ɒ/
/ʊ/	54.58%	#	5.03%	3.82%	12.24%	20.83%
/ʌ/	#	49.13%	5.64%	4.17%	9.81%	18.98%
/ɪ/	17.08%	17.71%	64.32%	34.72%	30.47%	12.27%
/ɛ/	4.03%	7.29%	8.33%	48.96%	6.94%	#
/a/	10.69%	13.89%	12.15%	8.33%	35.16%	14.93%
/ɒ/	13.61%	11.98%	4.51%	#	5.38%	32.99%

Table 8: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Since the number of vowel tokens differs across short vowel categories, no statistical comparisons were made for confusions of short vowels.

For long vowels, Mauchly's test of sphericity shows no significant differences, χ^2 (5) = 3.625, p = 0.656. Vowel quality is not statistically significant F(3) = 0.577, p = n.s. Post hoc multiple comparisons for /iː/ /uː/ /ɑː/ and /ɔː/ indicate that there are no significant differences between recognition proportions for CVVs at different gate intervals. Listeners rely on acoustic similarity in making judgements on vowel quality.

4.5.2 An Examination of the Results between Gates

Long Vowels

Average vowel confusion proportions in the recognition of /iː uː ɑː ɔː/ at the 10ms gate

10ms	/iː/	/uː/	/ɑː/	/ɔː/
/iː/	65.28%	36.11%	19.44%	11.11%
/uː/	22.22%	51.39%	27.78%	31.94%
/ɑː/	8.33%	8.33%	36.11%	23.61%
/ɔː/	4.17%	4.17%	16.67%	33.33%

Table 9: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /iː uː ɑː ɔː/ at the 20ms gate

20ms	/iː/	/uː/	/ɑː/	/ɔː/
/iː/	70.83%	23.61%	19.44%	9.72%
/uː/	12.50%	52.78%	26.39%	43.06%
/ɑː/	11.11%	9.72%	33.33%	13.89%
/ɔː/	5.56%	13.89%	20.83%	33.33%

Table 10: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /iː uː ɑː ɔː/ at the 30ms gate

30ms	/iː/	/uː/	/ɑː/	/ɔː/
/iː/	79.17%	31.94%	15.28%	1.39%
/uː/	15.28%	34.72%	31.94%	48.61%
/ɑː/	5.56%	12.50%	36.11%	11.11%
/ɔː/	0%	20.83%	16.67%	38.89%

Table 11: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /iː uː ɑː ɔː/ at the 40ms gate

40ms	/iː/	/uː/	/ɑː/	/ɔː/
/iː/	84.72%	23.61%	18.06%	1.39%
/uː/	12.50%	52.78%	27.78%	30.56%
/ɑː/	1.39%	12.50%	26.39%	20.83%
/ɔː/	1.39%	11.11%	27.78%	47.22%

Table 12: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Short Vowels

Average vowel confusion proportions in the recognition of /ʊ ʌ ɪ ɛ a ɒ/ at the 10ms gate						
10ms	/ʊ/	/ʌ/	/ɪ/	/ɛ/	/a/	/ɒ/
/ʊ/	47.78%	#	4.86%	8.33%	12.50%	18.98%
/ʌ/	#	44.44%	8.33%	8.33%	8.33%	13.43%
/ɪ/	20.00%	20.14%	54.86%	36.11%	39.58%	17.59%
/ɛ/	7.22%	11.11%	8.33%	40.28%	7.64%	#
/a/	12.78%	9.03%	17.71%	6.94%	26.04%	19.44%
/ɒ/	12.22%	15.28%	5.90%	#	5.90%	30.56%

Table 13: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /ʊ ʌ ɪ ɛ a ɒ/ at the 20ms gate						
20ms	/ʊ/	/ʌ/	/ɪ/	/ɛ/	/a/	/ɒ/
/ʊ/	53.33%	#	4.17%	2.78%	10.42%	16.20%
/ʌ/	#	47.92%	5.21%	2.78%	11.11%	24.07%
/ɪ/	19.44%	17.36%	63.54%	34.72%	29.86%	12.96%
/ɛ/	1.67%	7.64%	9.38%	50.00%	5.90%	#
/a/	11.67%	16.67%	12.50%	9.72%	36.11%	19.91%
/ɒ/	13.89%	10.42%	5.21%	#	6.60%	26.85%

Table 14: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /ʊ ʌ ɪ ɛ a ɒ/ at the 30ms gate

30ms	/ʊ/	/ʌ/	/ɪ/	/ɛ/	/a/	/ɒ/
/ʊ/	55.00%	#	5.56%	2.78%	13.54%	26.85%
/ʌ/	#	50.00%	4.51%	2.78%	12.15%	22.22%
/ɪ/	15.56%	19.44%	65.28%	30.56%	29.51%	10.19%
/ɛ/	4.44%	3.47%	9.03%	58.33%	6.25%	#
/a/	11.11%	13.89%	9.72%	5.56%	34.38%	13.43%
/ɒ/	13.89%	13.19%	5.90%	#	4.17%	27.31%

Table 15: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Average vowel confusion proportions in the recognition of /ʊ ʌ ɪ ɛ a ɒ/ at the 40ms gate

40ms	/ʊ/	/ʌ/	/ɪ/	/ɛ/	/a/	/ɒ/
/ʊ/	62.22%	#	5.56%	1.39%	12.50%	21.30%
/ʌ/	#	54.17%	4.51%	2.78%	7.64%	16.20%
/ɪ/	13.33%	13.89%	73.61%	37.50%	22.92%	8.33%
/ɛ/	2.78%	6.94%	6.60%	47.22%	7.99%	#
/a/	7.22%	15.97%	8.68%	11.11%	44.10%	6.94%
/ɒ/	14.44%	9.03%	1.04%	#	4.86%	47.22%

Table 16: Proportion of correct and incorrect percepts for each vowel sound (produced stimuli in rows and heard stimuli in columns)

Tables 13-16 are organised similarly to the ones in 4.5.1. The findings for the gate intervals display the following results for long vowels:

At the 10ms gate, Mauchly's test of sphericity shows no significant differences, χ^2 (5) = 2.439 p = 0.816. Vowel quality is not significant F(3) = 0.000, p = n.s. Post hoc multiple comparisons for /iː/ /uː/ /ɑː/ and /ɔː/ indicate that there are no significant differences at gate number 1.

At the 20ms gate, Mauchly's test of sphericity shows no significant differences, χ^2 (5) = 4.045p = 0.600. Vowel quality is not a significant factor, F(3) = .000, p = n.s. Post hoc multiple comparisons for /iː/ /uː/ /ɑː/ and /ɔː/ indicate that there are no significant differences at gate number 2..

At the 30ms gate, Mauchly's test of sphericity shows no significant differences, χ^2 (5) = 5.225, p = 0.455. The factor of vowel quality is not significant F(3) = .000, p = n.s. Post hoc multiple comparisons for /iː/ /uː/ /ɑː/ and /ɔː/ indicate that there are no significant differences at gate number 3.

At the 40ms gate, Mauchly's test of sphericity shows no significant differences, χ^2 (5) = 4.297, p = 0.567. The factor of vowel quality is not significant F(3) =.000, p = n.s. Post hoc multiple comparisons for /iː/ /uː/ /ɑː/ and /ɔː/ indicate that there are no significant differences for confusion effects at gate number 4. The results are similar at all four gate intervals.

Overall, these results strongly suggest that listeners rely on acoustic similarity between vowel types in making judgements in recognition. There are no systematic perceptual biases with respect to vowel confusions.

On the whole, the way in which listeners confuse different vowel types for each other does not change significantly over time. The proportion of confusions for vowel qualities *other than* the vowel actually heard remain fairly constant. Although statistical tests were not performed for short vowels with respect to perceptual confusions in this research (due to the difference in stimulus numbers across categories), the values evidenced in tables 13-16 for each of the six short vowel types are quite similarly spread across acoustically similar vowels. For example, /ʊ/ is always readily confused with /ɪ ɒ/ (whose formant relationships may be quite similar

to mid-high back vowels in English varieties), whereas /ʊ/ is relatively rarely responded to as /a ɛ/.

When we look at the results in tables 7-16 in detail, we can still see that front vowels are more likely to be confused for front vowels than for back ones and vice versa, and frontness can be seen to take precedence over height in the ways in which confusions occur. For instance, /iː/ is much more likely to be recognised as /uː/ than as /ɑː/ or /ɔː/, while /ɪ/ is a common confusing option for /a/ (and vice versa), a result which is not true for e.g. /ɪ/ contra /ʊ ɒ/. Listeners find it straightforward to discriminate for frontness, while height is more difficult to recognise correctly. In sum, whichever vowel is considered at a given point in time, the degree of acoustic similarity of any given response option to the vowel actually heard has the strongest bearing on what a listener's final response will be. Having described aspects relevant to confusion effects connected with vowel recognition, we will now describe aspects of recognition relevant to lexical frequency.

4.6 Lexical Frequency

This section shows to what extent lexical frequency affects the recognition of monophthongs. The choice of splitting all the stimulus words according to their syllabic shapes (CVVs contra CV-/p t k/ shapes contra CVNs) in this subsection can be partly justified on presentational grounds and by the fact that the FPD of these syllable shapes often differs significantly, and may affect the timing of vowel recognition (see e.g. the results in subsections 4.2.3 and 4.4). Since the previous subsection shows that the proportions of vowel responses are quite evenly spread across different vowel sounds, we will not compare lexical frequency in such a respect here. Given this result, the

only possible variable where lexical frequency might be playing a role in vowel recognition will relate to the structural aspects of CV(V)/Cs

For all stimulus categories examined and tested at all four gate intervals, linear regression analyses were used to ascertain whether frequency significantly affects vowel recognition. Since the listeners were also exposed to the written forms of stimuli (not just auditory forms) and much of human perception is visual (cf. e.g. Goldstein, 2013), the written frequencies) listed in the British National Corpus (BNC) are examined in this subsection (as opposed to spoken ones). The BNC is the largest British English corpus of spoken and written texts. Figures 55-59 are organised as follows: correct recognition proportions for different vowels can be found on the y-axes, with lexical frequency on the x-axes. The trendlines represent the general recognition tendencies for each syllable shape. For example, for CVVs in figure 55, we can see that on an average the more frequent a stimulus, the more often it is recognised correctly. The results of the regression analyses for each syllable shape are reported below figures 55-59:

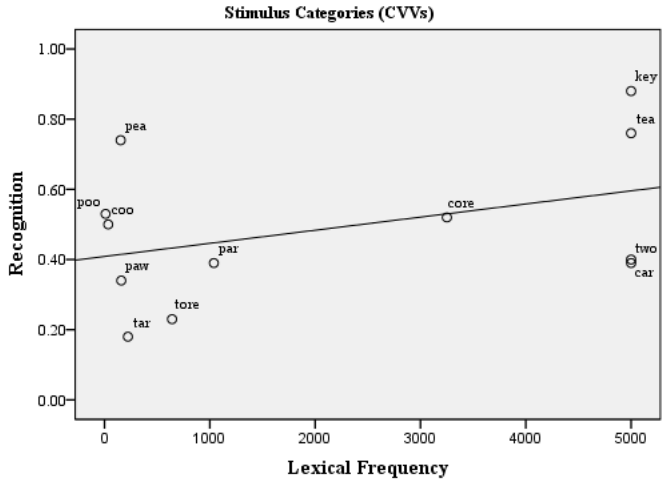


Figure 55: Vowel recognition according to lexical frequency in all CVV stimulus tokens

a) For CVVs frequency explains 40.1% of the variation in recognition, R^2 =.401, F(1,10) = 1.913, p = n.s. It was found that frequency did not significantly predict recognition of CVVs. Lexical frequency does not have strong links with vowel recognition in CVVs.

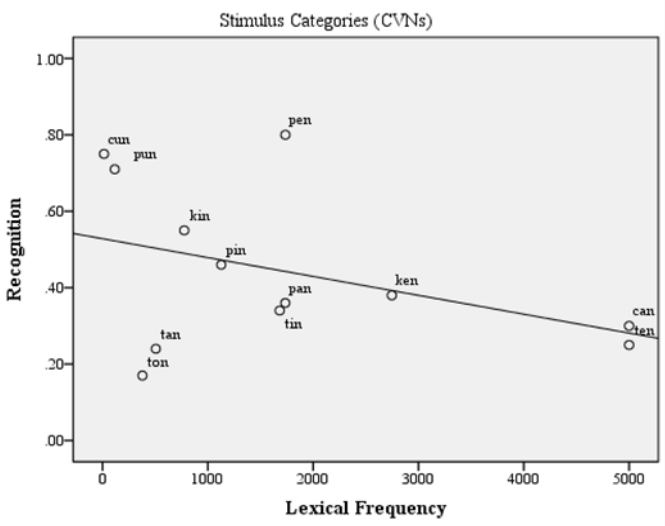


Figure 56: Vowel recognition according to lexical frequency in all CVN stimuli

a) For CVNs, frequency explains 39.9% of the variation in recognition, R^2 =.399, F(1,10) = 1.892, p = n.s. It was found that frequency did not significantly predict recognition of CVNs. Lexical frequency does not share a close relationship with recognition in CVNs.

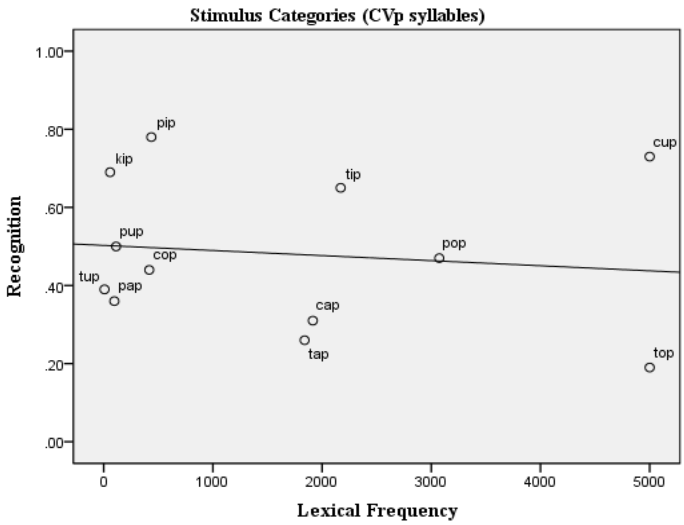


Figure 57: Vowel recognition according to lexical frequency in all CVp syllables

b) For CVp syllables, lexical frequency explains 12.5% of the variation in recognition, R^2 =.125, F(1,10) =.158, p = n.s. It was found that that frequency did not significantly predict recognition of CVp syllables. The potential link in this instance is very weak considering the low value of the R statistic. There is no link between frequency and recognition for CVp syllables.

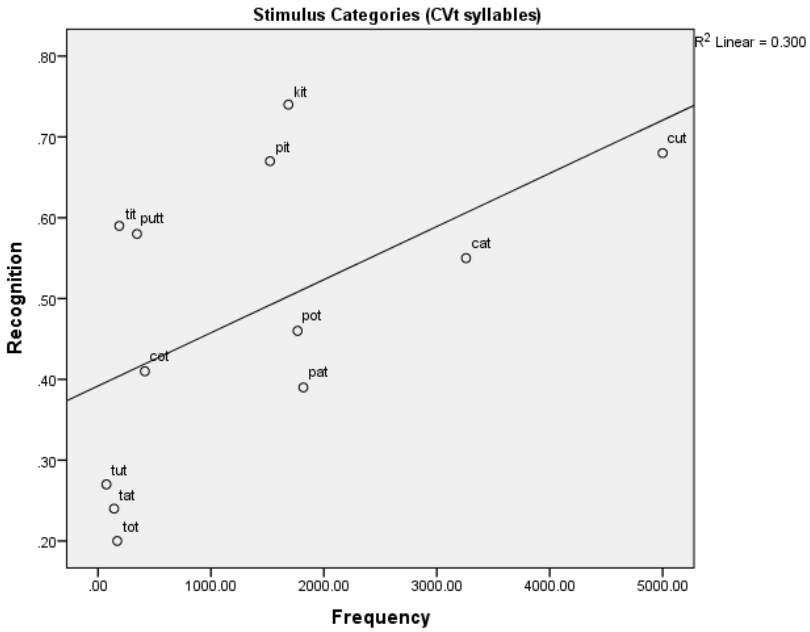


Figure 58: Vowel recognition according to lexical frequency in CVt syllables

c) For CVt stimulus tokens frequency explains more of the variation in recognition than for other syllable shapes: $R^2 = .547$, $F(1,10) = 4.277$ $p = 0.065$. It was found that that frequency did not significantly predict recognition of CVt syllables. Since the p value is indicative of a statistical tendency, there may still be some kind of link between recognition and frequency for CVt syllables (bearing in mind that more than half of the variation in recognition values is explained by the model). The results can be seen to suggest that, to some extent, the more frequent the CVt, the more reliably it is recognised by listeners as well (cf. the trendline in figure 58).

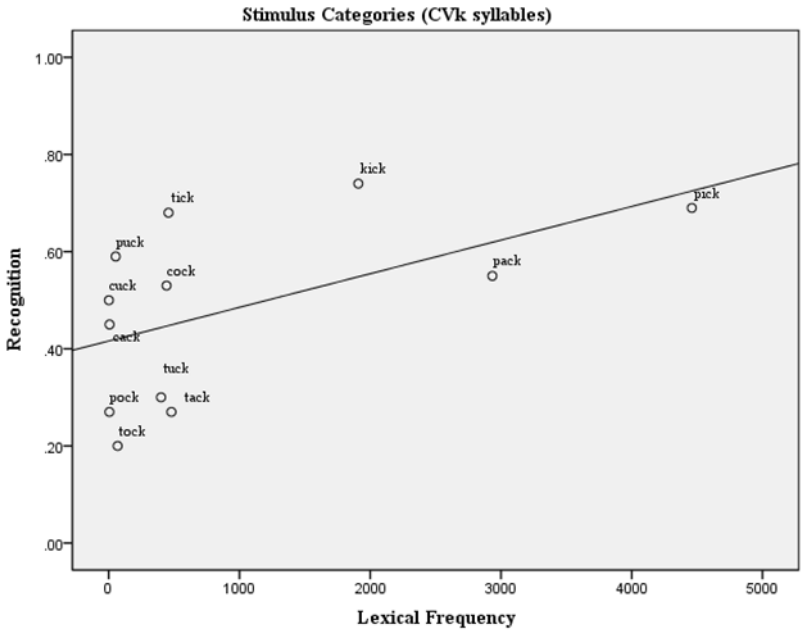


Figure 59: Vowel recognition according to lexical frequency in CVk monosyllables

d) For CVk stimulus tokens frequency also explains somewhat more of the variation in recognition compared to CVV, CVN and CVp syllables: R^2 = 538. F(1,10) = 4.083 p = 0.071. It was found that frequency did not significantly predict recognition of CVt syllables. Since the p value for CVk is only slightly higher than for CVt tokens and the correlation statistic is similar, there might be a link between recognition and frequency for CVk syllables. The results give some indication for the suggestion that the more frequent the CVk, the more reliably it is recognised by listeners (cf. the trendline in figure 59).

In summary, the results on lexical frequency offer little support to lexical frequency being an important perceptual variable in this kind of an experiment, since none of the five comparisons is significant (with three of them having p values

of 0.2 or above). This aspect of perception seems particularly relevant from the perspective of the way in which recognition evolves temporally and with respect to the bidirectionality of coarticulation (cf. the different syllable types displayed in figures 55-59). Acoustic similarity between vowels is a much more important factor in vowel recognition than lexical frequency, as the results displayed in tables 7-16 confirm (= phonetically similar responses to a heard stimulus are strongly preferred in recognition).

Nevertheless, for some of the most frequent syllable shapes in speech (like CVt and CVk syllables), some kind of link between recognition and frequency might exist in terms of a) the time course of recognition and b) how lexical frequency might be related to stimulus structure. We will now summarise the results displayed in this chapter in the next subsection.

4.7 A Summary of the Results Presented in Chapter 4

This subsection summarises the results presented in this chapter from the viewpoint of production (4.7.1) contra perception (4.7.2). The final subsection of this chapter revisits the hypotheses presented in chapter 2 by showing whether and to what extent they have been supported.

4.7.1 Production Results

In summary, the evolution of formant values differs for low, mid and high vowels, so that F1, F2 and F3 are most variable for high vowels in spectro-temporal terms across time. There is more moment-to-moment variation in the centre frequencies of the first three formants for high vowels than for mid and low vowels. This result applies to all three main syllable shapes

examined (CVVs, CVNs and CV-/p t k/ syllables, cf. figures 26-45).

For bidirectional variation in vowel recognition, no significant effects are evidenced for /p t k/-V-/p t k/ syllables: the potential phonetic co-extensiveness between coda plosive portions and plosive onset portions do not significantly affect vowel recognition. However, CVNs display significantly more variable vowel formant trajectories compared to CV-/p t k/ monosyllables and especially CVVs. This result is evidenced despite CVNs having short vowels with less variable VISC patterns than CVV syllables, which have long vowels (cf. figures 42-45).

4.7.2 Perception Results

On an average, recognition becomes very significantly more reliable between 30ms and 40ms into the aperiodic phase, showing that as listeners hear more of the vowel *at a specific point in time*, recognition becomes much more reliable (cf. figure 46).

Onset place of articulation comprises a very highly significant cue to recognition, with alveolar plosive onsets trailing bilabial and velar onsets in this respect (cf. figure 47).

Vowel quality has an effect on recognition, so that high vowels tend to be recognised more reliably (and earlier) than mid and low vowels. Low vowels are sometimes recognised at near chance level during the early part of the aperiodic phase (cf. figure 48). In sum, despite there being more moment-to-moment variation in VISC for high vowels in terms of production, the longer time that is needed to move the jaw from its neutral position to produce mid and especially low vowels has an overall negative effect on the time course of perception.

Low and mid vowels take somewhat longer to recognise compared to high vowels.

The individual vowel results for CVNs need to be teased apart in order to compare the results for coda quality in terms of exponency as well as perception. This claim is particularly true for CVNs, whose recognition timing properties differ depending on the phonetic quality of the nucleus (cf. figures 50-54). For stimuli with nasal codas, the temporal co-extensiveness between parts of the onset and coda is dependent on vowel quality and the overall articulatory constellation behind a given CVN. For example, the southern and northern variants of the 'pun' vowel and also 'pen' engender much more reliable levels of recognition than 'pin' and 'pan'. It is important to bear in mind the limitations of this claim with respect to 'pun', whose exponents differ in northern and southern accents (meaning that no statistical comparisons were made).

In terms of distinguishing different vowel categories, the results displayed in 4.5 suggest that the more acoustically similar the vowel heard to a given response option, the more likely that response option will be a listener's final response choice for a CV(V)/C (cf. tables 7-16). Since listeners were not asked to distinguish between long and short vowels (see subsection 3.5.3), this result lends good support to the claim that listeners tend to select choices that are phonetically similar to the heard stimulus than choices which are auditorily more distinctive. For example, [iː] is more likely to be recognised as /uː/ or than as /ɑː/ or /ɔː/ given a choice between e.g. the four words 'pea-poo-paw-par'.

Lexical frequency does not have significant effect on recognition, though CVt and CVk syllables come close to reaching statistical significance in terms of recognition (cf. figures 58-59). The effects observed may be systematic in this

instance, but almost certainly not perceptually significant across *all* stimulus types. Having presented and statistically analysed all the results of this study, we will move on to discuss them in the next chapter.

5. Recognising and Building Representations for Vowels through Time

5.1 Overview

The aim of this chapter is to discuss the results presented in chapter 4 by presenting a workable model for how vowel recognition evolves through time, and which recognises that "there is no single applicable unit that can faithfully mimic the rhythmic-temporal organisation of speech" (Local and Ogden, 1997, p. 110). This issue is relevant because it reinforces the idea that subtle FPD can significantly affect perception of vowel timing and the temporal dynamic correlates of vowel sounds.

It is important to show at the outset of this chapter to what extent the results presented in the results chapter extend our understanding of vowel recognition timing. It will also be described to what extent the aims of the research have been fulfilled. The structure and contents of this chapter is described as follows: 5.2 describes the relationship between the research questions, main aims and hypotheses outlined in chapters 1-2 and the results presented in chapter 4 on the one hand and the findings of previous literature on the other (see especially 2.2). 5.3 shows that the phonological modelling of vowel recognition requires perception to be simultaneously relative to many levels and elements of phonological structure, such as syllables, onsets, rimes, nuclei and terminal nodes. This dependency of recognition on the overlaying of consonants upon vowels (e.g. Coleman, 1998 and Öhman, 1966) is exemplified through feature sharing between abstract phonological categories and phonological rules underlying the recognition of syllable constituents, such as onsets and rimes. Since phonological processing comprises a mirror image of the non-segmental definition of coarticulation in speech production given in 1.3.1,

5.3 shows step-by-step how the processing of representations from time-varying input shares a direct relationship with the way recognition evolves temporally. The dependency of phonological processing on the FPD of the aperiodic phase and properties of the acoustic input is given an explicit statement. The way in which representations for vowels and other syllable constituents are updated through time is illustrated. Vowel recognition is shown to i) mirror the acoustic input in the same way that the acoustic output mirrors production and ii) that such modelling is directly dependent on having a sufficiently broad view of coarticulation and the declarative rules underlying feature sharing.

A very important aside in the context of 5.3 is that the relevance of the statements on abstraction rules, feature sharing and representations in 5.3-5.4 are closely tied to the findings on vowel length, and in particular what level in the syllable it should be represented at. Since the perception of vowel quality and its time course are sensitive to the exponency of vowel length, it is possible to highlight the perceptual significance of spectro-temporal variation in VISC. Since the perceptual input from long and short as well as high and low vowels differs, the dynamic variation related to VISC is reflected in the time course of recognition. This variation can therefore be seen to affect the representation of vowel length phonologically.

As in the previous chapter, the results are discussed separately for males and females. To the extent that the underlying trends for males and females are not statistically significantly and/or qualitatively different (see 4.2), 5.3.6 rounds up the perception/production timing model by outlining some similar general conclusions we can draw for stimuli as produced by female speakers. The reason for including this section on female stimuli is related to allowing us to have as comprehensive a picture of vowel recognition timing from CV(V)/Cs as possible. To a lesser extent, the section also

highlights some of the phonetic differences relating to aspiration and vowel resonances in female stimulus productions (which mainly relate to the difficulties in measuring F1 reliably). Thus, the reasons for including subsection 5.3.6 relate more to methodological issues and aspects of illustrating the main trends for vowel recognition comprehensively rather than assuming that recognition from male and female stimuli differs.

Subsection 5.4 presents a discussion of how the abstraction rules presented in 5.2 can be applied to recognition from CVNs. It is shown that the type of spectral distortions related to the exponency of vowel height in CVNs can be related to previous findings on perception by Hawkins and Stevens (1985), as well as related findings by Cohn (1990) and Chang et al (2011) on production. The conclusions on nasality in this chapter thus reflect the results described in 4.4 and 4.7. The findings for CVNs are secondary compared to those for length, since they also reflect the underlying syllable structure. In sum, this chapter will a) show how the results presented in the previous chapter extend our understanding of the perception of coarticulation and b) propose a model for solving specific problems relating to phonetic exponency in vowel recognition and apply the model as an example in context.

Thus, before commencing the analysis and applying key findings to the model, we will consider in 5.2 to what extent the main findings extend those of previous similar studies and in what ways the aims of this research have been fulfilled.

5.2 Extending Our Understanding of the Perception of Coarticulation and Vowel Recognition

5.2.1 A Re-examination of the Hypotheses Presented in Chapter 2

In summary, the following hypotheses were drawn in section 2.6 on vowel recognition timing and the structural as well as

phonetic aspects related to this phenomenon. Each hypothesis is summarised briefly first. We then say whether and to what extent each hypothesis is supported.

a) It was hypothesised that listeners will achieve reliable vowel recognition 30ms subsequent to plosive release. Since recognition becomes significantly more reliable *between* the 30 and 40ms gates, this hypothesis (although giving a fair approximation of the time course of recognition) is not fully supported.

b) It was suggested that acoustic similarity between heard stimuli and the resulting percepts is the most important criterion in recognition (rather than e.g. lexical frequency or the structural aspects of the syllable). The results in tables 7-16 confirm that this hypothesis is supported.

c) The third hypothesis made the prediction that carryover and anticipatory coarticulation both significantly affect vowel recognition, and may have simultaneous effects on recognition timing. Since the results presented subsections 4.2.2-4.2.3 and 4.4 on the productions of different stimulus structures (CVN/CVC and CVV) and vowel recognition in CVNs both confirm this hypothesis, it receives good support.

d) It was hypothesised that alveolar onsets give rise to significantly fewer correct vowel responses than either bilabials or velars (with no significant difference between the latter two). Since the results and statistical tests displayed in subsection 4.3.2 affirm that onset place of articulation may be a very highly significant factor in vowel recognition, this hypothesis is supported.

e) The fifth hypotheses related vowel quality and vowel features to the time course of vowel recognition, so that short vowels are recognised earlier and more easily than long ones. Since the results displayed in subsection 4.3.2 show that for 3 out of 4 long vowels (/iː/ forms an exception due to its very low F1 value as a high vowel), recognition levels are lower than for 4 out of 6 short vowels, this conclusion receives relatively good support. The fact that the remaining two short vowels /a ɒ/ did not lead to more reliable and earlier recognition than the other four short vowels /ɪ ɛ ʌ ʊ/ can be explained by the fact that they are [-high] and will thus be more variable in VISC than high vowels like /iː ɪ/. It is important to bear in mind here that the inclusion of CVNs in this study will often have negatively affected the time course of recognition for short (i.e. early vowel recognition is delayed for short but not for long vowels). The fact that long vowels (especially /iː/) are more peripheral than short ones may also have influenced this result. For these two reasons and despite the small contradictions related to this result, the hypothesis can also be seen to receive some support from a broader theoretical viewpoint.

f) The sixth hypothesis made predictions on vowel recognition timing with respect to vowel height. High and especially high front vowels were hypothesised to be recognised more reliably than mid and especially low ones. The results presented on the recognition of individual vowel sounds in section 4.3.2 strongly support this hypothesis, since (on an average), high vowels are recognised more reliably and earlier than mid vowels, which in turn are recognised correctly more often and earlier than low vowels.

g) The final hypothesis made a specific prediction about the relationship between nasalisation present in a stimulus and the timing of vowel recognition. It was hypothesised that increasing presence of nasality during the aperiodic phase will distort listener ability to recognise the vowel correctly. This hypothesis is mostly supported. However, given the results displayed in subsection 4.4 on different types of CVNs, vowel quality *can* be a complicating factor with respect to how early listeners can recognise vowel from nasalised aspiration (but not always). The [+ back] nasalised vowels /ʊ ʌ/ engender much more reliable vowel recognition than the [- back] /ɪ ɛ a/. Therefore, the overall articulatory constellation behind a CVN may have different implications for vowel recognition timing.

How are we to account for the kinds of results described in this chapter in nonsegmental terms and with respect to phonological processing? How do the results displayed in this chapter extend the findings of previous research on vowel recognition from plosives and studies on vowel timing/VISC? How do the results align with the aims set out in chapters 1-2? The next section describes a) the relationship between the findings displayed in this chapter and previous research, as well as how the aims align with the findings.

5.2.2 Reconciling the Aims and Results of this Study

Firstly, the main aim of this study has been to act as a springboard for further research into coarticulatory and phonological processing and especially to extend previous findings on these two areas concerning vowel recognition. The two main findings on length and long-domain coarticulation in CVNs highlight the relatively limited understanding of speech processing in conventional theories. Moreover, this suggestion

also applies to radical non-segmental polysystemic research studies, albeit to a lesser extent. The theoretical stances in both strands of research (see e.g. Browman and Goldstein, 1986 and Liberman and Mattingly, 1985 contra Coleman, 1998 and Hawkins, 2003) seem too polarised, given the results of this study. Given our limited knowledge of the underlying representations and neural processes relevant to speech production and perception, a more fruitful direction for future research would be to move away from the phoneme vs. prosody and articulatory/gestural vs. auditory perception debates, for example. Instead, it will be better to focus on what the interaction between different linguistic systems can tell us about coarticulation and its relationship with phonological processing and FPD.

Secondly, the other main aim of this research has been to steer away the debate from exercising complete control over different variables in experimental linguistic research. For example, when working with real words and/or online experiments, we will *always* have to accept at least some degree of uncertainty in how precisely a given set of findings actually represent a linguistic feature or phenomenon. If we attempt to exercise control over everything, we limit the conclusions we can draw, risk alienating the wider linguistic audience and/or make claims that may have little bearing on the questions we are attempting to answer (cf. e.g. Tekieli and Cullinan (1979) and Cullinan & Tekieli's (1979) claims on the *phonemic* value of duration in *nonsense syllables*).

5.2.3 Main Findings

The main aim of this research is to explore the timing patterns associated with the recognition of English monophthongal vowels from plosive onsets. The main theoretical goal is to show how a non-segmental polysystemic view of the

perception of monosyllabic word forms can provide us with helpful insights into the perception of coarticulation and vowel recognition in the following respects:

a) the perception and representation of vowel timing
b) coarticulatory and listening strategies, *and*
c) the coarticulatory properties of complex sounds.

The ways in which variation in FPD within the early part of the aperiodic phase of a plosive onset can exemplify articulatory and related phonetic aerodynamic differences in the realisation of English CV(V)/Cs enables us to account for the main two new significant findings of this thesis:

i) syllable shapes with less phonetically complex vowel sounds such as short and high vowels engender earlier vowel recognition (because their moment-to-moment variability is less extensive spectro-temporally).

ii) lack of spectral indication of oncoming nasalisation within the early part of the aperiodic phase lends itself to more reliable recognition earlier in time.

Although the finding on CVNs is secondary, it has implications for the main finding on cues to length (high vowels being shorter than low ones), since nasality *can* serve to delay recognition. The main secondary finding shares a close theoretical connection with the main finding of this research.

Before outlining the contributions to knowledge in this thesis, we will consider a general theoretical and a methodological issue related to the findings on vowel length. VISC, which reflects the phonetic encoding of vowel length, has a significant bearing on recognition and its time course.

The findings in this research that are related to this phenomenon stand in contrast to those of similar previous studies, which make no mention of the importance of moment-to-moment variation in vowel formant centre frequencies (see e.g. Cullinan and Tekieli, 1979, Winitz et al, 1972 and Tekieli and Cullinan, 1979). Previous studies on the perception of coarticulation from English plosives pay little attention to the perceptual role of VISC in vowel recognition. However, at least some of the discrepancies relating to how duration affects recognition can be explained by the fact that most of the previous literature does not highlight the encoding of length and especially the fact that previous studies have not investigated differences between the perception of long and short vowels in any detail (only Winitz et al, 1972 use one short contra several long vowels). Instead, CVs with short vowel sounds rather than CV(V)/Cs have been the main focus of previous research on vowel recognition.

Another key point of discussion that needs to be highlighted before summarising and discussing the main findings is related to the temporal dynamic exponents of vowel height. This issue is particularly important from viewpoint of the temporal properties of vowel recognition and especially the phenomenon of VISC. High vowels require virtually no jaw movement. It will therefore be easier for listeners to recognise vowel quality earlier in high vowel contexts, as they are shorter in duration than low vowels. These two phenomena reflecting different phonological features share a close relationship in terms of recognising their time-varying exponents, both of which reflect VISC.

5.2.4 The Way Recognition Evolves Through Time

The more general contribution of this study can be described as follows: this study forms the first documentation of vowel

recognition from English aspirated plosives in varieties spoken in England. This research also comprises one of the first non-segmental studies on the perception of coarticulation. Previous studies on the perception of coarticulation for English have incorporated segmental-phonemic frameworks. There is no previous research on prosodic and/or non-segmental phenomena in vowel recognition from aspirated plosives for any variety of English.

One of the most important contributions to recognition in this study is the fact that although it is possible to represent vowel recognition in CV(V)/Cs as a whole in temporal terms, the phonetic encoding of phonological and syllable structure must be taken into account in investigating vowel recognition timing. The new main finding on length, according to which this feature should be represented at the highest syllabic node rather than at the nucleus level, shows that timing information on vowels in CV(V)/Cs is spread throughout the phonetic exponents of monosyllables. In other words, the finding shows how different properties on the different sounds in monosyllables are *always* to a certain extent intermingled with each other, regardless of the phonetic properties of the individual sounds present. Although this finding is potentially not very surprising from the viewpoint of research into coarticulatory phenomena, it shows some of the limitations of our knowledge of vowel perception timing and especially how it relates to phonological processing.

Most importantly, the dynamicity behind vowel articulations remains a key point in vowel recognition. This claim is particularly true for VISC, which varies spectro-temporally depending on length and vowel height. Although previous research (e.g. Nearey and Assmann, 1986) has shown that the underlying formant relationships for vowel sounds can often be reliably recognised ca. 30ms into a vowel sound, this research extends this finding to English varieties spoken in

England and also to aspirated plosives. However, the temporal locus point for reliable vowel recognition in this study is located slightly later in time than this (i.e. *between 30 and 40ms*). This difference between earlier research and the main finding suggests that the phonetic complexity of sounds in English varieties spoken in England and their coarticulatory strategies may delay vowel recognition somewhat more than in other varieties.

5.2.5 Contrast, Representation and Vowel Recognition

This thesis espouses the claim that both non-segmental and segmental representations and exponents may have a role to play in signalling contrasts and in phonological representation as well as phonological processing more generally. For example, the phonological model outlined in the next two major sections of this chapter is not considered as *the unparalleled solution* to vowel recognition. Rather, it reflects a non-segmental and polysystemic understanding of the results described in chapter 4.

Previous research on vowel recognition from aspirated plosives makes no mention about polysystemic or prosodic phenomena. These are two areas which this research delves more deeply into. However, phonological processing and representation do not simply reflect what sized/shaped perceptual targets listeners aim at. The discussion thus far strongly suggests that speech and vowel timing are more complex phenomena than previous research suggests.

5.2.6 FPD and Coarticulatory Direction Effects

This study offers the following answers to previous gaps in theory with respect to structural variation contra coarticulatory direction effects: i) it has been shown that anticipatory

nasalisation in CVNs has distinctive effects on recognition depending on vowel height ii) it has been shown in chapter 4 that the ways in which formant structure evolves through time may be relative to the structural aspects of a CV(V)/C (cf. figures 38-41). Together, these two findings show that structural variation related to the assignment of features to any given node in a CV(V)/C shares a close relationship with FPD and coarticulatory direction effects.

5.2.7 Phonological/Syllable Structure

Since recognition remains relative to different levels and aspects of structure, phonological and syllable structure together help to shape vowel recognition and its time course. For example, since English lacks stressed CV monosyllabic lexemes (with short vowels) but has CVVs, it is possible to account for the interdependence between phonetic interpretation of different aspects of structure and VISC. To account for vowel recognition timing in English CV(V)/Cs specifically, there is a need to highlight the coarticulatory FPD in CV(V)/Cs. The results shown in chapter 4 on CVVs and CVNs support this conclusion. In sum, adding a coda slot to a monosyllable does not necessarily complicate the listeners' task in vowel recognition. Rather, the main contrastive unit (the vowel), its structural specifications and phonetic complexity in relation to both the onset and coda steer recognition temporally.

5.2.8 Long-Domain Coarticulation and Airflow

This thesis has investigated whether a re-evaluation of the phenomenon of long-domain coarticulation is necessary within the context of coarticulatory and phonological theory. It is asked to what extent such aspects of coarticulation are reflected in the timing of vowel recognition.

The findings suggest that earlier models of long-domain coarticulation (see e.g. Hawkins and Nguyen, 2001, Goffman et al, 2008 and Cohn, 1990) do not go far enough in terms of showing how widespread phonetic influence from different parts of an utterance on other portions of the signal can be. Earlier models do not show in sufficient detail how the overall articulatory constellation in CVNs may serve to influence perception and production, something which this research does (cf. e.g. the results for nasalised vowels with different frontness and height values in chapter 4, figures 42-45). These two claims on long-domain coarticulation fit well together with the non-segmental framework of this study. Subsequent work on coarticulation should place a stronger emphasis on the interdependency between phonological representation, feature sharing and phonetic exponency in stimuli with complex phonetic exponents. Having outlined the findings of this study and their relationship to previous research and the aims of this thesis, we will apply the results to a non-segmental polysystemic phonological model of vowel recognition.

5.3 General Aspects of a Model of Vowel Recognition

In this subsection we will consider how listeners probabilistically abstract vowels from the acoustic signal and to what extent the temporal dynamics of vowel perception/production influence recognition in such terms. The primary focus of the model is on how phonetic information is distributed throughout the CV(V)/C syllable, and how this aspect of phonological processing allows listeners to abstract information for vowels and other accompanying constituents from time-varying phonetic exponents. This processing may be done largely in advance of the physical realisation of a given syllable constituent (such as an onset or a coda) based on the

coarticulatory information heard so far. Incremental information is seen to contribute to the time course of vowel recognition, since listeners may modify and update their abstractions based on the additional dynamic information received at a given point in time.

We will make some general observations and statements about the model before tackling the examples on the time course of recognition of CV(V)/Cs in the next subsection.

Before commencing the analysis, it is important to point out that what matters in the model espoused in this chapter are its general principles, not the precise details through which an abstraction is made. We now move on to discuss the phonological rules underlying abstraction from time-varying phonetic exponents.

The proposed model follows key aspects of declarative models (such as Polysp) in that listeners' awareness of properties of syllable structure allows projecting the phonetic properties of upcoming structures and constituents in advance of their physical realisation. This aspect of the temporal dynamics of CV(V)/Cs explains why listeners are often able to project vowel quality from monosyllabic utterances with a high degree of probability. The claims made in chapters 1-2 on feature sharing, as well as the fact that no single unit of timing is capable of fully accounting for the hierarchical organisation of speech timing (see Local and Ogden, 1997) are important buildings blocks for the phonological processing model

The model proposed follows a similar line of thinking to that in *YorkTalk* (see e.g. Coleman, 1992, Ogden, 1992 and Local & Ogden, 1997). Given the kind of exemplification of phonetic interpretation in previous non-segmental work (such as *YorkTalk*), we can model the frame around which listeners frame their abstraction rules as follows:

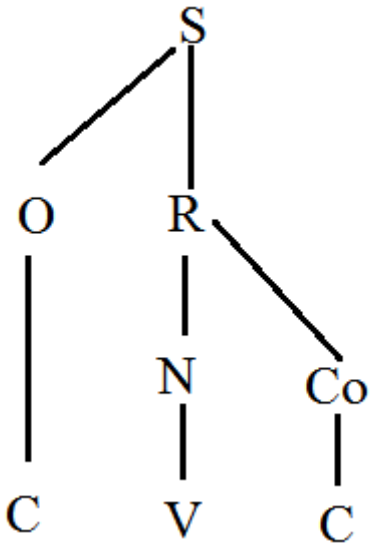


Figure 60: Syllabic tree for phonological abstraction

Listener recognitions of spectro-temporally varying speech signals can be seen to be comprised of abstractions over a specific type of syllable structure in English, such as the one described in figure 60 (see e.g. Coleman, 1998). Each constituent/node (e.g. onset and coda), which can be represented by graphs has a given number of features distributed over itself, such as [- voice] or [+ high]. The two nodes on the right (Co = coda and C = consonant) in figure 60 are optional and do not occur in CVVs such as ‘tea’ and ‘core’, as such word forms have no coda. The nodes on the left and middle stand for onset (O), consonant (C), syllable (S), Rime (R), Nucleus (N) and vowel (V), respectively. The only obligatory element is the head, which is represented as a vertical line in the bottom middle part of figure 60: only the vowel forms an obligatory category in the syllable.

Having described the syllabic frame around which feature sharing works, how does the listener go about building up representations like the one described in figure 60 from time-varying exponents? Since the kind of syllable structure described in figure 60 is here proposed as the frame around which speaker-listeners project phonetic information onto

abstract phonological representations, what deductions can we make about processing given the temporal co-extensiveness and overlaying of onset and coda exponents on the nucleus and vowel? Figures 61-67 and the accompanying commentaries clarify this problem:

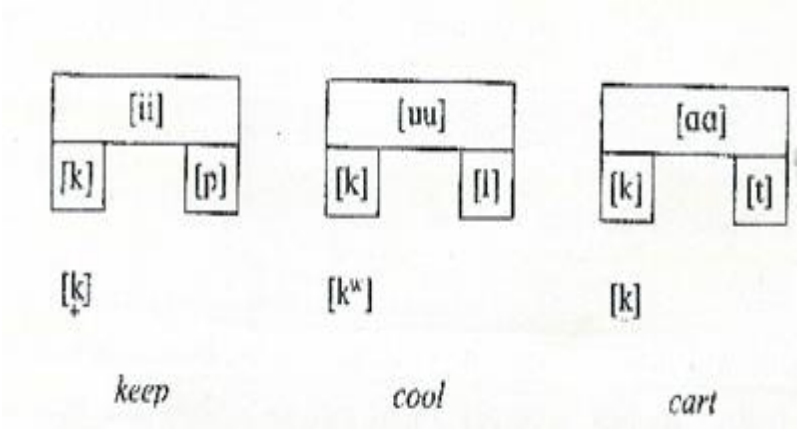


Figure 61: A coproduction exemplification of coarticulation in English CVC monosyllables (Coleman, 1992, p. 179), figure 5.4

Since the different sounds in a CVC are coproduced in the way described in figure 61 (with consonants being overlaid on vowels), the features between the nodes in the types of representations depicted in figures 60 and 61-63 can be modelled as being shared in phonological terms (= feature sharing). Figures 62-63 give an example of how this process functions in phonological processing and how it can be related to coproduction of plosive onsets with the vowels. Features can be modelled as being shared by different constituents. Figures 62-63 describe general phrase structure rules which can be used to relate phonetic input to abstract phonological representations. The phonological rules apply to the distributions of features in syllable nodes and are therefore applicable to both segmental and non-segmental models. The different colours denote headedness relations between different levels in structure. For example, green structures denote the syllabic (i.e. root) node, whilst red structures denote one of its

daughters, the onset. The rules in figure 60 are C → O and O → S (rather than vice versa) because the listener will know from having heard the carrier phrase + word stimulus that the first sound is a plosive consonant. S/he then needs to work out what its mother nodes (O → S, etc.) are, in order to arrive at a more a complete abstraction of each monosyllable.

C (+ cons) → **O (voice)**
(nasal)

C (+ cons) → **Co[18] (nasal)**

O (voice) → **S (long)**
(nasal) **(back)**
(high)
(round)

Figure 62: Step 1: projecting constituents in the syllabic tree: from daughters to mother nodes

S (long) → **O (voice) + R (voice)**
(back) **(nasal)** **(nasal)**
(high)
(round)

O (voice) → **C (+ cons)**
(nasal)

R (voice) → **N (Co[19])**
(nasal)

N → **(long) -> VV (back)**
(high)
(round)

Figure 63: Step 2: projecting constituents in the syllabic tree: mother nodes to daughters

[18] A coda (Co) remains a phonologically possible abstraction, since a *consonant* has been recognised by the listener.
[19] 'Co' is in brackets as it is not obligatory (Coleman, 1998)

The features [nasal],[voice], [long], [high], [back] and [round] at different levels in structure in figures 62-63 are not specified as '+, - or α' since they can have a feature present or absent, and a [-voice] onset can be [+ nasal] (i.e. the features need not agree). The rules exemplified in figures 62-63 relate to coarticulation in the sense that they enable listeners to project 'incomplete' acoustic input from time-varying exponents. This mutual dependency of perception upon input can be explained by feature sharing, which is exemplified in the rules described in figures 62-63, and which can be seen as a consequence of coproduction. Coleman (1990, pp. 14-15 and 1998, p. 179) has shown that coarticulation in CV(V)/C monosyllables can be modelled as coproduction, since "parametric[20] phonetic representations may be glued together in parallel, rather than simply concatenated" Coleman (1990, pp. 14-15).

The rest of this subsection shows two illustrations of how the listener goes about building up representations from 'incomplete' input for CV(V)/Cs from having heard only part of the onset portion. The illustrations in figures 64-66 are similar to the ones in figures 62-63, however they illustrate the temporal advance projection of syllable constituents in plosive-V(V)/Cs specifically, whereas the rules in figures 62-63 and 67 illustrate feature sharing more generally. A more general example is also included at the end of this subsection (cf. figure 67 and the accompanying commentary), which discusses and exemplifies feature sharing and temporal phonetic interpretation in more detail.

Having heard a transient around the moment of the plosive burst, the listener can deduce that the acoustic

[20] 'Parametric' is a term that relates to the temporal co-ordination of independent acoustic-articulatory parameters in a monosyllable.

properties heard thus far are consonantal. A consonant has been heard.

The following step in the abstraction process is to work out what the parent node of the consonant is (cf. figure 64). Is the consonantal portion just heard by the listener the daughter node of an onset or that of a coda? There is an ambiguity in working out what the parent node of the consonant is (i.e. O → C or Co → C) which is resolved by syntactic and phonetic detail. Listeners know from the syntax of the carrier phrase heard that a word is upcoming, which enables them to deduce that there will be a syllable. From the phonetics of the consonant, in turn, listeners will know that voiceless onsets will have longer closure durations than codas (e.g. Davis and Summers, 1989). For these two reasons, an onset is a more plausible abstraction than a coda:

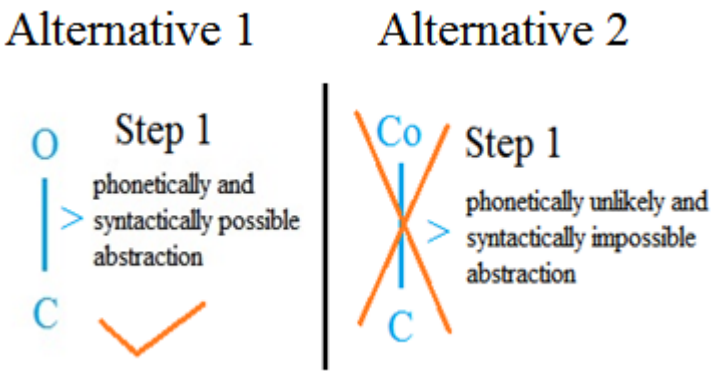


Figure 64: Abstraction step 1 in an English monosyllable

To briefly explain the thinking behind figure 64, we can say that since a new constituent cannot commence with a coda (Coleman, 1998), the listener can exclude such an abstraction. What else can listeners conclude from step one detailed in figure 62? As shown by e.g. Coleman (1998, p. 224), a syllable (S) must consist of an onset (O) and a rime (R). Listeners can work out the following from what is already available at time slot 1:

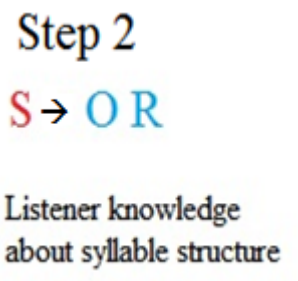


Figure 65: Abstraction step 2 for an English monosyllable

The red letter in figure 65 displays the mother node (syllable), which the listener can deduce from has been heard so far. The blue nodes show the daughter nodes, (i.e. the sisters O and R), which comprise the following steps in the abstraction process.

The listener cannot at this point in time be certain whether the upcoming syllable includes a coda, as the syllable commencing with an aspirated plosive can comprise a CVV, CVN or a CVC. Overall, the listener has arrived at the following abstraction just from hearing the transient at the plosive burst:

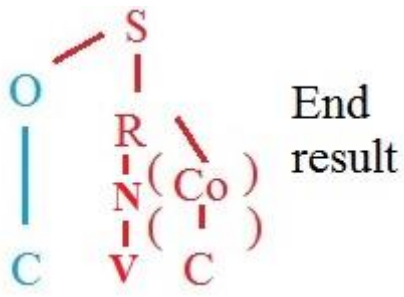


Figure 66: Abstracting new syllabic information from plosive onsets

The red and blue colours in figures 65-66 show the step-by-step processing that listeners perform by way of using the rules specified in figure 62-63, and in particular the fact that it is possible to in advance project the rest of the syllable constituents from 'incomplete' input. For example, having heard a new word commencing with a transient, the rules tell the listener that the consonant branches from the onset (O) node, and that the parent of the onset can only be the topmost

node, i.e. the syllabic node (S). Having established the rule-based interdependencies between these two parts of a syllable from the FPD available at the burst transient, the listener can recognise i) what vowel is being heard and ii) the structural properties of the syllable with a high probability.

This type of abstraction process demonstrates the perceptual significance of coarticulation for phonological processing, which is made possible through feature sharing (see e.g. Coleman, 1990, 1998 and Ogden, 1992). This process shows that listeners have rules for what types of syllable shapes are possible as abstractions given what has been heard so far.

Having exemplified the general principles behind phonological processing in the perception of coarticulation as well as those for CV(V)/Cs, we will round up this subsection by including an explicit statement of the relationship between temporal properties of input and phonological processing. Figure 67 illustrates this issue:

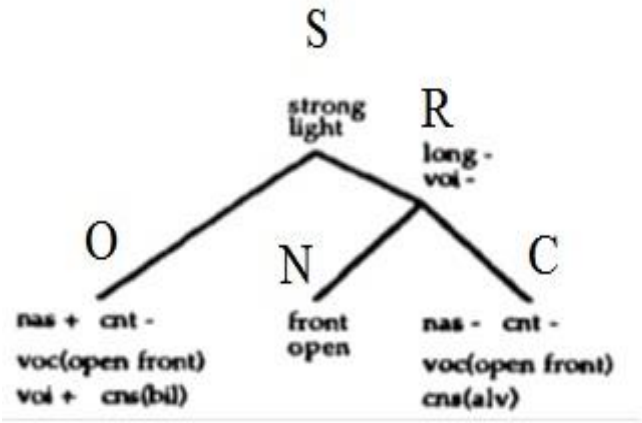


Figure 67: A partial phonological representation for ‘mat‘
(Ogden, 1992, p. 82, figure 1)

In figure 67, we can see a set of structured acyclic graphs representing the word “mat”, which have given features distributed over them, such as the [+ front] and [+ open]

features for the vowel. According to Ogden (1992, p. 82), given such a word having been produced by a speaker, we can say that listeners have a parser in their minds containing a grammar of English syllable, metrical and lexical structure. If not, words heard would be meaningless to listeners. Therefore, phonetic interpretation requires an explicit statement. Let us imagine that the onset portion [m] in /mat/ starts at time 0 while the coda portion [t] ends at e.g. time 350. Such values can be used as points of reference for the way in which phonological processing functions: the temporal co-extensiveness between the constituents in CV(V)/Cs may allow listeners to deduce the underlying abstract phonological categories and representations from partial input. The model described in this subsection contains abstract phonology, with structures and which has an explicit model of phonetic interpretation. We will use it in this chapter.

In sum, a model is needed that specifies the relationship between phonetic detail and abstraction using the kinds of <time, value> pairs exemplified for 'mat' in the paragraph immediately above. The 'time' parameter can in this chapter be equated with the moment of the plosive burst transient burst (time slot 1), while the gate intervals 10, 20, 30 and 40ms (time slots 2, 3, 4 and 5[21], respectively) can be considered as the 'values' that vowel recognition is projected from. In this sense, the 'time' parameter can be equated with the plosive burst, which is an anchor point for vowel recognition, whilst the gate intervals correspond to the 'value' points that reinforce the recognition of what can be projected in the perception of coarticulation from that anchor point.

[21] In the illustrations containing spectrograms and waveforms in this chapter, the four gates are referred to as t + 10/20/30/40ms, respectively, in order to facilitate the presentation.

We now move on to apply the rules and representations exemplified in this subsection to vowel recognition. The main goal is to show how perception is relative to feature sharing while also exemplifying the way in which recognition evolves through time. The example used here is /iː/ as in e.g. 'pea'.). 'Pea' also illustrates and exemplifies the production results detailed in 4.3 in terms of how vowel recognition evolves through time. The spectrograms in 5.3-5.4 thus necessarily reflect *individual* stimulus instances by a given speaker, whilst those in the previous chapter reflect production averages.

This approach allows us to generalise the results presented in 4.3 to vowel sounds in other word forms (e.g. 'paw-par-poo') and linking the FPD of production presented in the accompanying spectrograms in 5.3-5.4 with the way in which phonological-perceptual processing mirrors the acoustic output.

Both male and female production examples will be used to illustrate the relevancy of FPD for recognition. This strategy will allow for a comprehensive exemplification of the findings using particular examples of produced stimuli (as opposed to the production averages presented in chapter 4). However, since e.g. /iː/ remains the same vowel as heard from male and female productions, and since the production results in chapter 4 suggest that the qualitative distinctions between the two speaker types' productions do not differ significantly (see 4.2), we will only illustrate certain phonetic differences related to vowel recognition from stimuli as produced by female speakers.

Further consideration needs to be given to the illustrations given in the abstraction figures in this chapter. The figures describing what abstractions the listener should be aiming for (cf. e.g. figure 60) are coloured black. The other abstraction figures that describe recognition from the four gate

intervals (i.e. time slots 2, 3, 4 and 5) reflect the updates made to recognition by listeners through time using two distinctive colours in each abstraction figure: as the underlying vowel quality becomes more certain to listeners, the way in which recognition evolves through time is illustrated in figures 68-114.

5.3 Projecting Vowel and Syllable Structures Step-by-Step Using Incremental Dynamic Information

We will first briefly discuss two key asides concerning the materials used in this subsection. For reasons of generality, we should not treat the different syllable shapes (i.e. CVVs, CV-/p t k/ syllables and CVNs) as different subtests of the larger experiment, despite that /ɔ:/ occurred only in CVVs, /ɛ/ only in CVNs and /ɒ/ only in CV-/p t k/ syllables. In other words, the percentage values given in this chapter on recognition and confusor vowel options display results representing English CV(V)/Cs as a whole, rather than comparing the results for different CVNs and/or CV-/p t k/s with each other. However, as stated in the previous subsection, we will also discuss and illustrate the underlying perception-production trends using examples of both male and female stimuli.

As an important aside, the exemplification of the findings for /i:/ in 5.3 and /ĩ ã/ in 5.4 represent the results for the particular vowel sounds *in all word stimuli* having one of these three vowels, rather than results for the recognition of individual lexemes. The examples used illustrate general trends. We now move on to the 'pea' example. We begin by first illustrating male stimuli at the different gate intervals and

then move on to exemplify the equivalent female productions and gate intervals at t + 10, 20, 30 and 40ms.

5.3.1 Example: abstraction of 'pea'

Having characterised the abstraction process generally, let us examine the onset plosive looking only at recognition from the burst transient (= time slot 1). With little accompanying vowel resonance, listeners have six choices in English on which to base their abstraction, /p t k b d g/[22]. Before commencing the analysis, we present the abstraction the listener should be aiming for in 'pea':

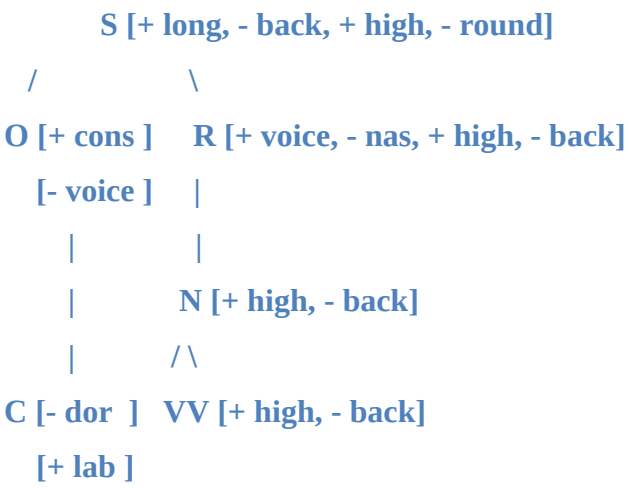


Figure 68: Correct phonological abstraction for 'pea'

The principles behind the representation of feature sharing in figure 68 are based on the abstraction rules displayed in figures 62-63 as well as on the way exponents of CV(V)/Cs are distributed throughout a syllable (cf. figure 61). For example, exponents of length are distributed throughout the CVV, while those for voice and nasality are shared between the onset and its daughters and the rime and its daughters, respectively. Figure 68 shows the features shared between the various nodes in the representation for 'pea' (such as the one for nasality), and

[22] The use of '/ /' brackets represents contrast in this chapter, *not* phonemes.

that the length feature is best represented at the syllable level (cf. the top of figure 68). This conclusion is based on the fact that the phonetic encoding of length is exponed distinctively in long and short vowels, so that the formant centre frequencies in the transitions for short vowels vary more rapidly temporally than in long vowels (cf. figures and commentary on VISC in subsection 2.1).

Although listeners may in principle be able to make good guesses as to the type of syllable that is upcoming at time slot 1, they cannot be fully certain as to the place of articulation of the onset, given that hold phase duration varies between speakers (and even within the same speaker's productions, Lisker, 1957, p. 43) and also because insufficient FPD is available on the formant transitions at time slot 1.

Therefore, the spectral properties of the received input point towards a percept that is to be built up as follows:

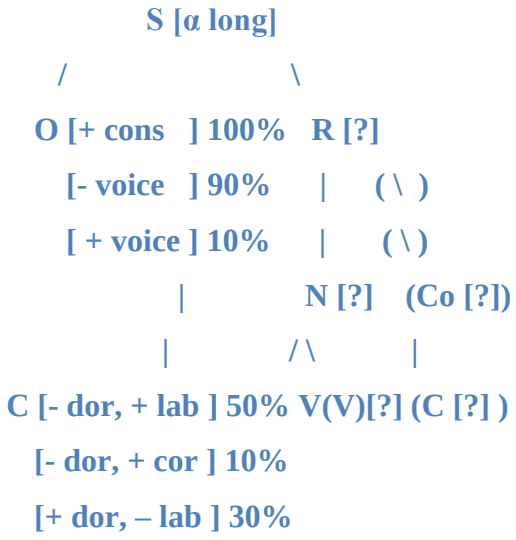


Figure 69: Set of abstraction probabilities for a monosyllable at t + 0ms

Since the [+/-voice] feature is specified as 10% at the O level, the percentage values for place of articulation add up to 90% in figure 69 (= a total of 100%). The question marks on the right-hand side of figure 69 at the rime, nucleus and vowel levels denote the listener not having sufficient phonetic information at this point in time with respect rime properties (i.e. at this point the vowel quality is underspecified).

Given the phonetic information available from the burst and closure duration, we can conclude that the listener abstracts a [- voice, - dor, + lab] plosive onset (cf. figure 69). Vowel quality remains more uncertain than in cases where there is more audible vocalic resonance (as e.g. at time slots 2 and 3). For example, the listener is not yet certain whether a long or short vowel is upcoming, since insufficient vocalic resonance is audible. Hence '?' signifies that both abstractions remain equally likely at time slot 1 (this type of illustration is particular this figure 69).

First, the issue of underspecification relates to the generality of this model rather than claiming that listeners had such choices available as potential responses in the experiment described in chapter 3. For example, given the differences in closure duration between voiced and voiceless plosive onsets (e.g. Lisker, 1957 and Davis and Summers, 1989), the fact that a [- voice] percept can be seen as 9 times as likely as a voiced one (90% vs. 10% respectively, cf. figure 67) means that voiced abstractions are equally as unlikely. That is, ca. 50% / 9 gives us a ca. 5.6% probability for a voiced bilabial, for instance.

Second, the abstraction reached by the listener (see e.g. figures 68-69) always comprises the vowel with the highest percentage abstraction probability. Uncertainty is built into the model, and a listener will abstract the vowel that is most likely given the phonetic information heard so far. Similarly, such a strategy in this model helps to account for the results displayed in 4.5 on perceptual confusions, since it allows a reliable and adequate explanation for why listeners rely most on acoustic similarity in making choices on different vowel rather than on any other criteria.

For example, having heard a relatively long hold phase of ca. 100-140ms in the plosive at time slot 1, listeners may be more likely to choose [– dorsal, + labial] as their abstractions

rather than [+ dorsal, - coronal] and [+ coronal, - dorsal] whose plosive realisations would have shorter hold phases (see e.g. Lisker, 1957 and Stevens, 1998). The hold phase for a voiced plosive will comprise a shorter mirror image of its voiceless counterpart (e.g. Lisker, 1957). It is important to bear in mind that voiced plosives as produced in varieties spoken in England may have little or no voicing during the hold phase (see e.g. Docherty, 1992, pp. 115-117). For this reason, /b d g/ remain possible abstractions.

Given these types of cues to place of articulation and voicing in 'pea', the probabilities for abstraction at time slot 1 can, for example, be said to be 50% for a bilabial voiceless plosive, 30% for a velar voiceless plosive and 10% for an alveolar one. By analogy, voiced bilabial, voiced velar and voiced alveolar could be said to have ca. 5.6%, 3.3% and 1.1% probabilities as possible abstractions at time slot 1, due to the differences in hold phase durations between voiced plosives and their voiceless counterparts.

At this point in time, the listener can only be fairly certain that a voiceless plosive has been produced and fully certain that a syllable is upcoming. We now move away from a more general account to discuss recognition at time slot 2.

5.3.2 Abstraction at Time Slot 2 (Burst Transient with 10ms Vowel Resonance)

We can use the spectrogram in figure 70 to represent the acoustic evidence a listener has access to having heard 'I think you s**ay p**' + 10ms of vocalic resonance (all spectrograms and waveforms in this chapter are organised similarly):

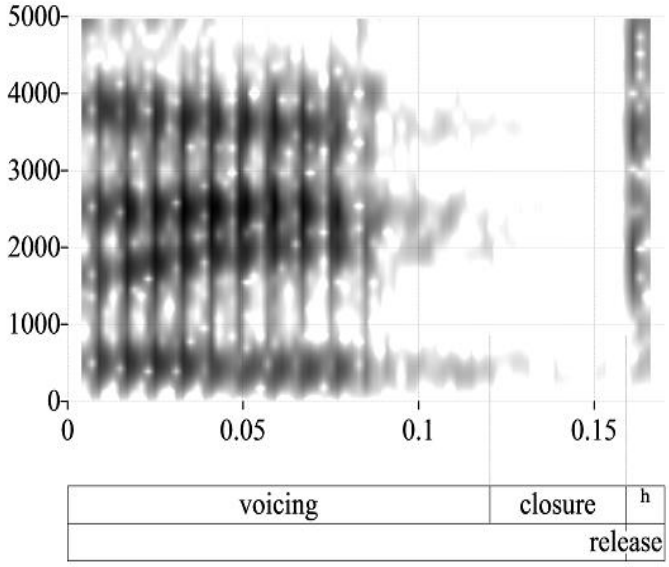


Figure 70: A partial segment of the '[eɪ pʰ]' portion in 'I think you s**ay p**ea' (10ms gate) produced by a southern male speaker

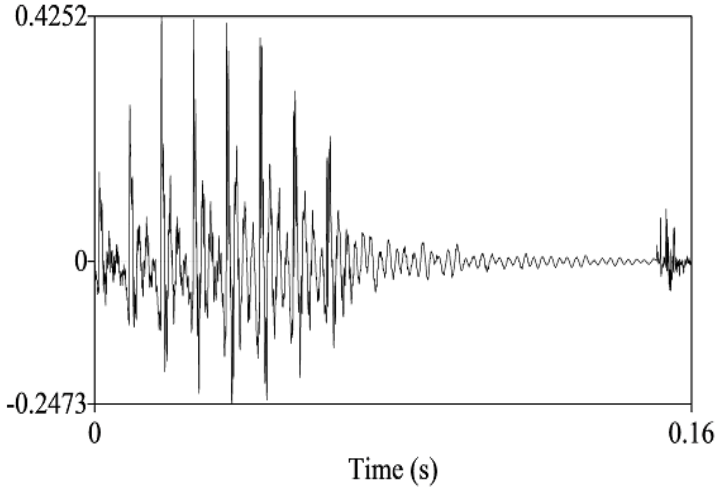


Figure 71: A waveform of '[eɪ pʰ]' in 'I think you s**ay p**ea' (10ms gate) produced by a southern male speaker

Having heard a long hold phase of about 100-140ms and 10ms of the initial transitions into the vowel from the transient in the onset plosive at time slot 2, the listener can be more certain that the onset is [- voice], though it may not yet be possible for the listener to decide firmly on its place of articulation, because i) insufficient perceptual detail on the

initial formant transitions out of the burst may be available and ii) due to the fact that closure durations for /p/ may vary, even for the same speaker's productions of the same word form (Lisker, 1957, p. 43). Despite these caveats, the probability of a [- voice] percept has been increased to e.g. 95% while the listener might only be able to work out that the plosive seems to be a labial one, with some degree of uncertainty. We now illustrate the perception of vowel quality in more detail at time slot 2.

Listeners have a 66.7% chance at time slot 2 to recognise 'pea', and opt for the following abstraction (cf. 4.5):

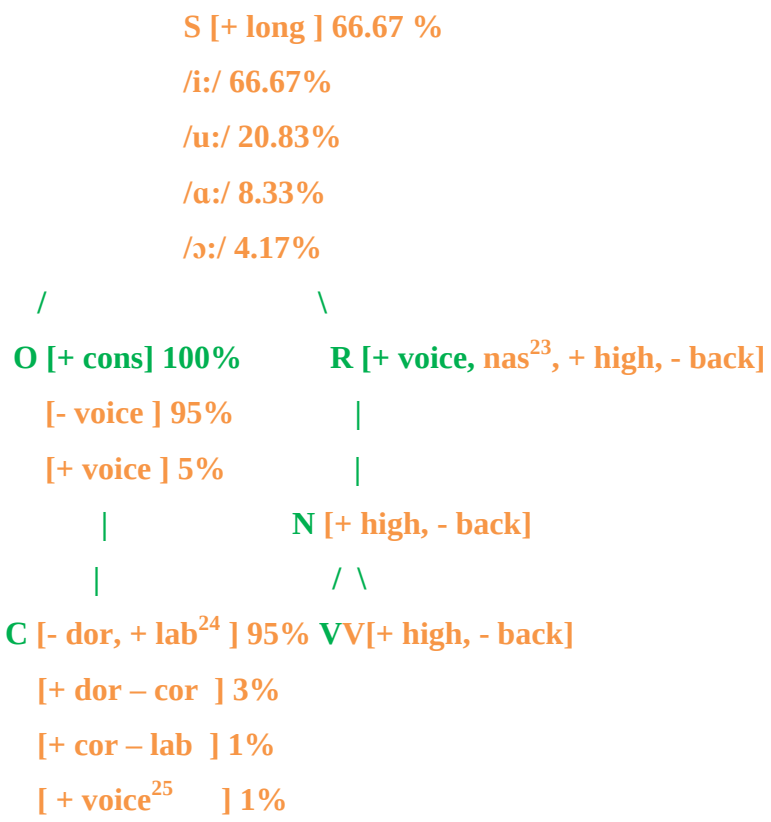


Figure 72: Abstraction for 'pea' at t + 10ms

[23] Nasality is not specified featurally until t + 30ms, which is considered the earliest point in time the feature can be distinguished (see e.g. Ali, 1971)

[24] Given the phonetic combination of closure duration and the spectral location of the burst, a [+ labial, - voice] percept is highly likely.

[25] The percentages in figure 72 for voiced places of articulation mirror the cue of closure duration similarly as for figure 69

The colours in figure 72 show the updates made to the representation moving from time slot 1 to slot 2: orange reflects updated representations, whereas green shows already deduced structures and constituents. The same principle as for figure 71 applies to the other abstraction figures in this chapter (though the colours differ in each case).

The listener is already ca. 66.7% certain that a high vowel has been produced rather than a low one, since all phonetic information for F1 points to such a percept (cf. the results detailed in subsection 4.5.2). Since English varieties spoken England generally do not have /ɛː/ or /eː/-like monophthongs (Wells, 1982), this conclusion is reinforced.

For /iː/, the most important combination of acoustic energy for F2-F3 can be seen to occupy an area at a relatively high frequency for a vowel, while F1 is low in its centre frequency. The energy minimum will occupy a relatively wide area at the middle of the 0-3500 Hz spectral area that is most important to vowel perception (see Harris and Lindsey, 1995, p. 18). Such quantitative variability can aid listeners in the abstraction process, which requires relating time-varying signals to qualitatively different types of target. This is an important point that applies to the exemplification of both male and female stimuli in this chapter, and through which the findings related to both types of speakers can be discussed similarly, as the underlying qualitative trends do not significantly differ (cf. chapter 4).

However, the most important point in this context is at what level we represent the length feature and to what extent it is shared with the consonantal slots in a CV(V)/C. This phonological feature is reflected in properties relating to duration, which correlate with VISC. Since duration correlates with VISC, so that the movements of the main formants

become slower for long than for short vowels, listeners find it harder to deduce the underlying vowel quality as rapidly and reliably. Long vowels take slightly longer to recognise reliably compared to short vowels. However, as the phonological literature on the phonetic encoding of length suggests (see Coleman, 1998 and figures 18-19, 69) as well as the claims on VISC in 2.1, length is exponed across the phonetic exponents of monosyllables. The phonetic encoding of duration as a marker of temporal dynamicity in CV(V)/Cs can be seen as the most significant polysystemic finding in this research, since length can be represented as a syllable level feature rather than at the nucleus level. It is clear from previous research by Coleman (1990, 1998), Ogden (1992) and Local and Ogden (1997) that some vocalic features, such as backness and/or height, should be represented at the syllabic level. This claim is not as true for length, however (cf. e.g. Coleman, 1990, 1998). The rapidly time-varying phonetic exponents of vowel duration can be seen to significantly affect the time course of recognition and the projection of upcoming constituents from the aperiodic phase.

From this claim on the phonetic encoding of length at the syllabic level in English follows that listeners can project CVC structures in advance of their physical realisations just as well as CVVs. Although listeners were not asked to distinguish short and long vowels in this study, the rapidly time-varying exponents in VISC give listeners sufficient access to length cues to make such a distinction very early on.

In sum, since the listener knows from the rapidly time-varying properties of VISC that the upcoming vowel is a short one, s/he can project a CVC just as well from 10, 20, 30 and 40ms of vocalic resonance as for a CVV. Nevertheless, since a [+ long] abstraction is the most likely choice given by listeners, the optional coda is rendered implausible as a response choice, given that short vowels only occur in closed syllables in

English (Coleman, 1998). We now move on to briefly discuss the other possible response choices given by listeners at time slot 2.

In the context of time slot 2, the listener has no access to robust phonetic evidence for the types of exponent (such as a high F1 and more variability in VISC in lower and longer vowels) that normally accompany a low vowel (i.e. [+ open] abstractions are likely to be excluded), with stronger evidence for a /i:/ percept than /u:/, for which F2 and especially F3 would have lower centre frequencies than for /i:/. The potential /u:/ abstraction relates closely to the /u:/-fronting in varieties spoken in England (see e.g. Wells, 1982, p. 294 and Foulkes and Docherty, 1999, p. 7): F2 can reach values of up to ca. 1900Hz, as spoken by males. The phonetic exponents of /u:/ are highly confusable with /i:/ in most varieties of English studied in this research. There are two likely parses at time slot 1, one of which is less likely than the other (66.7% contra 20.83%).

For this reason, we can conclude that the pull that is exerted towards the high front area of the vowel space for long vowels is partly explained by the phonetics of F2 in long high vowels in varieties spoken in England. In summary for time slot 2, recognition of 'pea' can be seen to exert a pull towards the high front area of the vowel space very early on during the aperiodic phase, with 7 of 8 of responses being for high vowels.

It is suggested here that we can in large part account for the remaining 12.5% of responses as relating to the types of order effects described in chapter 2 (which make neighbouring vowels more confusable) and/or lapses in concentration by listeners.

At time slot 2, more information is available about the phonetic and phonological identities of a) the type of plosive

heard and b) the quality of the upcoming vowel. As the syllable structure and vowel quality unfold through time from time slot 1 to 2, the listener has become more certain about vowel quality and the more general aspects of syllable and phonological structure.

Overall, the listener can more reliably project vowel features from the available phonetic information compared to time slot 1, with little access to information on vowel quality.

We now move to look on at how the unfolding of temporal dynamic information on the vowel can be abstracted by listeners when incremental vowel resonance information is being heard at time slot 3.

5.3.3 Abstraction at Time Slot 3 (Plosive Burst with 20ms Accompanying Vowel Resonance)

At 20ms, the listener has the following spectral cues available (since 10ms more vowel resonance is now audible):

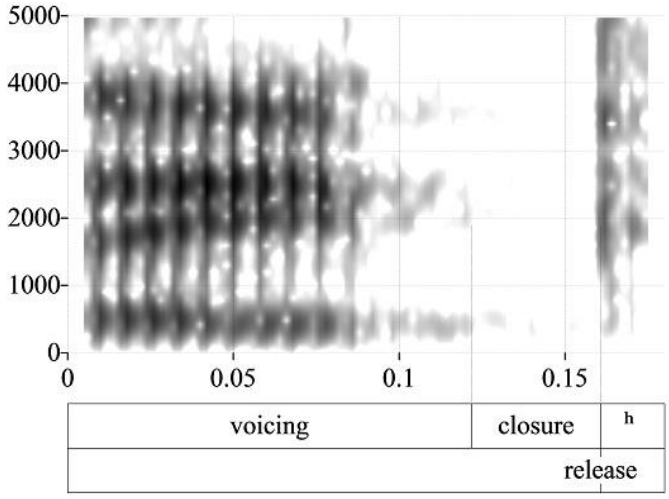


Figure 73: A partial segment of the '[eɪ pʰ]' portion in 'I think you s**ay p**ea' (20ms gate) as produced by the southern male speaker

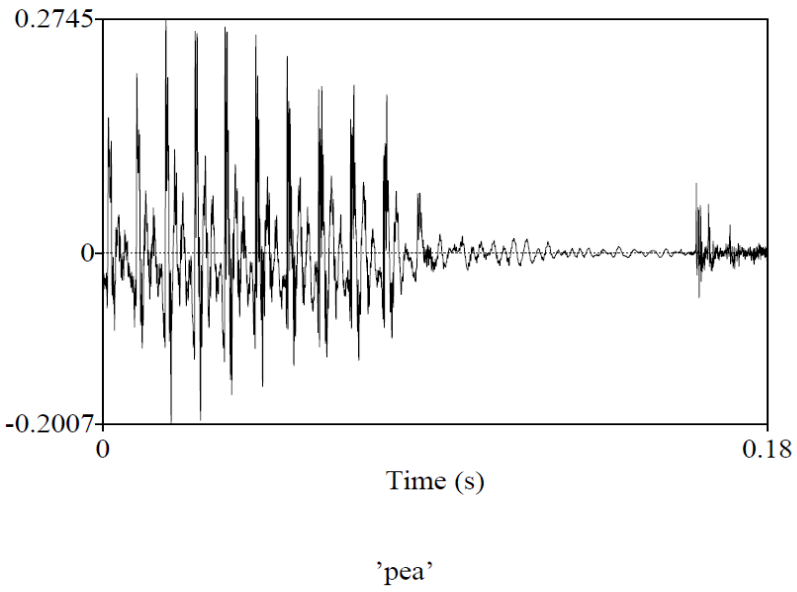


Figure 74: A waveform of '[eɪ pʰ]' in 'I think you s**ay p**ea' (20ms gate) as produced by the southern male speaker

Given the incremental information provided by the 10ms of additional vocalic resonance takes the listener a step closer to deducing the overall relationship between the main formants, since a larger proportion of vocalic information is audible. The listener has a better chance of correctly abstracting the underlying structures, while making more compatible phonological bifurcations from the longer duration of audible vocalic resonance. The listener has begun to receive more reliable indications as to the trajectories of the main formants (F1, F2 and F3). This acoustic property relating to both the spectral location of the burst (see e.g. Hillenbrand et al, 2001) and changes in VISC gives the listener an even stronger perceptual cue to place of articulation. The listener is now fully certain that the onset should be heard as [+ labial, - coronal][26]. This claim can be defended by the fact that the F2 offglide frequency from the burst portion of the realisation of the voiceless plosive is located ca. 5.4% higher in frequency than

[26] This observation is meant as a general statement that is not applicable to the results but rather to the phonological phrase structure rules set out in 5.2.

for its voiced cognate (e.g. Stevens, 1998: 362-365). This additional spectral cue is useful in bifurcating between [+ / - voice] at the onset level. But it can also be used at time slot 2 in terms of excluding [+ coronal] and [+ dorsal] structures as possible options at the consonantal level of abstraction, since the offglide F2 frequencies for alveolar and velar plosives in the context of a [+ high - back] vowel tend to be located higher than for [+ labial] plosives (see e.g. Stevens, 1998, pp. 362-365 and 371-374).

The spectral locus for the burst is not only typical for a labial plosive (e.g. Stevens, 1998), but the way in which the main formants have started evolving is typical for a vowel that is realised in the high front area of the vowel space. The listener has become yet more certain that a quality approximating towards /iː/ is being heard, though listeners cannot be entirely certain about this abstraction at this point in time in the absence of more solid evidence from the way in which VISC is distributed spectro-temporally.

Since the listener has access to VISC, it is possible to make a deduction from the spectro-temporal variation in 'pea' to distinguish for vowel length: the real-time changes in VISC are slower than for /ɪ/, for example, and do not start to evolve as rapidly in the early part of the aperiodic phase. This conclusion applies to the overall trajectories of F2 and F3 in particular. The trajectories of the main formants may comprise a key indication for the listener in terms of making a reliable bifurcation as to the phonological feature of length, enabling the following set of possible abstractions[27]:

[27] The colour scheme in the abstraction figures differs for each gate since different updates to recognition probabilities are made at each time interval.

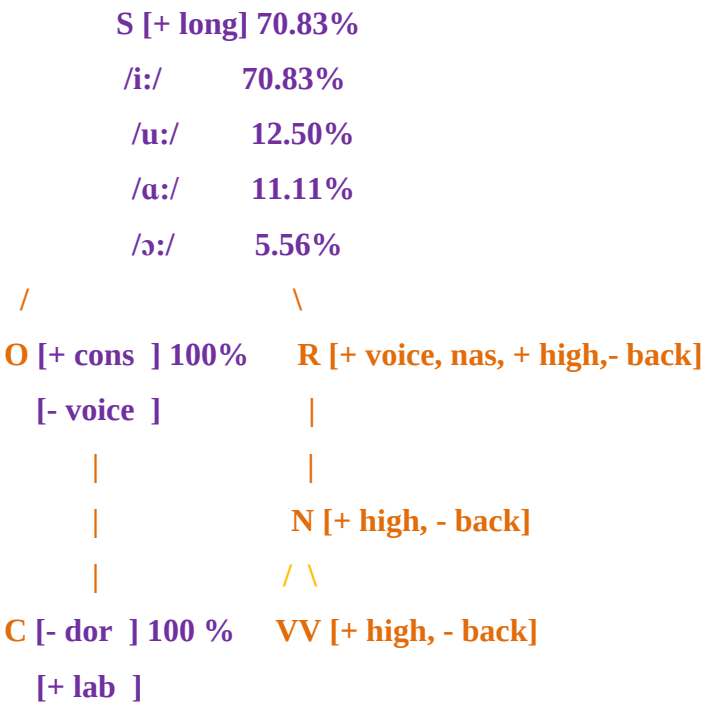


Figure 75: Abstraction for ‘pea’ at [t + 20ms]

The listener is now ca. 70.83% certain about the phonetic identity of the vowel (see subsection 4.5.2), having initially been somewhat more ambivalent. Two conclusions can be drawn from the results at time slot 3 for the [+ high, - back] vowel option: despite the additional 10ms of vocalic information at time slot 3, recognition has increased by only ca. 3.2%. This conclusion may in part be explained by the types of order effects attributed to vowel perception as presented in chapter 2 as well as the general non-linear nature of speech and vowel perception (see e.g. Moore, 2008, and Rosner and Pickering, 1994). For example, despite the fact the additional 10ms of vocalic resonance represents a doubling of the magnitude of cues to the underlying vowel, the recognition reliability has increased only slightly from time slot 2 to slot 3. [ɑ:] remains almost as viable a confusor as [u:], with only a ca. 1.4% difference (12.5% vs. 11.11%) as responses. This finding could mainly be explained by the lack of rounding in [ɑ:], whereas [u:] is rounded. In the presence of 10ms

additional vocalic information, listeners also have more robust access to F3 cues than at time slot 2.

5.3.4 Abstraction at Time slot 4 (burst transient + 30ms vowel resonance)

Time slot 4 represents an additional 10ms increase in the duration of audible vowel resonance compared to time slot 3; yet, the timing increment is smaller proportionally than between slots 2 and 3 (i.e. 100% added duration of timing information contra 50%). We can use the properties discernible in figure 76 to identify the most important pieces of acoustic information that listeners can use in recognition:

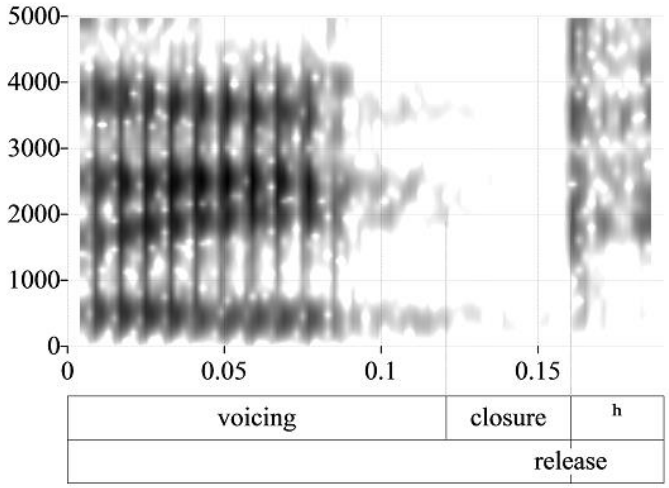


Figure 76: A partial segment of the '[eɪ pʰ]' portion in 'I think you s**ay p**ea' (30ms gate) as produced by the southern male speaker

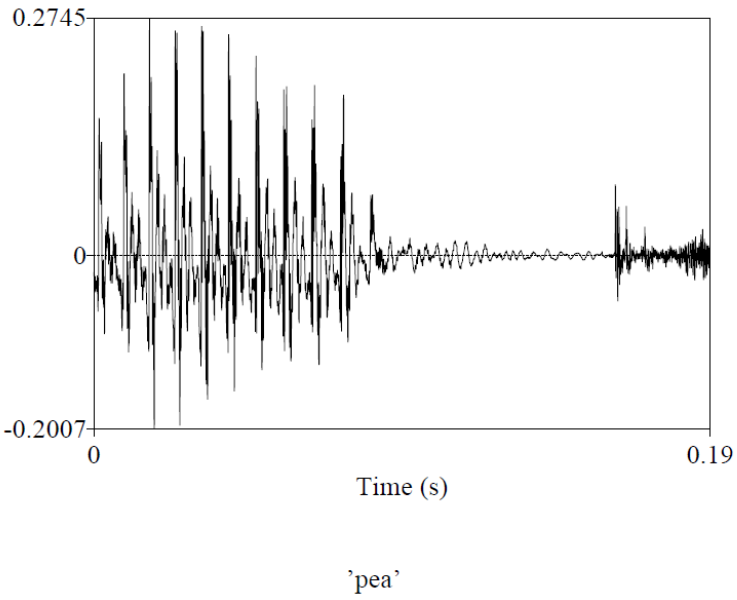


Figure 77: A waveform of '[eɪ pʰ]' in 'I think you s**ay p**ea' (30ms gate) as produced by the southern male speaker

Having heard a longer portion of the aperiodic phase of the onset at time slot 4, the listener is 100% certain as to its place of articulation. The listener also has partial access to F2-F3 as the main formants are manifested during the aperiodic phase (around 0.16-19 seconds in the bottom right-hand corner of figure 76). For example, figure 76 shows that the listener has more extensive (i.e. longer) spectro-temporal access to the trajectories of F1, F2 and F3 from the [eɪ pʰ] portion in the utterance 'I think you s**ay p**ea'. The energy minimum which is characterised by the lack of dark striations between ca. 500-2000 Hz as we approach the mid part of the aperiodic phase is typical of a high front vowel (e.g. Harris and Lindsey, 1995). These properties can be used by the listener to match against the following set of possible abstractions:

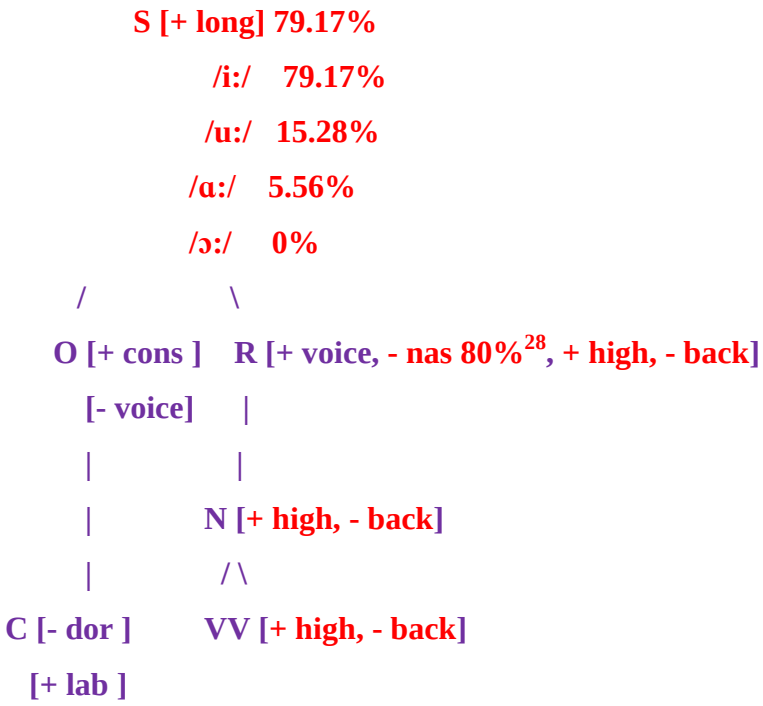


Figure 78: Abstraction for 'pea' at [t + 30ms]

The main conclusion we should make at time slot 4 is that listeners have updated the reliability of the initial projections for vowel quality, phonological features and syllable structure as follows: from figure 78, we can see that the increase in the reliability of detecting vowel quality has been increased by ca. 7.5% compared to the 3.2% increase between time slots 2 and 3 (cf. figures 72 and 75). This increment represents a 250% comparative addition in the reliability that the listener can recognise vowel quality (i.e. when we compare the increase in the reliability of recognition between time slots 3 and 4). This finding reinforces the claims made by Rosner and Pickering (1994) and Nearey and Assmann (1986) on the perceptual significance of the trajectories of the main formants ca. 30ms into the aperiodic phase. At this point in time the trajectories of F2 and F3 (and to a lesser extent, F1) begin to approach the steady state values more rapidly. As Rosner and Pickering

[28] The % values for recognition of the nasal feature at t + 30/40 ms correspond to the suggestion by Ali (1971), in that nasality can be reliably distinguished ca. halfway through the aperiodic phase in plosives.

(1994, p. 330) have suggested that detecting vowel quality involves performing an averaging of the magnitude of VISC during the entire duration of a vocalic gesture, the increase in formant movement velocity at time slots 4 and 5 shows the perceptual significance of the temporal evolution of VISC in the perception of coarticulation. We now move on to discuss the other responses that listeners gave at slot 4.

The probability of /ɔː/ as a response option has been reduced to 0% and that /uː/ remains the most likely confusor (though /ɑː/ receives a small number of responses). The importance of this finding is that it is indicative of the importance of phonetic similarity between vowels. Since /ɔː/ tends to have a very low F2 in varieties spoken in England, it is phonetically very distant from /uː/, rendering an /ɔː/ abstraction unlikely given the magnitude of additional coarticulatory information available to listeners at t + 30ms. We can see that the full duration that temporal dynamic change encompasses can be used by listeners to abstract vowel quality.

In summary, at time slot 4, the listener is able to recognise key properties of the upcoming vowel based on the fact that the consonants and vowel are coproduced, which is mirrored in recognition. The availability of cues to vowels in the aperiodic phase becomes particularly evident at time slot 4, since the listener has heard sufficient information at this point to reliably deduce the underlying formant relationships (Nearey and Assmann, 1986).

Having heard a larger proportion of the transitional part of the aperiodic phase, the listener is able to narrow down his vowel choices to a quality approximating towards /iː/. The listener opts for /iː/ with a. 79.17% probability.

The second conclusion we can draw relates to the listener having more reliable access to whether an oral vowel has been heard or not. Since subsection 4.4 will show that recognition of vowel quality is harder and slower from CVNs than from CV-/p t k/ monosyllables and given the conclusions by Rosner and Pickering (1994) on formant velocities at ca 30ms subsequent to the burst, the listener can be fairly certain at t + 30ms (say 80%, cf. figure 78) that a [- nasal] rime has been heard. The listener has heard no indication of the type of spectral distortions that are typical cues for [+ nasal] vowels, such as extra resonances or zeroes (see the reviews of Hawkins and Stevens' (1985) study on vowel nasalisation and Stevens, 1998). Since this question is a secondary one in this study and since listeners were not asked to distinguish oral and nasalised vowels in the experiment described in chapter 3, this conclusion and especially the probability assigned to the [nasal] feature remain tentative. Nevertheless, this issue is significant for the main finding on length, because it helps to show that given sufficient temporal dynamic information on properties like VISC (which correlates with vowel length) and aperiodic friction, listeners may be able to relate their response choices to a very small set of phonetic correlates. We will now show how recognition is updated at time slot 5.

5.3.5 Abstraction at Time Slot 5 (Plosive Burst with 40ms Accompanying Vowel Resonance)

The addition of temporal information on vowel quality from time slot 4 to 5 represents a yet smaller proportional increase in the duration of audible vowel resonance (i.e. compared to slot 3 contra 4). Are there further practical implications for recognition with respect to the evolution of formant information and FPD that require an account? In order to answer this question, we can use the properties discernible in

figures 79-80 to identify the most important pieces of acoustic information that the listener can use in recognition:

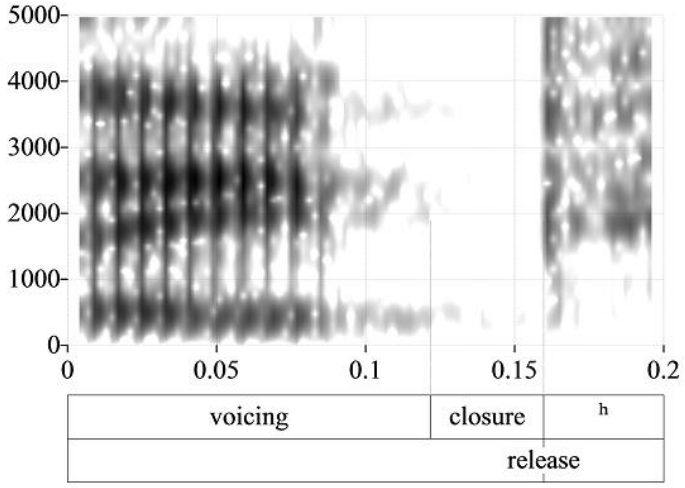


Figure 79: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay p**ea' (40ms gate) as produced by the southern male speaker

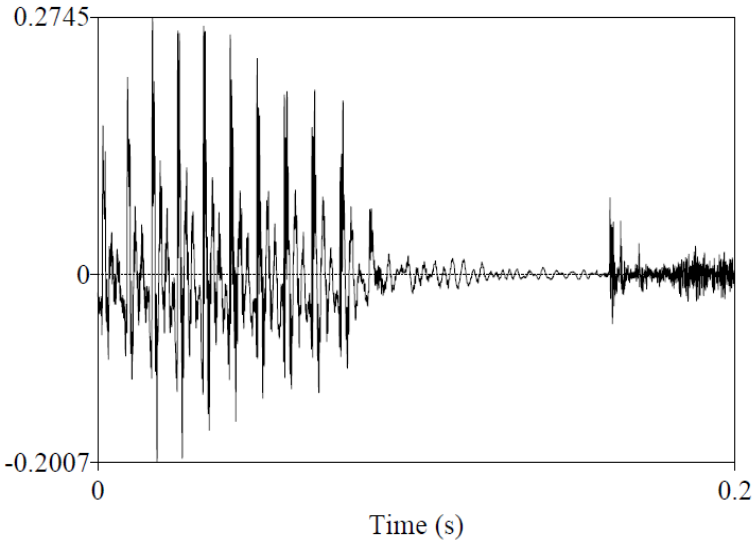


Figure 80: A waveform of '[eɪ p^h]' in 'I think you s**ay p**ea' (40ms gate) as produced by the southern male speaker

The most important conclusion we can draw at t + 40ms is that F2 and F3 are beginning to give listeners sufficient indication as to the final trajectory towards the steady state for

highly reliable recognition to be possible. When we inspect the figure 79 showing the initial parts of trajectories of F1[29]-F3 during the early part of the aperiodic phase, this issue concerning vowel categorisation becomes clearer: given the longer and clearer spectral cues on the evolutions of the main formants (cf. the right-hand side of figure 79), listeners have more robust and reliable access to vowel quality and length (as well as nasality) at time slot 5 than earlier on. Since there is now 10ms additional audible vocalic resonance and we have just passed beyond the main 30ms transitional part of the aperiodic phase referred to by Assmann and Nearey (1986), the listener is able to arrive at the following set of possible abstractions:

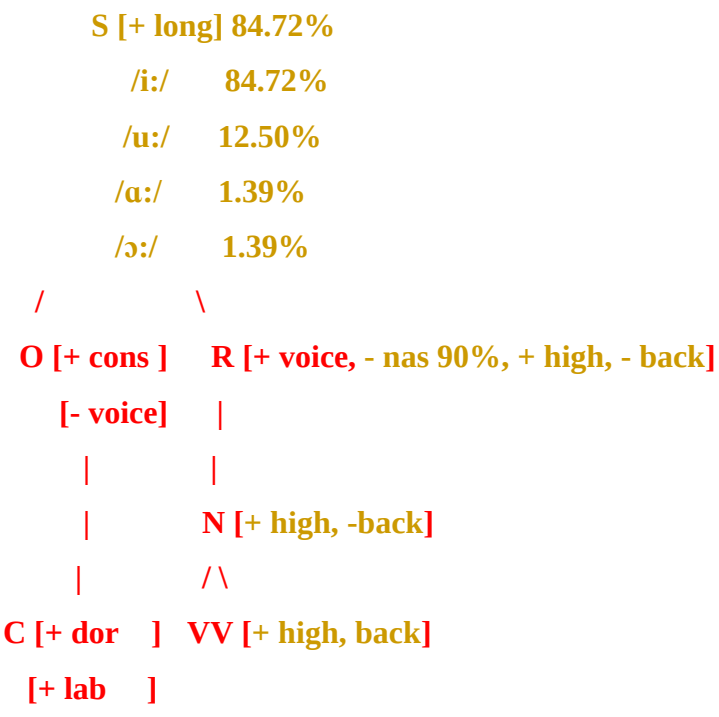


Figure 81: Abstraction for ‘pea’ at [t + 40ms]

The main conclusion we can draw from the values for the confusing vowel options [u:] [ɑ:] [ɔ:] at 40ms is that [ɑ:]

[29] Although F1 is shown in many of the spectrograms in this chapter, it may is variably present and can also be hard to estimate reliably in Praat.

[ɔ:] represent just 2/72 responses at time slot 5. The likelihood of a correct recognition been increased by ca. 5.5% at the 40ms gate, and listeners are usually able to recognise vowel quality accurately at time slot 5.

Given 10ms additional audible vowel resonance between time slots 4 and 5, the listener is now much more certain about his projection of the upcoming vowel as [- nasal], with a ca. 90% probability (cf. the right-hand and bottom right-hand parts of figure 81). Such a reliable projection might initially seem unlikely at time slot 5, however since Ali (1971) has shown that listeners *can* reliably distinguish for nasality ca. halfway through the aspiration portion, this claim receives good support.

To round off example no 1 for long vowels, the key claim that emerges from the 'pea' example is that since listeners know by rule that stressed open syllables in English are accompanied by long vowels, the phonetic encoding of length allows projecting the entire syllable structure in advance of most of its phonetic realisation. Listeners are quite certain that a CVV rather than a CVC is upcoming prior to the onset of vocal fold vibration in the vocalic portion, since they have access to VISC. Having applied the model fully to an example as produced by a male speaker, we will consider potential differences applicable to vowels as produced by female speakers by referring to equivalent productions of /i:/ as produced by the northern female speaker in the next subsection.

5.3.6 Applying the Temporal Abstraction Model to Female CV(V)/Cs

Mainly for the sake of variability in presentation, we will look at instances of 'tea' in this subsection. However, as in the

previous section, we will treat recognition of /iː/ as an average across ‘pea-tea-key’ rather than distinguishing the three onset types and their effects on recognition specifically. What remains most important in this context is the generality of the model to vowel recognition as a whole across time rather than individual structural specifications in onsets or codas.

Abstraction figures and proportions are not included in this subsection, since the underlying trends in FPD and especially formant relationships are qualitatively similar between male and female speaker stimulus productions (see subsection 4.2), and in order to avoid repetition. For example, even if we did find *some differences* between the proportions of correct recognitions of e.g. male and female /iː/, it would seem peculiar in the extreme to place too much emphasis on gender-related variation in this type of vowel recognition study. Any differences found would almost certainly not be significant in terms of categorisation, since listeners will pay more attention to other features in recognising speech as male or female, such as pitch, loudness and voice quality. We begin by considering ‘tea’ at t + 10ms for females. The organisation of all spectrograms and waveforms in this subsection is similar to that in the previous one displaying male productions.

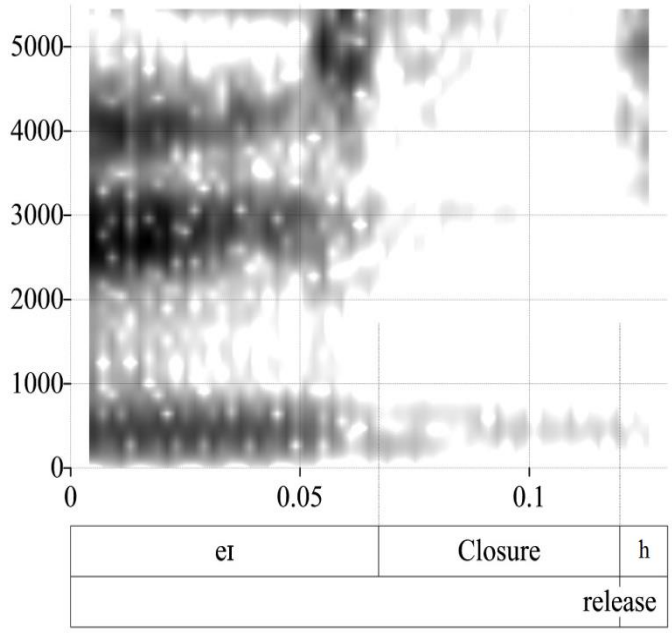


Figure 82: A partial segment of the [eɪ tʰ' portion in 'I think you s**ay t**ea' (10ms gate) as produced by the northern female speaker

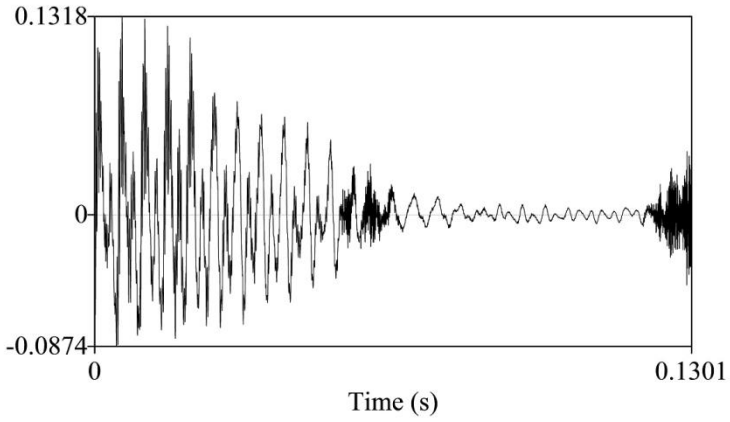


Figure 83: A waveform of [eɪ tʰ] in 'I think you s**ay t**ea' (10ms gate) as produced by the northern female speaker

When we look at the FPD of the various sound types produced at t + 10 ms in the female CVV in the spectrogram in figure 80, we can contrast it with the equivalent male production in figure 68 (reproduced immediately below) as follows:

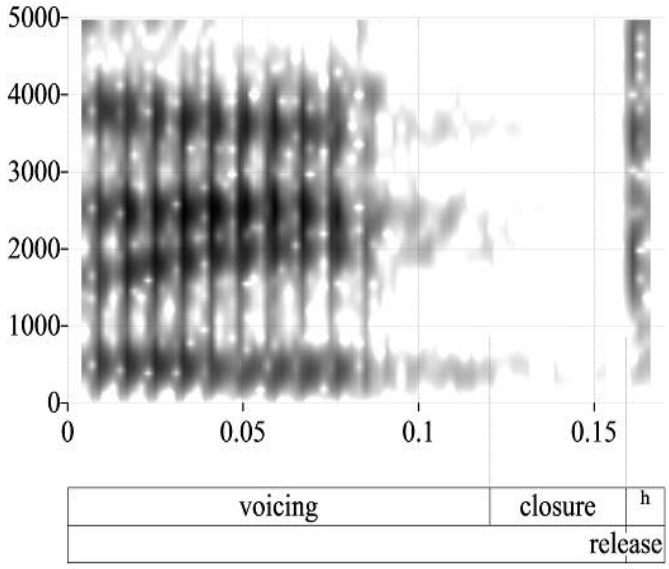


Figure 84: reproduction of figure 71, representing the beginning part of male /iː/ at t + 10ms

Albeit that the onset plosives in figures 82 and 84 are distinctive (cf. right hand sides of each spectrogram), and the spacings of the formants in the beginning portion of the female CVV in figure 82 will be somewhat larger (etc.), we can note at least some differences in FPD that could in principle influence vowel categorisation very early on. For example, when we look at the aspiration immediately subsequent to the plosive release in figure 82 at ca. 0.16 seconds, we can distinguish a dark band of relatively intensive energy, consisting of aperiodic friction (and some initial vowel resonances). For the female equivalent in figure 80, however, F1 and especially F2 at ca 0.13-0.14 seconds are difficult to discern, whilst the spectral area containing F3 has an intensive release at the same point in time. In sum, these types of differences may reflect the larger open quotient in female speech, leading to more critical damping of F1. However, this does not imply that female /iː/ will be harder to recognise reliably at t + 10ms than from equivalent

male productions. Instead, it may that since F3 is more strongly represented in the signal relative to F2 and F1 in female than in male speech at t + 10ms, listeners may on average find it slightly easier to distinguish female /iː/ from female /uː/ at this point in time than when hearing /iː/ as produced by a male speaker. This claim does not necessarily imply that the same will apply to the other two options available to the listener /ɑː ɔː/. It may be that the internal distribution of responses across vowel response options is slightly different in male vs. female stimulus productions. Alternative claims and suggestions are possible in this instance, however given the spectral differences between male and female productions of aspirated plosives in English and the FPD observed from figures 80 and 82, the claim receives some support. Now is a good time to move on to the next time slot (t + 20ms) and compare to what extent the same conclusions will apply at a point 10ms later in time:

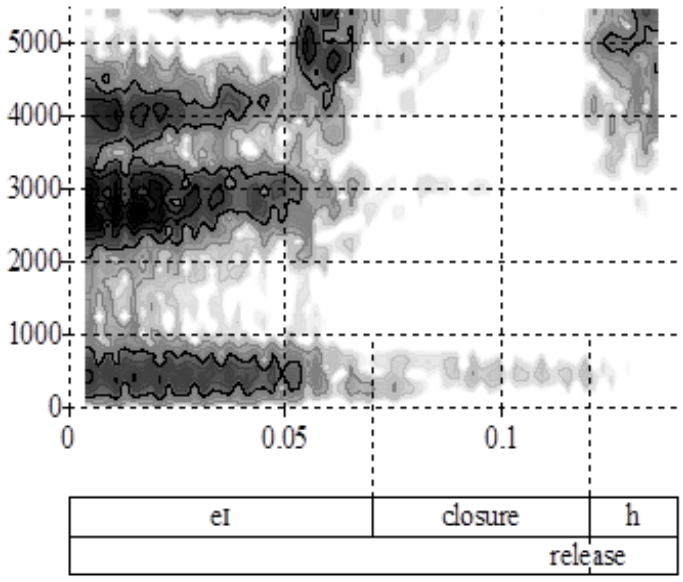


Figure 85: A partial segment of the [eɪ t^h] portion in 'I think you s**ay t**ea' (20ms gate) as produced by the northern female speaker

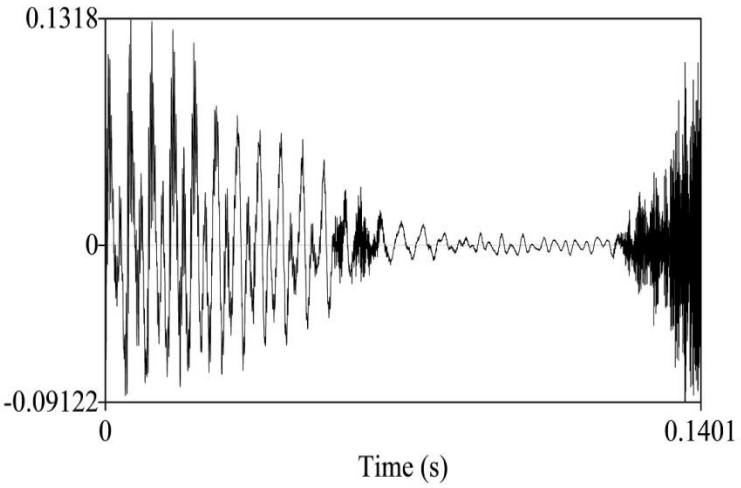


Figure 86: A waveform of [eɪ t^h] in 'I think you s**ay t**ea' (20ms gate) as produced by the northern female speaker

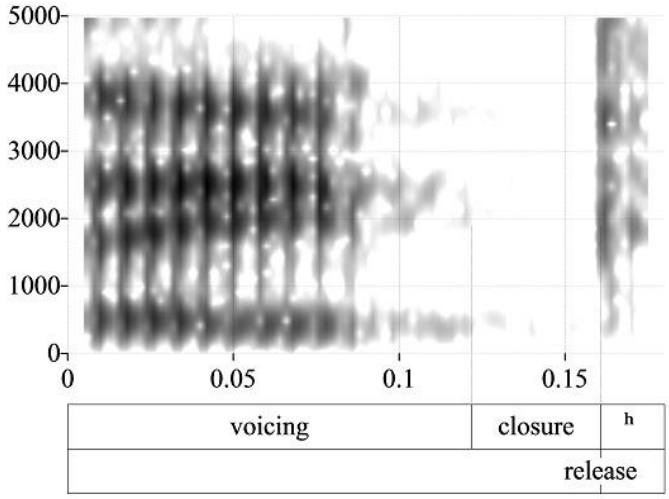


Figure 87: reproduction of figure 73, representing the beginning part of male /iː/ at t + 20ms

At t + 20ms for male contra female stimuli (cf. figures 85-87), the differences in FPD exhibited in the female variant of /iː/ is broadly similar to that observed in figures 81-83, both in spectral and temporal terms. The spectral region comprising the higher formants during the early part of the aspiration in the

female CVV at ca. 0.13-0.14 seconds has relative strong acoustic energy in the ca. 200-2.500Hz area, which is dampened in female /iː/. It is possible that such differences in FPD reflect other combinations of FPD than damping of e.g. F1 (such as intensity differences at given moments in production[30]), however since similar spectral properties are evidenced early on after plosive release in e.g. figures 6 and 7 representing southern female productions of 'car' and 'coo' (despite the difference in onset place of articulation), the claim is supported. Very early on, the listener is faced with a subtly different type of perceptual problem in recognising female /iː/ than male /iː/, regardless of onset place of articulation. We will now discuss the next time slot (t + 30ms) and compare to what extent the same conclusions apply at time slot 4.

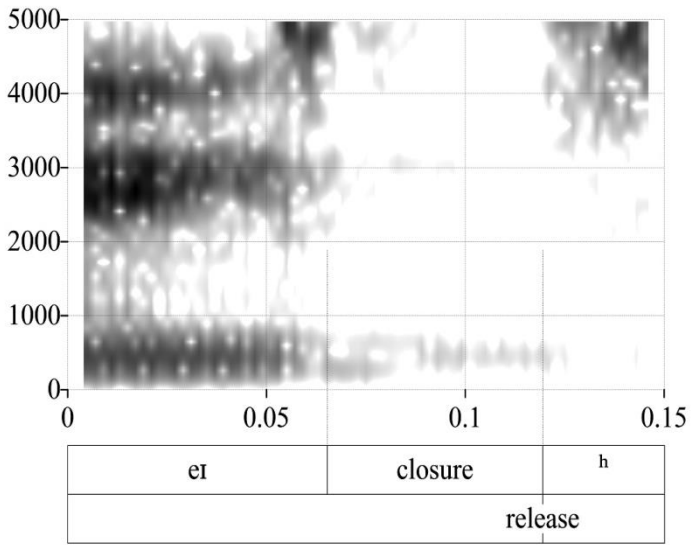


Figure 88: A partial segment of the [eɪ tʰ] portion in 'I think you s**ay t**ea' (30ms gate) as produced by the northern female speaker

[30] A comparison with the waveforms of male productions in figures 72 and 75 confirms this proposition not to apply, however.

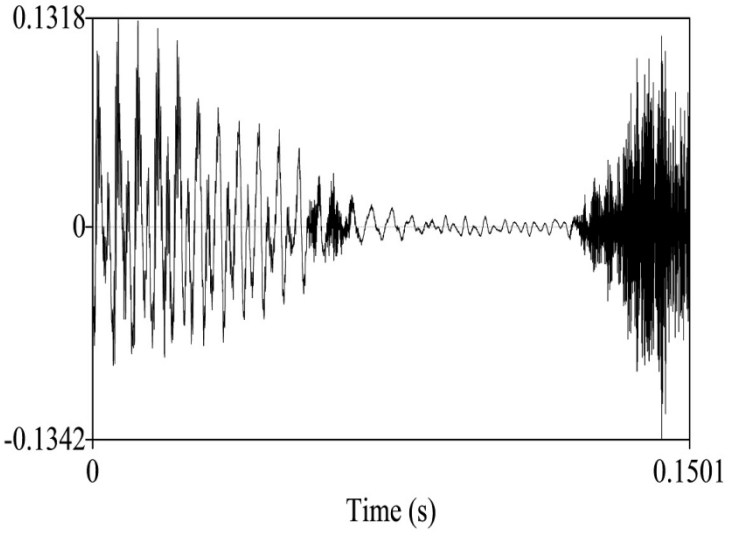


Figure 89: A waveform of [eɪ t^h] in 'I think you s**ay t**ea' (30ms gate) as produced by the northern female speaker

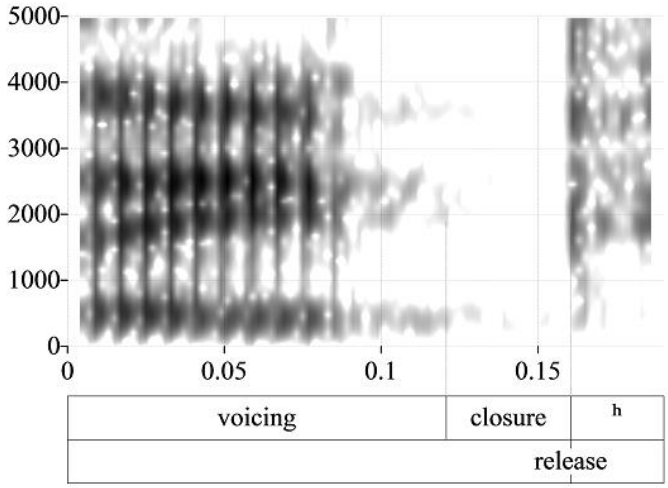


Figure 90: reproduction of figure 77, representing the beginning part of male /i:/ at t + 30ms

When comparing the emerging formant structure in the male production of 'pea' as displayed in figure 90 against that of female 'tea' in figure 88, similar conclusions can be arrived at as for time slots 2 and 3. There is still relatively little evidence of the kind of spectro-temporal continuity in F2 and especially F1 in female 'tea' between ca. 200 and 2.500Hz as in the equivalent male vowel in 'pea'. Where the male variant can

be seen to have a relatively clearly emerging formant structure at low frequencies, female /iː/ has little acoustic energy between ca. 200 and 2500Hz as estimated by Praat. It has been established thus far in this subsection that listeners face a slightly different task in phonetic terms in recognising the female long high front vowels compared to male long high front ones. There can be little doubt of the validity of this claim. However, it was also suggested in chapter 3 that subglottal and/or other resonances and effects (such as the damping of F1) might complicate the interpretation and measuring of female formant peaks at low frequencies. Since we are well on the way towards the vowel steady state portion at t + 30ms, the formant estimation method in Praat for measuring resonances in aperiodic friction as produced by females may engender a) more inaccurate estimates of formant peaks or b) not find them. We can say beyond doubt that recognising female /iː/ at t + 30ms is not exactly the same thing phonetically as perceiving male /iː/ at the equivalent point. However, the spectrograms and waveforms on male vs. female /iː/ in figures 82-90 may not tell us the whole story behind the phonetic differences in equivalent male and female vowels as produced in aspiration. For this reason, when we assess such spectro-temporal differences in the evolution of formant information through time in male vs. female vowels, it is important to bear in mind the limitations of current measurement methods. We will now move on to discuss the remaining time slot at t + 40ms.

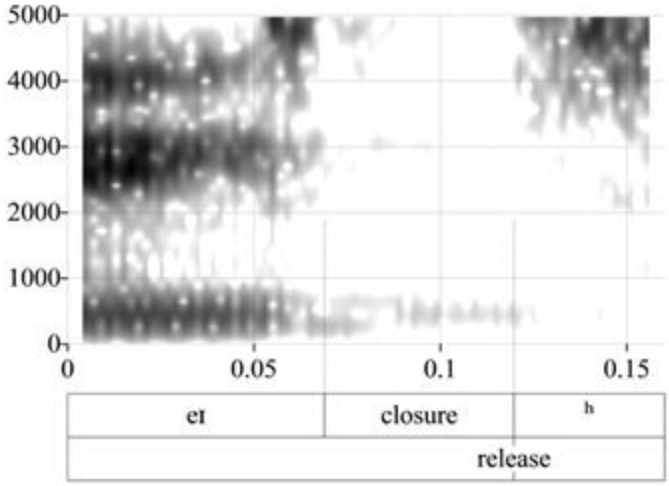


Figure 91: A partial segment of the [eɪ t^h] portion in 'I think you s**ay t**ea' (40ms gate) as produced by the northern female speaker

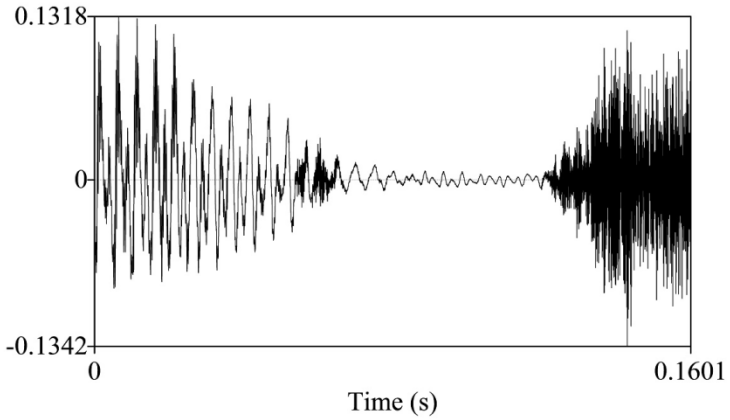


Figure 92: A waveform of [eɪ t^h] portion in 'I think you s**ay t**ea' (40ms gate) as produced by the northern female speaker

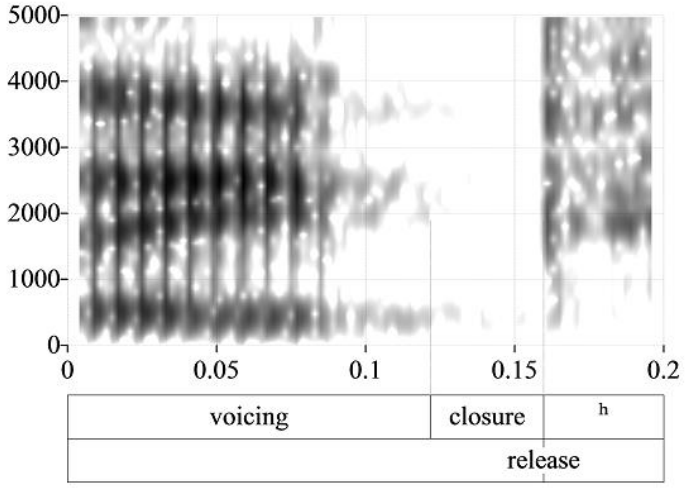


Figure 93: reproduction of figure 80, representing the beginning part of male /iː/ at t + 40ms

In summary for the spectro-temporal differences between male and female /iː/ at t + 40ms, we can note similar distinctions as at the shorter gates. There is little or no evidence for an emerging F1 in the female variant (cf. figure 91), which is probably due to the heavy damping of F1. However, at ca. 0.15 seconds in the 2.500-2.700Hz region, we can begin to see clearer traces of F2 at t + 40ms. In summary, these three differences between female /iː/ contra its male variant displayed in figure 93 aptly show the typical spectral differences between male and female CVVs as vocalic information evolves through time. As far as the role of F1 is concerned, we can be less certain, since Praat may not be able to clearly locate the centre frequency of F1 in aspiration (especially as produced by female speakers). Therefore, despite the phonetic differences noted between male and female /iː/ in figures 82-93, recognising /iː/ from female speech is much the same thing temporally as from male productions. The spectral distribution of phonetic properties and/or resonances in a vowel

sound at its relative location in spectro-temporal space may be slightly different in such productions.

Overall, the resonances and other key properties in vowel resonance as it evolves through the beginning portion of the aperiodic phase still have a) highly similar exponents in qualitative terms (whatever absolute values they take) and b) listeners will have certain representations built in their minds on how female and male vowels differ. For these two reasons, it should be emphasised at the end of this subsection that recognising a given vowel from female or male stimulus productions remains much the same perceptual task. This claim applies despite the fact that we may not be able to reliable measure F1 in female productions and whatever minor differences might be found in the internal distribution of responses in the type of forced choice experiment described in chapter 3. Having fully applied the phonological model generally to /iː/ as produced/perceived in English CV(V)/Cs, we move on to discuss the recognition of vowels from CVNs. Since we have already established whether and to what extent recognising vowels from aspiration differs temporally with respect to male vs. female speech, we will only consider male productions for CVNs in the next section.

5.4. Perceptual Implications of Rime Nasality for Vowel Recognition

5.4.1 Overview

We now move on to discuss whether vowel nasalisation serves to disrupt and/or delay listener capability of recognising vowel quality in CVN contra CVC syllables.

First, nasality cannot be represented at the syllabic root node in English (see e.g. Coleman, 1998). Rather, it is

represented at the onset and rhyme nodes which share the feature [nasal] with the consonantal nodes and the coda.

Second, despite the fact that nasality is not contrastive for vowels in English (Coleman, 1998), its phonetic encoding does have important implications for recognition, as we will show in this subsection. We will refer to our findings by applying the example ‘pin’ to the model detailed in 5.2 and contrasting their findings (cf. the abstraction figures in this subsection) with the equivalent [- nasal] rime in ‘pit’). It is important to bear in mind that the examples used are meant to apply across all onset places of articulation for CVCs contra CVNs in this instance, and that the example stimulus words are partly used for illustrative purposes.

Thirdly, we will also compare findings for these two vowels against ‘pun’ in 5.4.7. We will thus look at whether vowel height and backness have significant effects on recognition. The most important thing to observe in this subsection is that these examples *do* represent general trends and are in many cases statistically significant, as the results in subsection 4.4 confirm.

Lastly, nasality may delay recognition of vowels by introducing certain distortions into the FPD of the aperiodic phase. The main goal is to show that such results are peculiar to high front and low front vowels in this study (cf. e.g. Beddor and Krakow, 1999, Krakow, 1994, 1973 and Hawkins and Stevens, 1985), but not to other front and/or back vowels.

However, it seems very unlikely that listeners are able to recognise other properties such as place of articulation for codas early on. It is suggested here that listeners cannot make reliable distinctions about nasality until ca. 30ms into the aperiodic phase *at the earliest* (as also detailed in the abstraction figures in 5.3.4-5.3.5 above). This claim is in line with Ali’s (1971) findings on listener ability to distinguish for anticipatory nasalisation ca. halfway through the aperiodic

phase in CVNs. The statement also reflects Cohn's (1990) findings on the temporal co-extensiveness between nasal codas and the aperiodic phase of plosives and Nearey and Assmann's (1986) claims about listener's ability to distinguish vowels with sufficient access to VISC.

The discussions for each time slot in this subsection present two things: a) comparisons between oral and nasalised vowels for CVN and CVt stimuli, which helps a) to eliminate any spectral discrepancies arising from place of articulation, and b) allows a comparison of the set of abstraction probabilities between CVNs and CVC monosyllables. Since previous research (e.g. Beddor and Krakow, 1999, Krakow, 1994, Schourup, 1973 and Hawkins and Stevens, 1985) shows that perceived vowel height and the magnitude of nasalisation may be significantly affected for high and low front (but not for mid front vowels), the main focus will be on the perceptually distorting effects of nasalisation with respect to vowel height. The production data detailed in 4.2.3, 4.3.3 and the perception results detailed in 4.4 support this conclusion, since the findings are significantly different for /ɪ/ and /a/ (cf. subsection 4.2.4 for production and 4.4 for perception). We now move on to point out a few key issues about modelling the relationship between vowel height and nasalisation before commencing the discussion.

5.4.2 Modelling the Relationship between Vowel Recognition and Nasalisation in CV(V)/Cs

The differences in the results between CVC monosyllables and CVNs detailed in subsection 4.4 and 4.7 demonstrate that for CVNs with /ɪ/ and /a/ vowels, it can be very hard to maintain recognition at a similar level either at individual gates or on an average through time for CVNs as for CVCs. Before

commencing the analysis, we will briefly highlight one issue concerning the abstraction figures in this subsection: since the main purpose of the abstraction figures in 5.4.3-5.4.7 is to demonstrate the differences between CV(V)/Cs *with and without nasal rimes*, only nasal and non-nasal vowel abstractions are distinguished (i.e. not all vowel choice probabilities are listed in each of abstraction figure).

5.4.3 Recognition at time slot 2 for 'pin' (plosive burst + 10 ms accompanying vowel resonance)

The listener should aim for the following abstraction for 'pin':

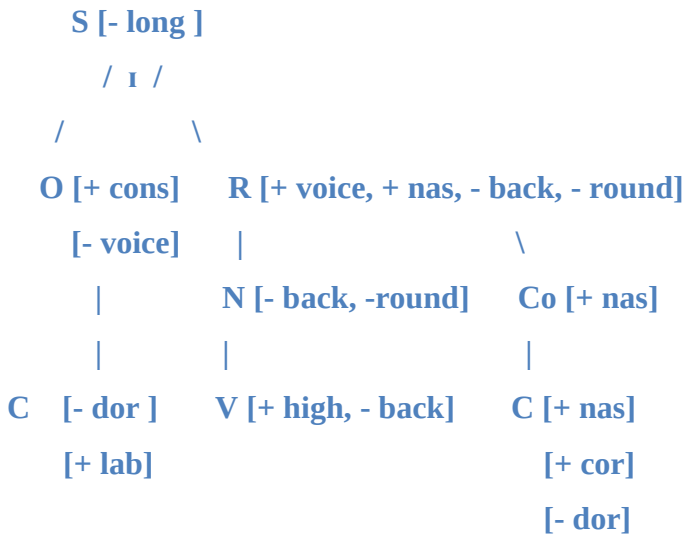


Figure 94: Partial phonological abstraction for 'pin'

It was established in chapter 2 that vowel height may have certain negative implications for vowel production in nasalised vowels (see e.g. Hawkins and Stevens, 1985). We will mainly illustrate and explore this issue with reference to 'pin' contra 'pit'. However, the fact that listeners may not be able to distinguish for nasality until time slot 4 is only partly relevant at time slots 2 and 3, since the results (cf. 4.4 and 4.7) show that [+ nasal] rimes still disrupt and delay the time course of recognition *at all four* gate intervals compared to stimuli with [-

nasal] rimes. The fact that listeners may not be able to distinguish for nasality at t + 10-20ms does not mean that it has no perceptual implications. Therefore, the significance of the percentage values assigned to recognition of nasality in rimes at time slots 4-5 is that the values are consistent with previous findings by e.g. Ali (1971), and little else. Thus, whether or not listeners are able to distinguish for nasality 10 or 20ms into the aperiodic phase is not relevant from a purely theoretical viewpoint if presence of nasality and/or nasal exponents delays vowel recognition significantly. We will now consult spectrographic evidence at time slot 2 to offer evidence for the claims on nasality.

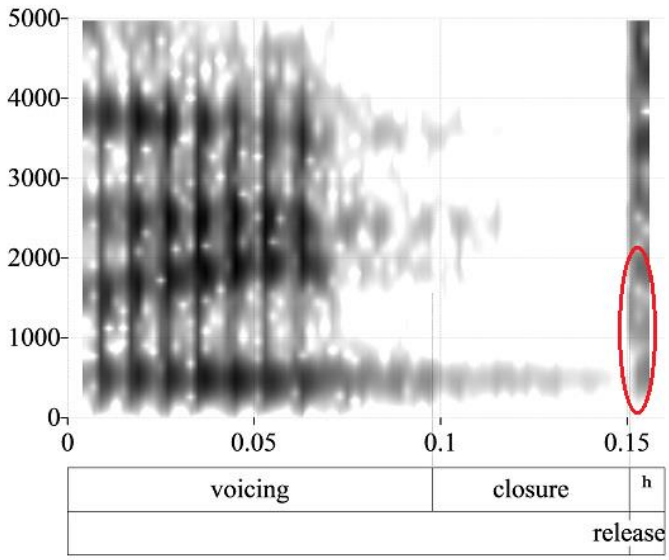


Figure 95: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay p**in' (10ms gate)

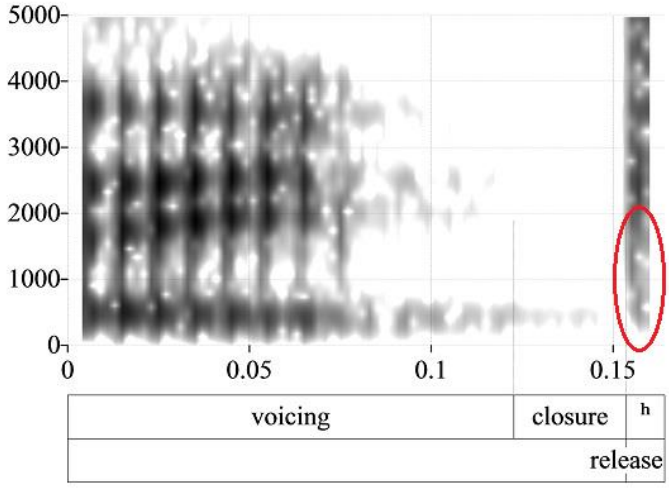


Figure 96: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **p**it' (10ms gate)

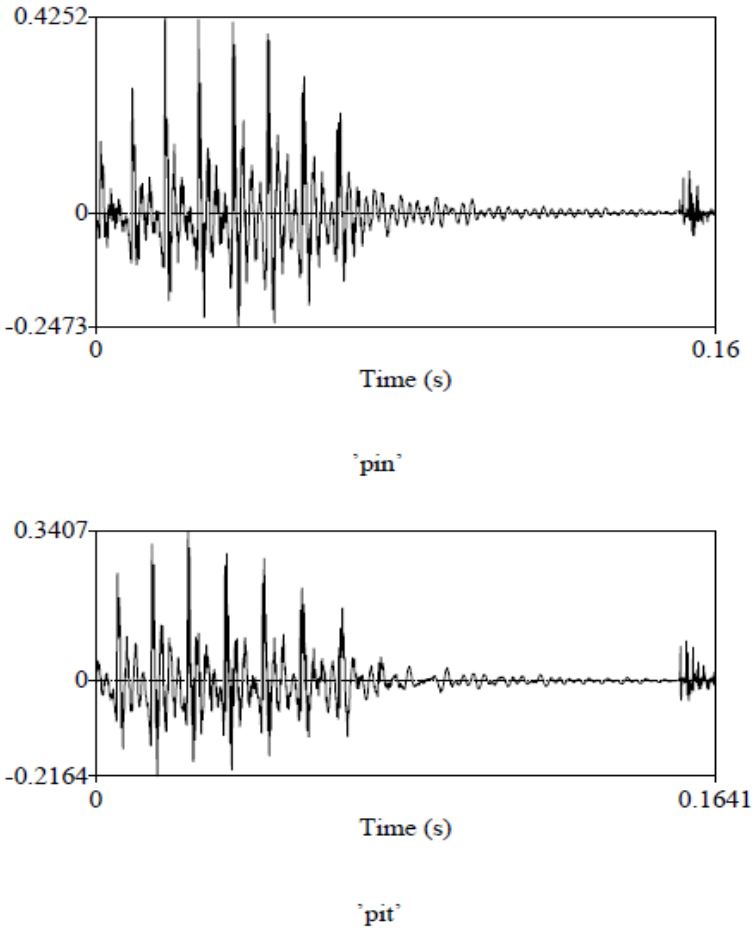


Figure 97: Partial stimulus waveforms for 'pin' (top) and 'pit' (bottom) at t + 10ms

What conclusions can we draw from the spectral evidence observable on the right-hand sides of figures 95-96 (and which listeners had access to)? When we zoom in closely to the area inside the red circles between ca. 100-2000Hz in both ‘pin’ (top) and ‘pit’ (bottom), we may be able to distinguish that the formant structure is emerging more clearly in ‘pit’ than in ‘pin’, with pockets of low energy around 200-500 Hz. Since the formants are more clearly distinguished in terms of their spacing in ‘pit’, listeners may find it easier to perceive their spectral relationships, that is. We do need to bear in mind that only 10ms of vocalic information is audible at t + 10ms, and that any spectral evidence we may able to discern on nasality might be difficult for listeners to recognise so early on, as in in figures 95-96 on ‘pin’ and ‘pit’. Yet, such variation can be significant perceptually, as the results of the statistical tests in chapter 4 confirm.

We will now compare how the spectral distortions in CVNs affect recognition probabilities in figures 98-99:

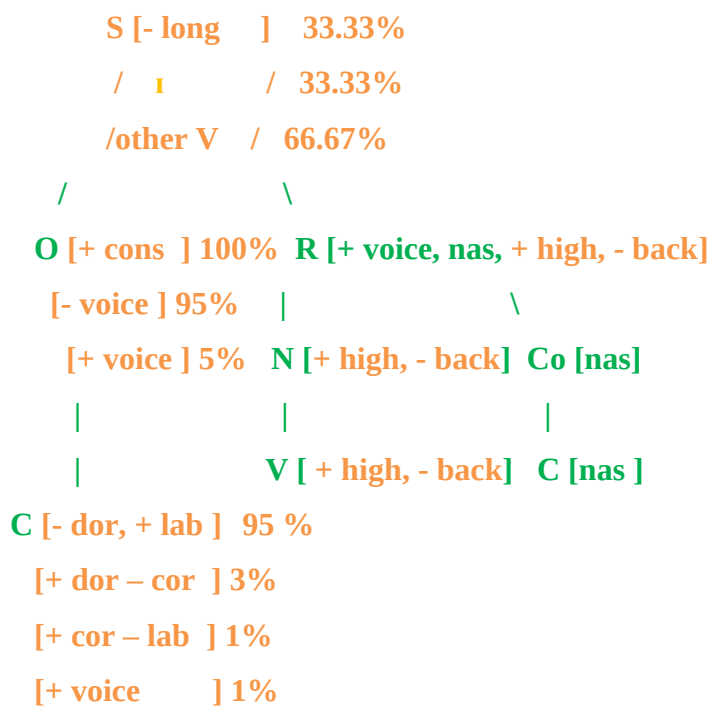


Figure 98: Abstraction of ‘CɪNs’ at t + 10ms

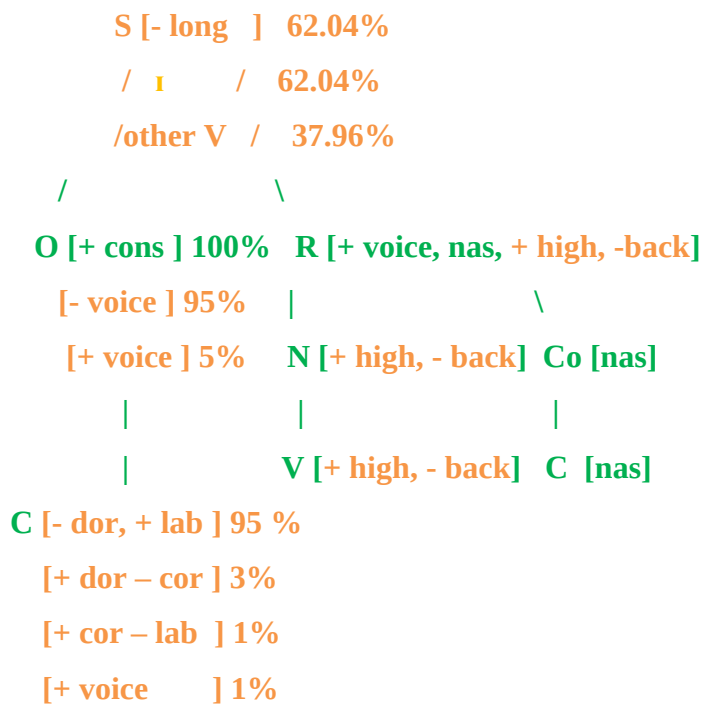


Figure 99: Abstraction of ‘CɪCs’ at t + 10ms

In sum, the listener has a much better chance of deducing the underlying vowel quality from ‘CɪCs’ rather than ‘CɪN’ stimuli (the difference in recognition being substantial at 33.33% contra 62.04%). We can conclude that especially very early on, high front vowels may exhibit relatively strong nasalisation and/or significant amounts of introduced nasal zeroes, as suggested in the accompanying discussion to figures 95-96. This type of FPD can affect recognition to the extent that the aperiodic phase in a preceding onset is distorted in its FPD. Such acoustic distortions make it much harder for the listener to deduce the phonetic identity of the upcoming vowel, since the formant structure of the vocalic portion is obscured (cf. figures 95-96).

We should note that such phonetic influence does not mean that we should model the potential spreading of FPD from resonant sounds (such as nasals and liquids) at the syllabic level. Rather, the finding shows that the claims made by Goffman et al (2008), which were introduced in chapter 1 enjoy

good validity and a good grounding in the actual realisations of speech rather than in canonical forms. This exemplification does not apply to the findings of this chapter alone. We now move on to time slot 3.

5.4.4 Recognition at time slot 2 for 'pin' (plosive burst + 20 ms accompanying vowel resonance)

At time slot 3, 'pin' displays the following phonetic exponents:

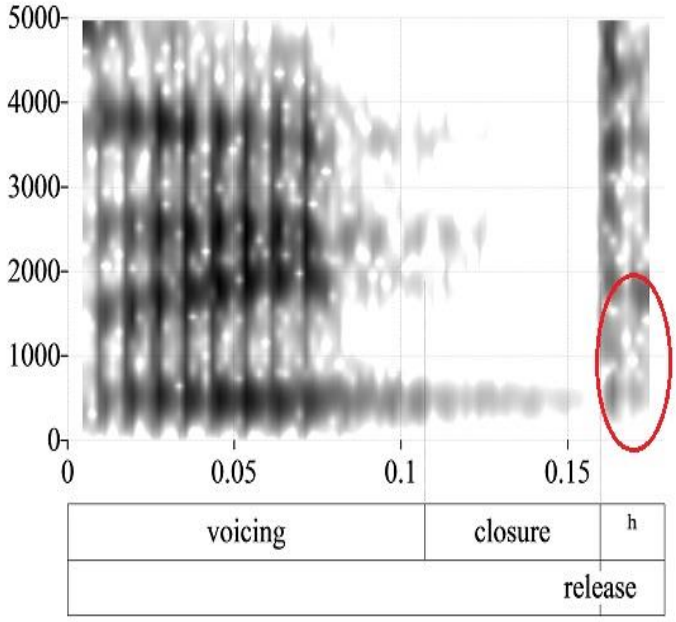


Figure 100: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **pi**n' (20ms gate)

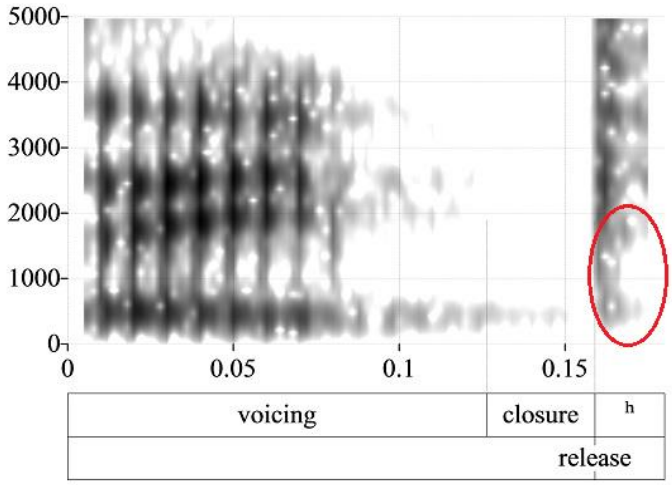


Figure 101: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **pi**t' (20ms gate)

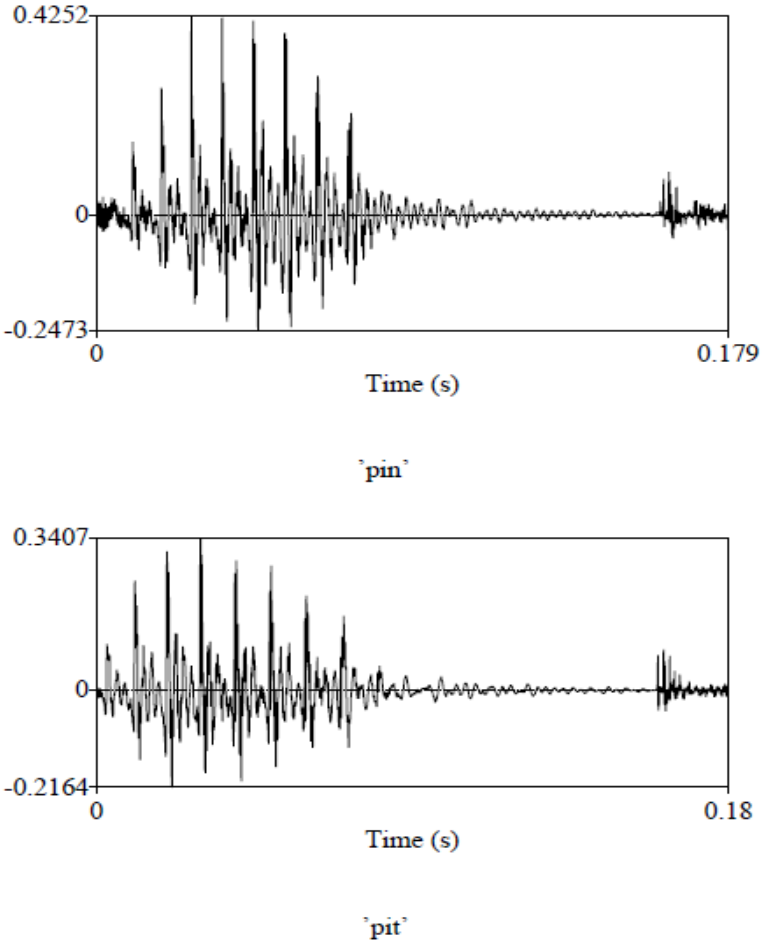


Figure 102: Partial stimulus waveforms for 'pin' (top) and 'pit' (bottom) at t + 20ms

When we compare the initial parts of the aperiodic phases of the onset plosives on the right-hand sides of figures 100-101 and the waveforms in figure 102, we can make some important phonetic observations already at t + 20ms. For example, between ca. 0.16-0.17 seconds in ‘pin’ (cf. the areas inside the red circle in figure 100), the listener does not yet have access to an emerging mid-frequency peak having low acoustic energy (see e.g. Harris and Lindsey, 1995). This acoustic distinction between ‘pin’ and ‘pit’ suggests that listeners are faced with a more significant challenge in recognising vowel quality from CVNs, as the temporal evolution of the formant trajectory in ‘pin’ is not as transparent to the listener as in ‘pit’. On the other hand, in ‘pit’, such a mid-frequency peak emerging between ca. 0.16-0.18 seconds *can* be evidenced between ca. 800 and 1700 Hz (cf. area inside the red circle in figure 101). What abstractions can be made for CɪCs contra CɪNs at this point in time? Figures 103-104 contrast these recognition probabilities:

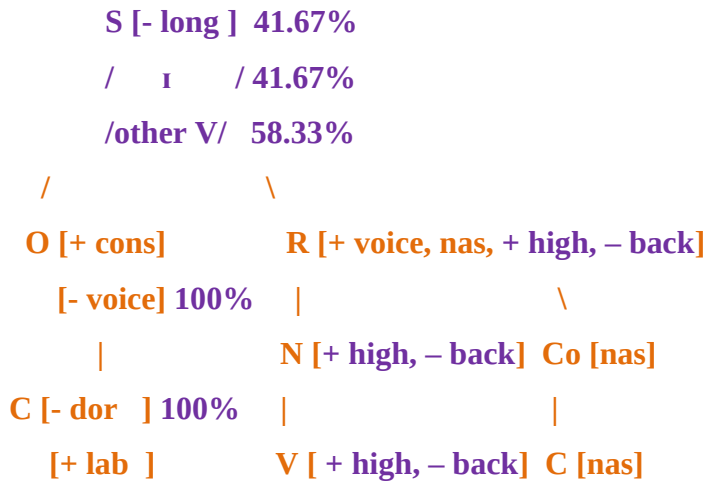


Figure 103: Abstraction of ‘CɪNs’ at t + 20ms

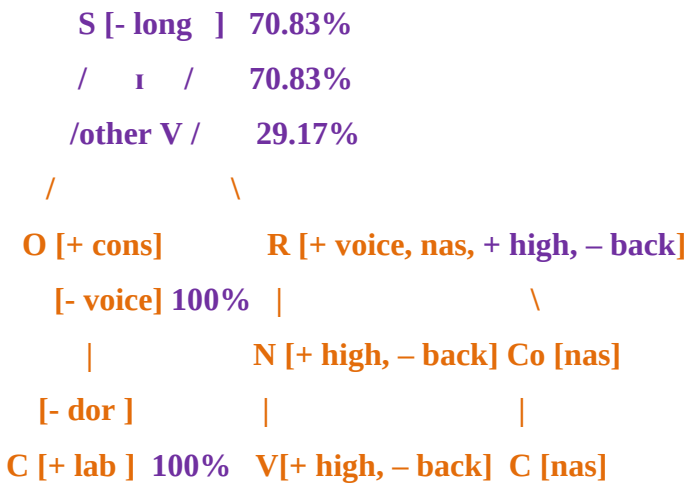


Figure 104: Abstraction of ‘CɪCs’ at t + 20ms

At time slot 3, the difference in the probabilities of arriving at the correct abstraction is ca. 28%, being slightly smaller than at time slot 2 (41.67% vs. 70.83%, respectively for ‘pin’ contra ‘pit’). Since the incremental difference in recognition between time slots 2 and 3 contra 3 and 4 amounts to only ca. 1.5%, we can draw similar conclusions as at time slot 2: the increment in the reliability is slightly higher for CɪNs than for CɪCs. Since the difference in the recognition reliability proportionally remains almost ¾ of the probability for the CVN (i.e. ca. 28% divided by 41.67%), the claims made at time slot 3 receive good support. That is, the type of FPD for ‘pin’ displayed in figure 98 *does* affect listener ability to achieve reliable vowel recognition early on. It would seem odd in the extreme not to take a ca. 30% difference so early on in time into account in theoretical terms. We will now look at time slot 4.

5.4.5 Recognition at time slot 4 for 'pin' (plosive burst + 30 ms vowel resonance)

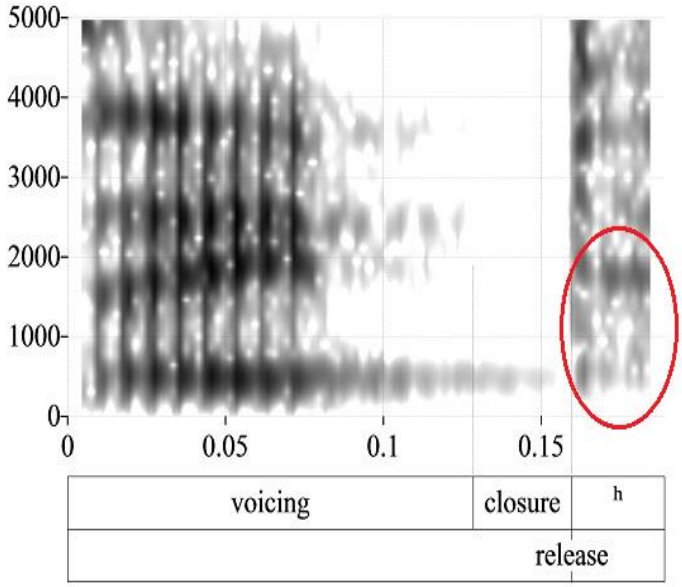


Figure 105: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **pi**n' (30ms gate)

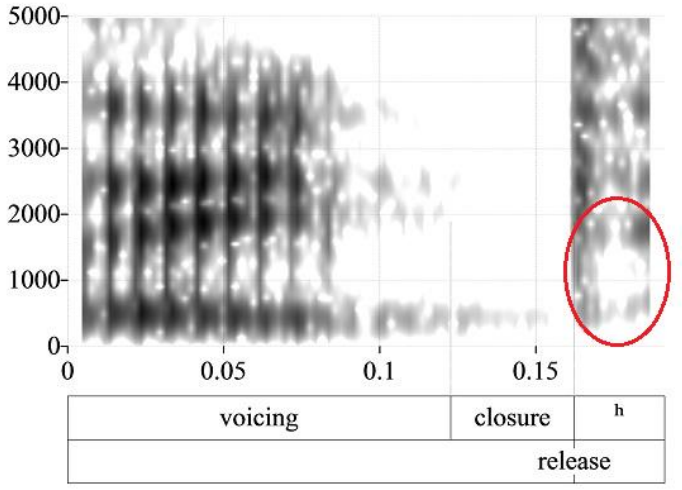


Figure 106: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **pi**t' (30ms gate)

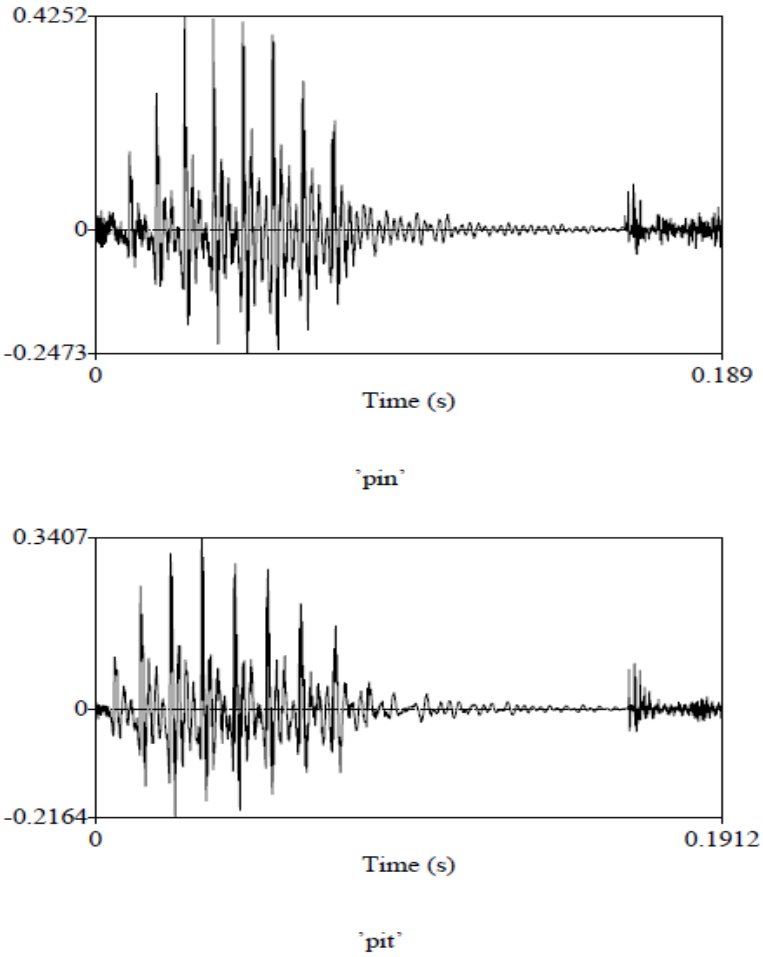


Figure 107: Partial stimulus waveforms for 'pin' (top) and 'pit' (bottom) at t + 30ms

When we observe the equivalent parts of the aperiodic phases at time slot 4 in figures 105-106, we can draw similar conclusions as at slot 3. For example, F2 in /ɪ/ in 'pit' (cf. figure 104) has started its descent towards the vowel's steady state portion. For 'pin' (cf. figure 105) we can still not observe as clear a trace of an emerging main formant pattern (cf. the areas inside the red circles in figures 105-106). These two pieces of production data as well as the findings presented in 5.4 thus far support Harris and Lindsey's (1995) claims about the perceptual significance of the mid-frequency peak with low acoustic energy in the recognition of high front vowels. Since listeners do not have as clear access to such a peak from the aperiodic phase in CVNs, the time course and reliability of

recognition are delayed and affected negatively. The recognition probabilities displayed in figures 108 and 109 support this claim:

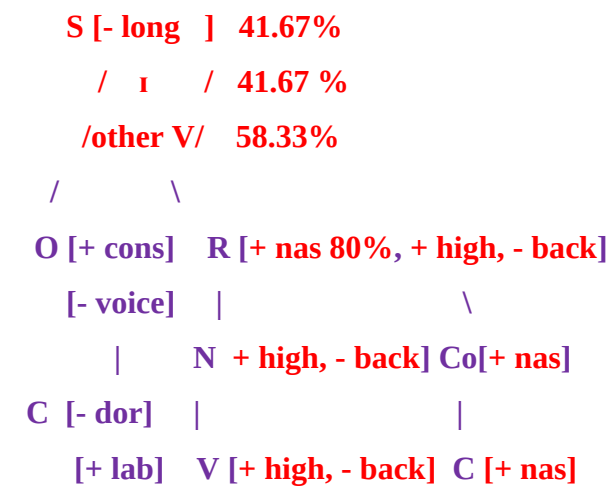


Figure 108: Abstraction of 'CɪNs' at t + 30ms

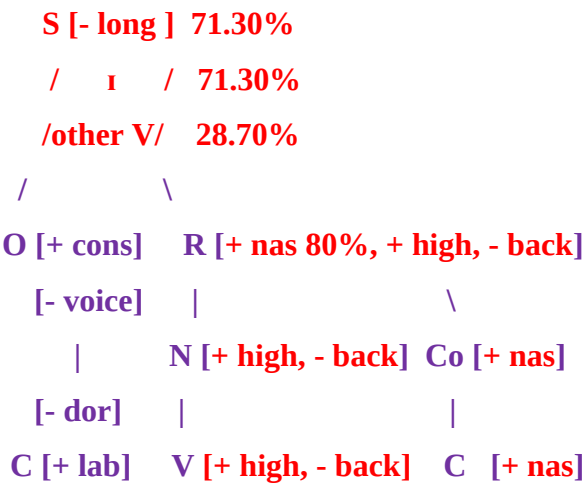


Figure 109: Abstraction of 'CɪCs' at t + 30ms

From the recognition probabilities displayed in figures 108-109, two conclusions can be drawn. The first one relates to the increment in recognition: the recognition reliability for the CɪN has improved ca. 5.5% from time slot 2, whereas that for CɪCs has remained constant at 71.30%.

The second conclusion relates to what this distinction can tell us about the perception of CɪNs. The comparison shows i) that recognition of CVNs as compared to CVCs functions differently at distinctive points in time, and that ii) the increments in recognition reflect this difference. For example, this finding also reinforces Nearey and Assmann's (1986) claims on the perceptual importance of the 30ms locus point for vowel perception, since recognition for CVNs has become more reliable, which is not true for CVCs. Since recognition is delayed through time for CVNs, adding temporal information is still perceptually significant at t + 30ms. This claim does not apply to CVCs, since they are devoid of nasality. Although this conclusion complicates the modelling of recognition from CVNs, it shows some of the limitations of our knowledge of anticipatory nasalisation and especially the extent to which it can affect the time course of vowel recognition. We now move on to look at time slot 5.

5.4.6 Vowel Recognition at time slot 5 for 'pin' (plosive burst + 40 ms accompanying vowel resonance)

Can we distinguish any other significant differences ca. halfway through the aperiodic phase in 'pin' and 'pit'? We can answer this question by inspecting the first 40ms in the aperiodic phases of the onset plosives in two instances of /ɪ/, one of which is [+ nasal] and the other [- nasal] (cf. the right-hand sides of figures 110-111 and waveforms in figure 112):

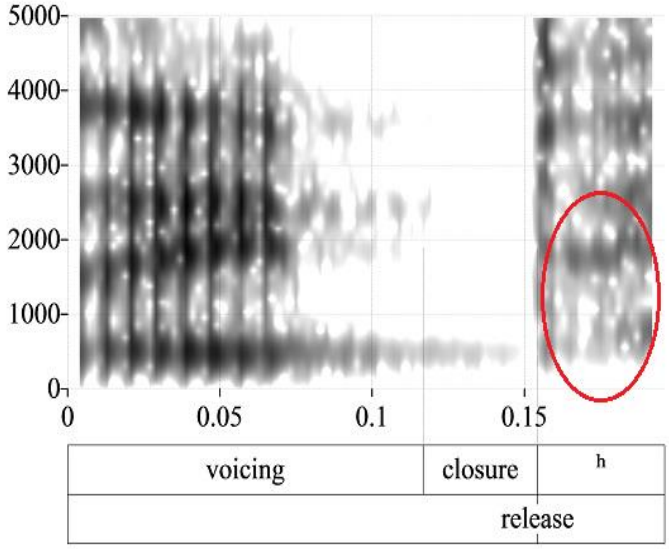


Figure 110: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **pi**n' (40ms gate)

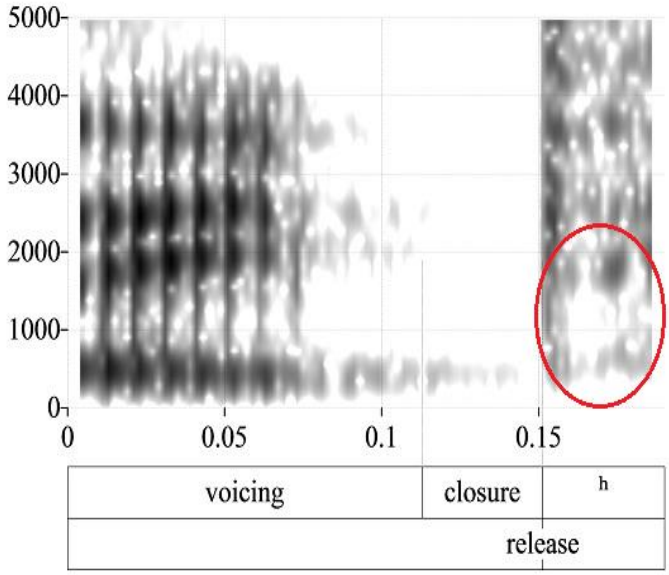


Figure 111: A partial segment of the '[eɪ p^h]' portion in 'I think you s**ay** **pi**t' (40ms gate)

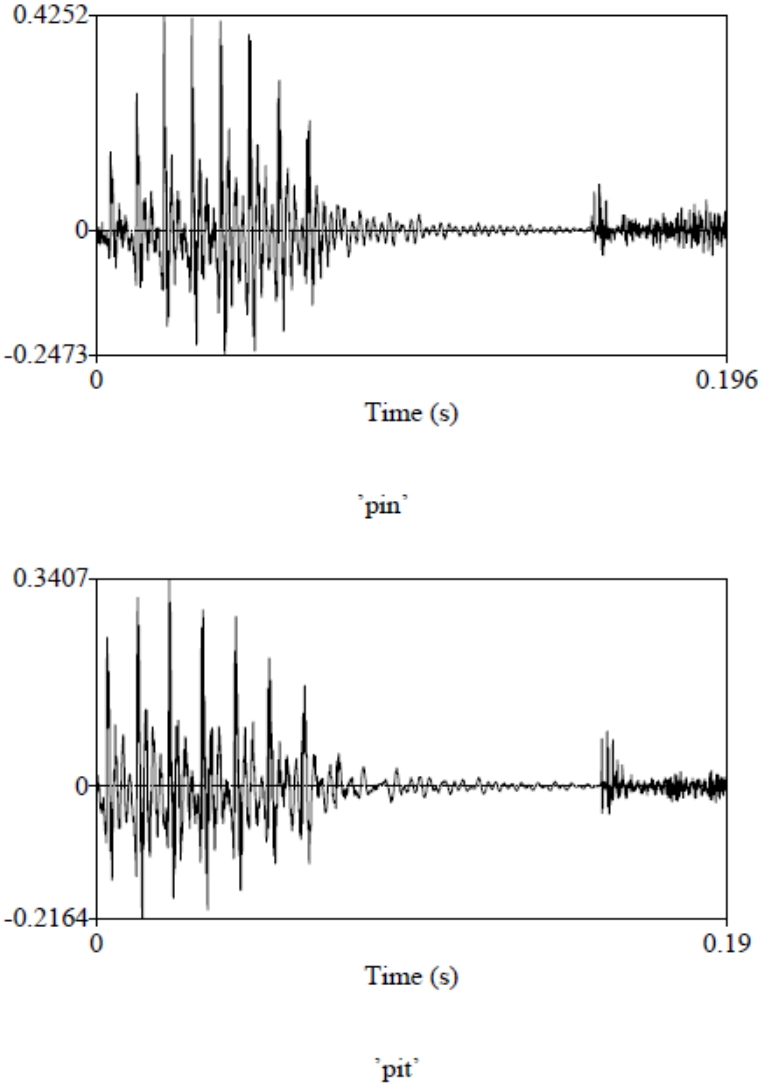


Figure 112: Partial stimulus waveforms for 'pin' (top) and 'pit' (bottom) at t + 40ms

When we examine the aperiodic phases of the onset plosives on the right-hand sides of figures 110-111 and the waveforms in figure 112, we can start to distinguish the emergence of comparatively more similar formant trajectories around 0.16 seconds in 'pit' and 0.18-0.19 seconds in 'pin' than at earlier time slots. When we examine the phonetic evidence available in the spectrograms in figures 108-109, we can also see that in the nasalised part of the aperiodic phase (cf. figure 110), the area for the typical mid-frequency peak has a higher F2 in 'pin' than in 'pit' (cf. the areas inside the red circles at ca. 800Hz

contra 1100Hz). This contrast means that the underlying formant structure in the nasalised vowel is somewhat obscured. For example, when comparing the more uniform resonance properties of the aspiration at 0.15-0.19 sec in ‘pin’ (cf. figure 110) with the more easily distinguishable formant structure at 0.15-0.19 sec in figure 109 we receive relatively good evidence for the suggestion that listeners find it harder to recognise vowel quality from CVNs, even at t + 40ms. Let us compare the recognition probabilities for ‘pin’ and ‘pit’ at time slot 5:

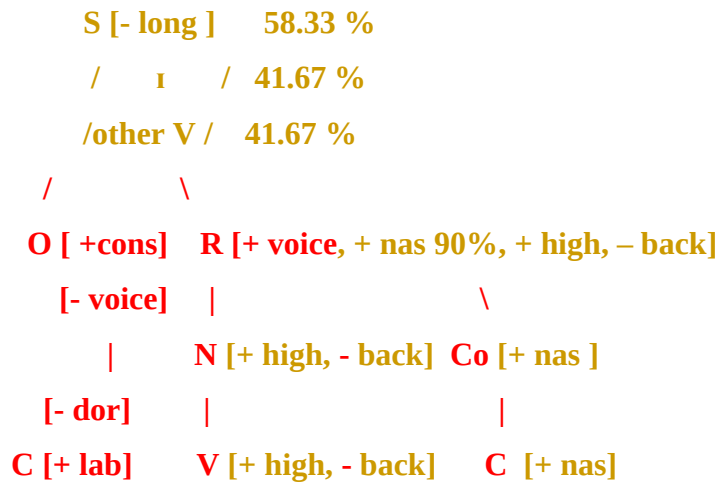


Figure 113: Abstraction of ‘CɪNs’ at t + 40ms

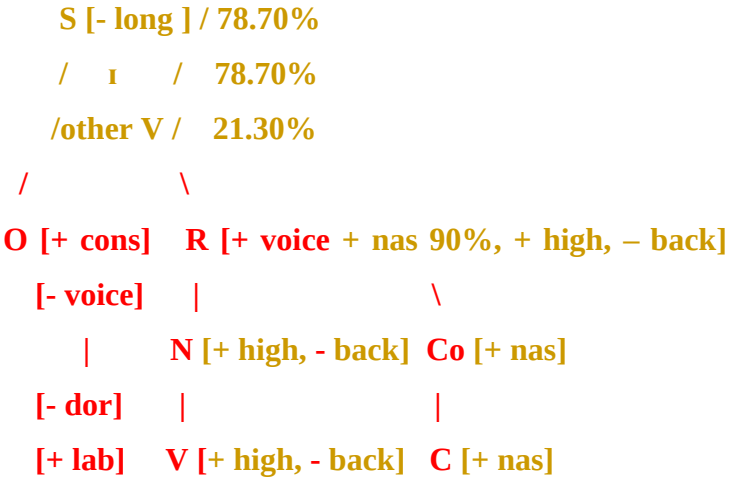


Figure 114: Abstraction of ‘CɪCs’ at t + 40ms

When we compare the recognition probability for ‘pin’ and ‘pit’ in figures 113-114, the difference in the reliability of recognition is ca. 20% (58.33% vs. 78.70%). Although it is difficult to say whether this difference is as theoretically significant as at earlier time slots, the difference observed at t + 40ms is much smaller than at shorter gates. For example, at t + 30ms the difference in recognition reliability between ‘pin’ and ‘pit’ is ca. 30%, which is 10% more on absolute level and ca. 50% proportionally. Considering this relatively large difference in recognition between nasal and non-nasal /ɪ/ at time slots 4 and 5, we receive additional support for the claims made on vowel recognition timing by Nearey and Assmann (1986) and also the claims made by other researchers on the production and recognition of nasalised vowels (e.g. Cohn, 1990).

Having fully discussed and exemplified the recognition differences between [+ nasal] contra [- nasal] vowels in CVNs contra CV-/p t k/ monosyllables, we will briefly discuss to what extent vowel quality in CVNs might be a significant factor in vowel recognition. Rather, we will focus on nasality in back vowels in CVNs.

5.4.7 Backness and Nasalisation in CVNs

There is no previous research on how backness might affect the perception of English vowels from plosives with nasalised aspiration. This may not be an obvious research question to ask, but since we cannot relate the findings on this aspect of recognition to previous research, we need to be more speculative about the results. Since the two back vowels’ realisations differ in northern and southern accents, this claim can be justified on theoretical grounds. It is not equally worthwhile to describe the abstractions made by listeners for

backness for this reason. Such explorations are best left for future research.

Figure 50 in chapter 4 shows that all front nasalised vowels had lower average recognition values (ca. 45.1%, 48.6% and 28.2% respectively) than in the context of the southern and northern variants of 'pun' (55.6% and 50.7% for northern /ʊ̃/ and southern /ʌ̃/, respectively). Although the number of stimuli for /ʊ̃/ /ʌ̃/ is half for that of their front vowel equivalent, the proportional recognition differences between back and front [+ nasal] vowels is quite large. It would seem odd to assign this difference due to the smaller number of stimuli for /ʊ̃/ /ʌ̃/. A more likely possibility is that the articulatory constellation for back vowels with different height values is not as conducive to nasalised aspiration as in /ã/ and /ĩ/. The most important finding in this context is the fact that the average recognition value for /ʊ̃/ is much higher than for its front counterpart, despite the fact that F1 and F2 will be much nearer to each other in /ʊ̃/ compared to /ĩ/. Therefore, the presence of nasality can be seen to be reflected in a reverse the timing of vowel recognition proportions, as in its absence oral /ɪ/ engenders more reliable recognition (ca. 64.32%) than oral /ʊ/ (54.58%, cf. subsection 4.5).

In summary for subsection 5.4, the recognition of nasalised vowels is not equal to that for oral vowels:

a) Speakers often tend to nasalise non-mid front vowels, which has significant effects on listener ability to recognise vowel quality early on.
b) Consequently, it takes at least 10ms longer for listeners to work out vowel quality as reliably as for oral vowels.

We now move on to summarise and evaluate the results and findings outlined and discussed in this chapter: in particular, we will consider the applicability of the model to other CV(V)/Cs and other languages, as well as coarticulatory models more generally. We will also highlight the model's agreement with findings in the previous literature, as well as the main finding on access to durational cues (such as VISC) and the representation of length. Some caveats concerning the extent to which the findings can be generalised are described.

5.5 Summary of Chapter 5 and the Model Behind Phonological Processing of vowels in CV(V)/Cs

A good and suitable assessment of the phonological processing of CV(V)/Cs requires us to show to what extent listener abstractions are sensitive to the phonetic exponency of monosyllabic utterances. Given a particular way of producing acoustic detail in CV(V)/Cs, listeners necessarily need a set of concrete declarative rules (as presented in 5.2) in order to be able to work out the interrelations between different sounds and constituents in a monosyllable, as well as how these dependencies shape the phonetic exponents of sounds at different points in time.

This chapter has presented a model of phonological processing, which expresses how listeners map from phonetic detail to phonological structures, using the same rules for production as for perception. The importance we should attach to vowel recognition from CV(V)/Cs can be summarised the way Polysp would have it (see e.g. Hawkins, 2003, 2010a, 2010b, Hawkins and Smith, 2001): if the sounds differ in an utterance, then structural factors and properties of such utterances must differ. For example, given more variable VISC variation during the aperiodic phase, listeners may have a way

to work out the underlying syllable structure reliably quite early on with a high degree of probability.

We still need to consider vowel recognition from a broad viewpoint, which attaches equal emphasis to phonetic and phonological properties of monosyllabic utterances, and which pays sufficient attention to phonetic detail. For example, the rules and figures exemplifying phonological processing in 5.2 may help to demonstrate the importance listeners attach to language/variety-specific coarticulatory and listening strategies. In this respect, the model developed in this chapter has answered the secondary research questions (see 2.5) in detail, whilst giving relatively straightforward answers to the primary research topic (see 1.2).

In sum, vowel recognition from CV(V)/Cs requires an explicit model that displays sensitivity towards both subtle and broader aspects of phonetic exponency and representation in phonological processing. The prosodic model developed in this chapter well exemplifies the reasoning and claims made on coarticulatory strategies and VISC in chapters 1-2. Two important new findings have been highlighted in this chapter,

i) on what level is length to be represented at (= the syllabic level rather than at the nucleus) and

ii) how temporal processing of vowels in CVNs can be delayed in the absence of clear access to the underlying formant pattern in CVNs.

5.6 An Evaluation of the Model on Vowel Recognition and Phonological Processing of CV(V)/Cs

It may be possible to generalise many of the findings to other languages, and in particular other varieties of English. This claim is particularly evident to the extent that coarticulatory

strategies and the phonetic patterns concerning VISC and nasality are similar. For example, Nearey and Assmann (1986) have already shown this for VISC in Canadian English vowels.

The model in this chapter has a broad scope from the viewpoint of vowel recognition, and in particular the complex properties we should attach to coarticulation and phonological processing. The model is similar to Polysp and probably equally generalisable. The model makes more specific predictions about the relationship between the bidirectionality of coarticulation and phonological processing. Since CVCs have been a source of great interest in recent research (and especially in Polysp, see e.g. Hawkins and Nguyen, 2001, 2004), this extension of Polysp is theoretically significant.

There are certain caveats to these claims. For example, it is not clear to what extent the results can be generalised to more complex syllable shapes, such as CVVN (e.g. 'corn'), CVVCN ('can't) and especially CCV(V)C(C) syllables (such as 'cringe' and 'scratch' in English, since their underlying VISC patterns as well as the required coarticulatory strategies in such syllables will differ (see e.g. Docherty, 1992). The general principles of the model developed in this chapter *can* be used for research into coarticulation and vowel timing, a claim which might also be generalisable to other languages (cf. e.g. the research on French and Taiwanese CVNs by Chang et al, 2011).

Having fulfilled and evaluated all the main aspects relevant to phonological processing in CV(V)/Cs we will round up the thesis in chapter 6.

6. Conclusion

6.1 A Summary of the Results

The following six points summarise the key results:

1) Vowel quality can be recognised reliably early on from the aperiodic phase of English aspirated voiceless plosives: ca. 30-35 ms into the aperiodic phase of the onset portion, recognition becomes significantly significantly more reliable (cf. 4.3.1).

2) The phonetic exponents of the onset, nucleus and coda all have a significant bearing on recognition and feature sharing in CV(V)/Cs:

a) The phonetic encoding of length for long vowels in CVVs and for short ones in CVNs and CVCs differs significantly, so that long and low vowels are more variable spectro-temporally in VISC than short and high vowels. This claim can be explained by the fact that consonantal exponents are overlaid on vocalic ones (e.g. Coleman, 1990, 1998). In sum, the time course of recognition reflects the encoding of VISC. This process takes effect so that short and high vowels tend to be recognised earlier than long and low ones.

b) Nasalisation from the coda portion into the exponents of the rime and the aperiodic phase of the onset portion significantly affects their FPD, so that the main formant patterns for F2, F3 and in particular F1 are obscured and/or dampened. The potential introduction of nasal zeroes contributes to such

acoustic-perceptual distortions, whilst additional nasal poles may make it more difficult for listeners to deduce the underlying formant relationships. This can have significant effects on the time course of recognition, reflecting the fact that the phonetic quality of FPD in CVNs comprises the main cue to vowel recognition.

3) High vowels offer better cues to recognition than low from the aperiodic phase. This result is explained by the general coarticulatory resistance that low vowels undergo: increasing the opening of the jaw requires additional temporal and physical adjustments to the articulation of CV(V)/C)s. This acoustic aspect is mirrored in a delay in recognition in word stimuli such as ‘par’, ‘cat’ and ‘top’.

4) Velar and bilabial onsets give more reliable cues to vowel quality than alveolar ones, which do not display a high degree of coarticulation. Since bilabials have no intrinsic tongue posture and velars display a high degree of coarticulation (with a wide area of contact between the tongue back and hard palate), vowel recognition can be achieved much earlier from these plosive sounds.

5) Phonetic and phonological context strongly affects recognition, regardless of sociolinguistic and extralinguistic factors. For example, the syllable shape underlying a gated stimulus significantly affects recognition in distinctive ways (see e.g. figures 49 in chapter 4).

6) The findings on long-domain coarticulation and vowel length are consistent with the previous literature on non-segmental phenomena in CV(V)/C syllables.For example, we would expect the kinds of findings by Cohn (1990) and Chang et al (2011) and on the co-extensiveness of coda nasality and aspiration in onsets to also be reflected in vowel recognition (and not just in the acoustics). The fact that consonants are overlaid upon vowels and affect their entire realisations (Coleman, 1990, 1998) reflects the functional encoding of length in CV(V)/Cs (also see Xu, 2009), which has implications for at what level vowel length should be represented phonologically. Thus, it is not surprising to confirm the perceptual significance of effects of vowel length and nasality, as they have already been deemed significant in terms of production in previous research on the *production* of CV(V)/Cs specifically.

6.2 Implications

Although the methodology, theoretical framework applied and the findings owe a lot to FPA, Polysp and DP, this research delves deeper than any of these theories in some respects, in particular with respect to the level of representation and exponency of length and the perceptual significance of coarticulatory strategies. Polysp, which is the most theoretically versatile and modern of these three theories, does not pay sufficient attention to the potential perceptual significance of coarticulatory distinctions and coarticulatory strategies. In particular, the thesis helps to show that the kinds of non-segmental effects noted by previous studies on long-domain resonance are not restricted to continuant sounds. Polysp does

not say a great deal about extending such findings to more complex syllable shapes and/or articulations (though see earlier research on production by Hawkins and Slater, 1994). This extension of non-segmental phenomena in English from liquid-V-C monosyllables to other more complex syllables shapes forms one of the main innovations of this research. Despite being indicative as a finding, the main secondary finding in this thesis on CVNs helps to show that we must not underestimate the significant amount of FPD that is needed in modelling coarticulatory phenomena in monosyllabic utterances. The same claim applies to the perceptual role of VISC as well, since moment-to-moment variation in vowel formant centre frequencies can have significant effects on recognition, as the discussion in previous chapters has shown.

However, we must also appreciate that the relationship between feature sharing and the spreading of exponents may be much more complex than previous studies suggest. For example, as has been shown in chapter 5, the phonetic encoding of nasality may have phonetic effects on sounds located 2-3 constituents away from the nucleus/coda in a CVN. Does such a result mean that we should specify spreading rules for such forms of phonetic influence? Such a proposition would be very hard to justify. Rather, it may simply be that feature sharing as a concept is much more complex in phonetic terms than previously envisaged. We now consider potential further research questions and directions for future research arising from this study on vowel timing and recognition in English.

6.2.1 Future Directions

This study has aimed to fulfil a gap in linguistic theory. The research questions have been answered in detail and I have provided a robust theoretical account of the phonological and phonetic phenomena that are associated with vowel

recognition. However, some of the methodological concessions and choices that had to be made (e.g. only having young native speaker-listeners and allowing for maximal articulatory freedom) leave many equally interesting questions unaccounted for. Therefore, one of the main aims of acting as a springboard for further research has been fulfilled in this study.

It is hoped that the theoretical framework and especially the main findings will help to broaden researchers' views of linguistics and of speech perception, both in a theoretical and in a more general sense. Even though the purpose of this research has little to do with FPA and Firthian linguistics as such, some of the claims made on the representation of vowel length, prosodies, and e.g. non-segmental phenomena in CVNs in this research have strong ties with the ideas of this 20th century form of British linguistics.

6.3 Suggestions for Further Research

This study advocates subsequent research to look at:

i) how the perception of diphthongs differs from the perception of monophthongs (though see Howell, 1981), and whether this aspect might apply distinctively to other varieties of English than ones spoken in England (such as GA or Australian English),

ii) how nasal place of articulation may affect vowel recognition (also cf. Chang et al, 2011),

iii) how noise and obstacles affect the perception of coarticulation,

iv) whether perception of coarticulation applies equally to conversational speech (also see Ostreicher and Sharf, 1976),

v) whether variables such as age, gender and social affiliation affect the perception of coarticulation (also see Nittrouer, 2007 and Parnell and Amerman, 1978),

vi) what the practical significance of recognition might be (whether technologically or clinically),

vii) how phonological and phonetic variation may influence the perception of coarticulation, especially as far as social phonological contrasts are concerned (cf. e.g. Foulkes & Docherty, 2006 and Ogden, 2006).

Other studies on the perception of coarticulation should investigate listeners' capability of recognising FPD in online lexical processing in more detail. For example, it would be interesting to know how far the coarticulatory effects of glides, fricatives, ejectives and clicks extend in English accents that have such sounds, considering their robust perceptual and acoustic properties (see e.g. Stevens, 1998). Such studies on different kinds of articulations could also form a good aid in developing an exemplar theory based on non-segmental phonology, since such an approach would allow more optimal modelling of how the perceptual system responds to qualitatively different speech stimuli.

Appendices

Participant Recruitment E-mail

The following e-mail was sent to the 24 listeners who took part in the perception experiment described in chapter three:

So just to sum up, your task in the experiment is to make a 4-way choice for the last word in sentential stimuli based on what you've heard so far.

For example, you might hear something like I think it's a

A t(urn)
B t(arn) > (poetic for 'lake)
C t(orn)
D t(een)[31]

That's more or less what it is. All the instructions are contained within the first few pages of the experiment and you need to answer a few questions about your age and where you were brought up (etc.) too. You also need headphones as well as real/flashplayer (or quicktime) to play the sound files.

Here's the link for you:

http://edu.surveygizmo.com/s3/851989/Perception-of-Vowels-from-Consonants-2

Best wishes,

Kaj

[31] Since CVVC words did not occur in this research, they comprised a good familiarisation set for potential listeners.

Stimulus Recording Sheet

As part of the recordings for the perception experiment described in chapter three, the following list was given to participants:

I think you say 'tock'
I think you say 'tap'`
I think you say 'pap'
I think you say 'top'
I think you say ‘tin’
I think you say ‘can’
I think you say 'pup'
I think you say 'cap'
I think you say ‘par’
I think you say ‘pan’
I think you say 'cut'
I think you say 'cock'
I think you say 'cack'
I think you say 'puck'
I think you say 'cat'
I think you say ‘tea’
I think you say 'cot'
I think you say ‘cun'
I think you say ‘pen’
I think you say ‘coo’
I think you say 'cop'
I think you say ‘car’
I think you say 'tuck'
I think you say 'kip'
I think you say 'tot'
I think you say ‘pun'
I think you say 'pip'

I think you say 'tut'
I think you say 'pick'
I think you say ‘pit’
I think you say ‘pea’
I think you say ‘two’
I think you say 'putt'
I think you say 'tup'
I think you say 'tat'
I think you say 'pock'
I think you say 'cuck'
I think you say 'tack'
I think you say 'kick'
I think you say 'tick'
I think you say ‘tar’
I think you say ‘pot’
I think you say ‘pin’
I think you say ‘key’
I think you say ‘core’
I think you say ‘tan’
I think you say ‘pat’
I think you say 'pop'
I think you say 'pack'
I think you say ‘ten’
I think you say 'tip'
I think you say ‘paw’
I think you say ‘tore’
I think you say 'ken'
I think you say 'ton'
I think you say 'kit'
I think you say 'cup'
I think you say 'poo’
I think you say ‘kin’
I think you say 'tit'

Nothing else was written on the recording sheet, however the speakers were asked to read the following set of instructions before each recording was initiated:

This test is designed to investigate the speech production patterns of speakers of English English, with particular reference to 1) consonants produced with plosion [p t k] and 2) vowels. The tests will be performed at the recording studio at a time to suit each participant.

After the recording equipment has been switch on by the experimenter, each participant will be asked to read each sentence written on a standard A4 paper using a standard and neutral intonation and rhythm. For theoretical reasons, it is very important that the sentences are produced as similarly as possible, especially with respect to intonation, rhythm and voice quality. Thus, the sentences should be read out clearly without hesitations and/or lengthy pauses in between each sentence. However, participants should still take time to produce each sentence neutrally and adequately. In other words, a brief (e.g. 2/3 second) pause must be reserved between the production of each sentence.

Participants are free to take as much time as they wish to complete the speech production test. The speech production test should take a maximum of 10-15 minutes.

Each speaker was also asked to fill in the following consent form:

Department of Language & Linguistic Science,
University of York
Heslington, York, YO10 3DD
***tel*: +44 1904 432650• fax: +44 1904 432673**

Consent to participate in research

Speech Production and Processing Research for PhD project Investigator: Kaj Nyman (supervised by Dr R. Ogden and Prof. J. Local)

I agree to take part in this test. I have been selected as a participant because I volunteered to take part.

I acknowledge that the investigator has explained

- what is involved in the test; ☐
- the purpose of the work in this area; ☐
- his commitment to preserving the anonymity of participants; ☐
- his commitment to using the information supplied by the test-subjects with confidentiality and impartiality. ☐

I am aware that I may withdraw my participation at any time, and that I am under no obligation to complete the required task. If I decide withdraw from the study, my data will also be removed.

I have had the opportunity to ask questions about this test, and I have received answers. I will also receive a general A4 size description of my contribution to the study once the analysis process has been completed by the researcher.

I am also aware that I agree to allow this data to be used for general linguistic research purposes (e.g. conference presentations). The data will be held indefinitely by the researcher, as the results generated by the experiments are not sensitive.

Signed ……..…

Name in block capitals……

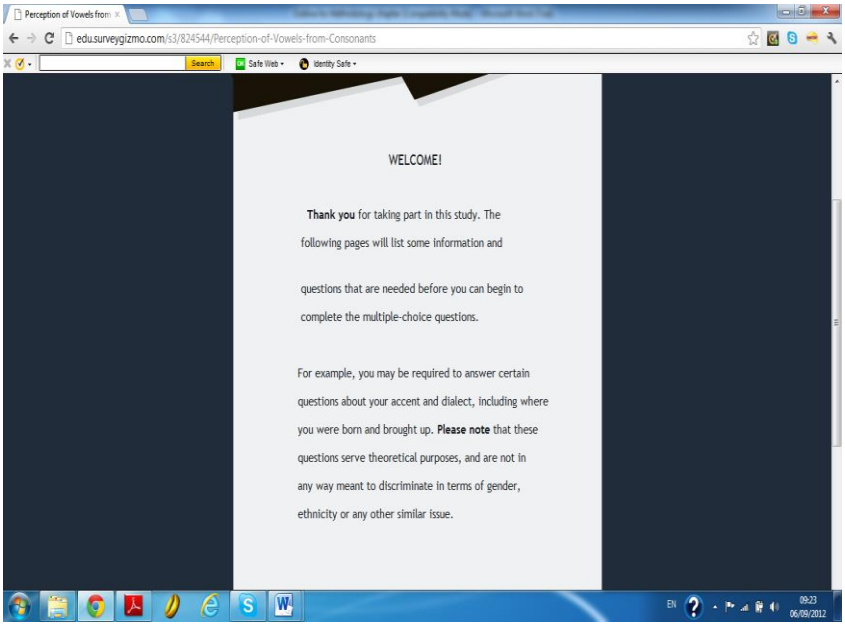
Researcher's signature

Date ……..

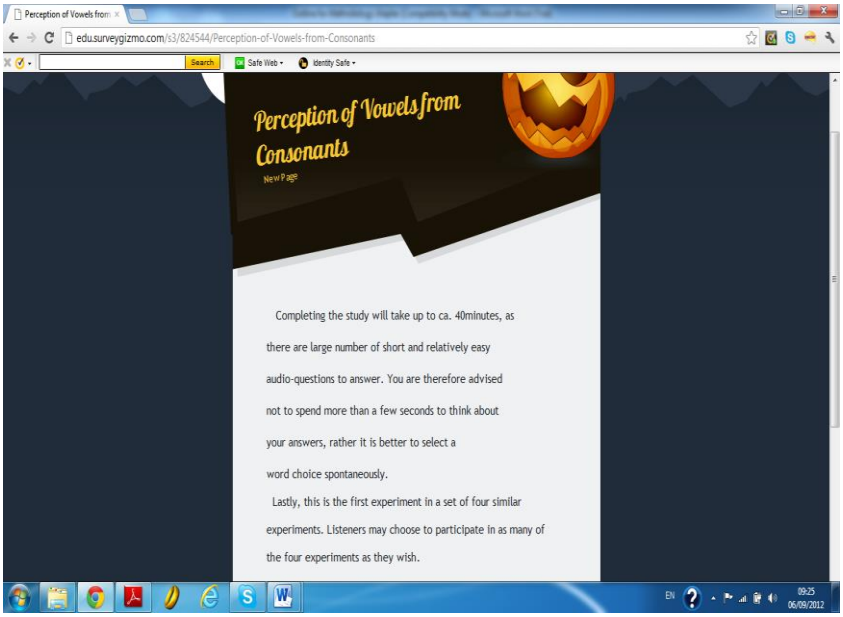
Approved by the University of York.

Experiment Sequence

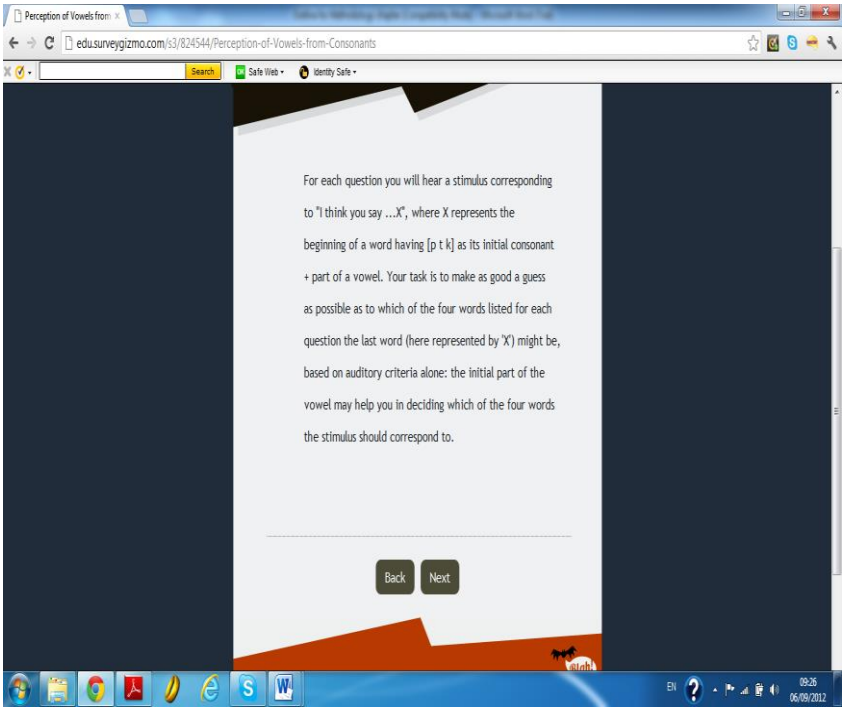
The following screenshots describe the continuation of the experiment as experienced by the 24 listeners participating in the perception experiment described in chapter three:



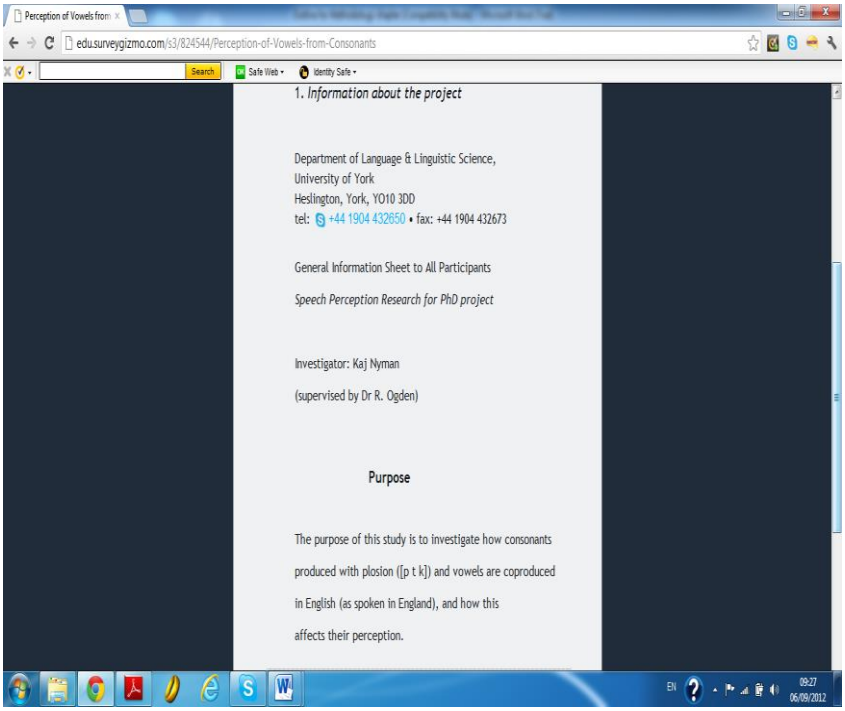
Page one:



Page two:

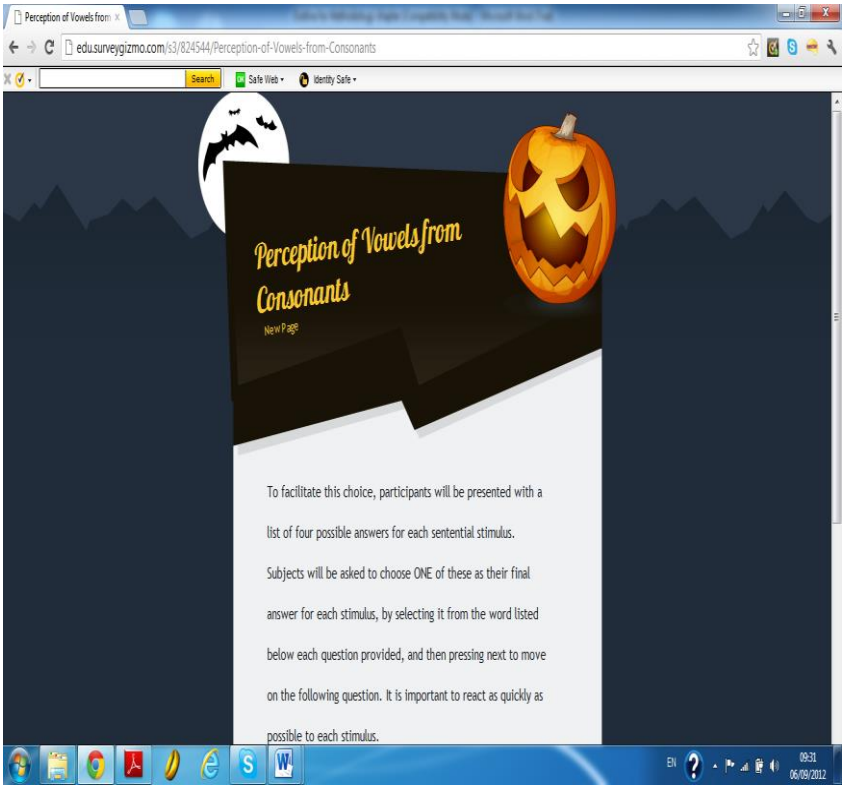


Page three:



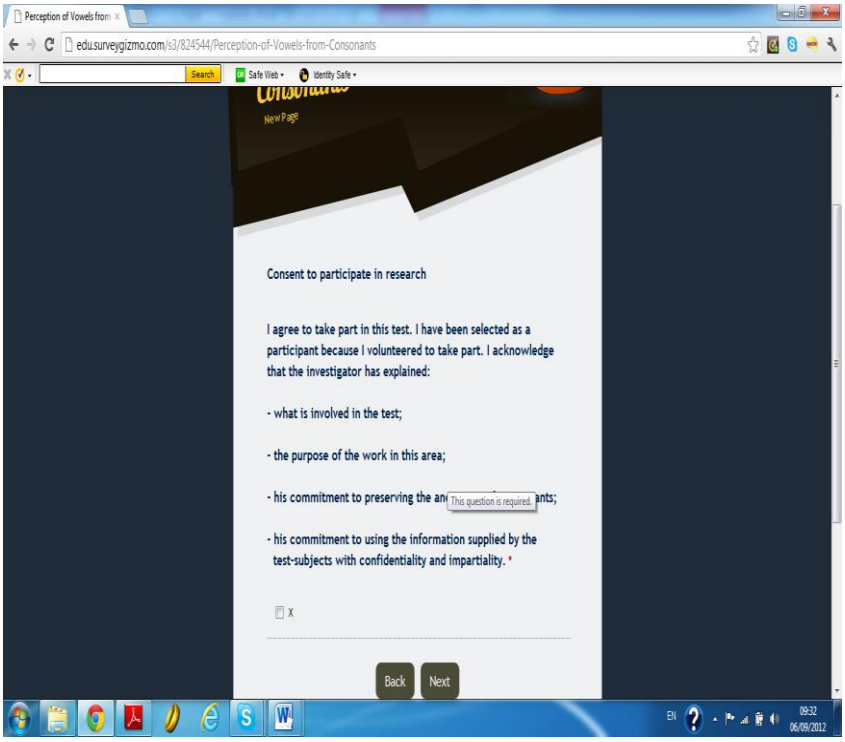
Page four:

Page five:

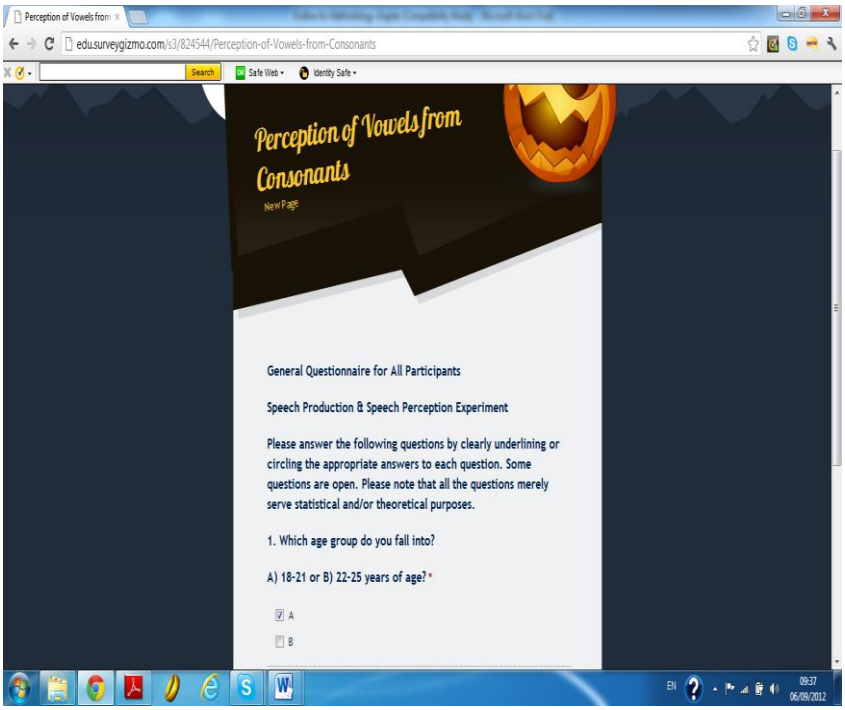


Page six:

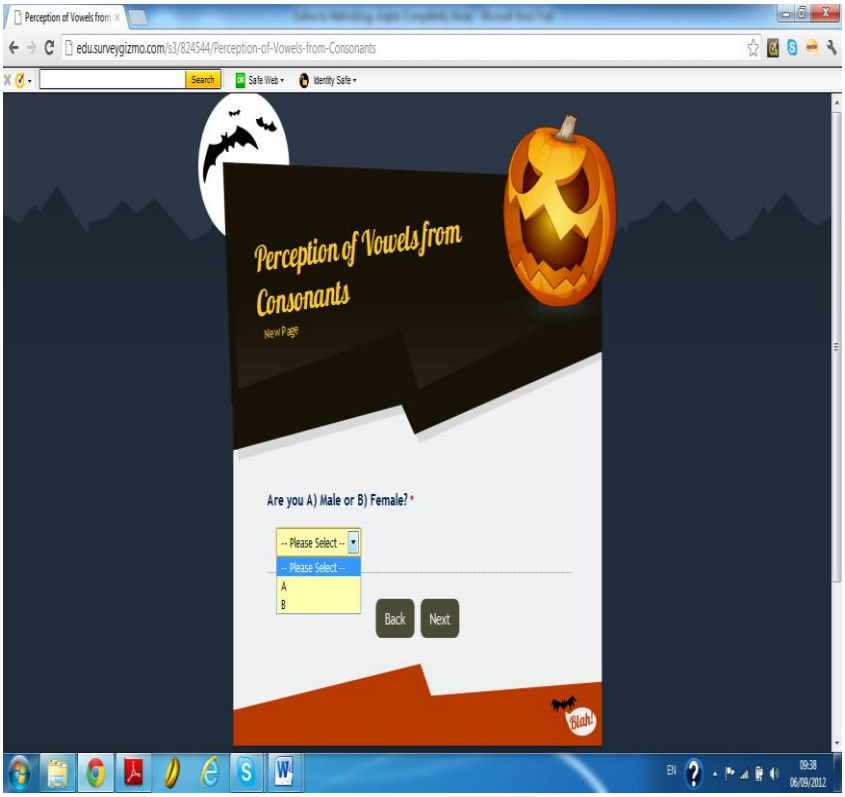
These first six pages were read by the listeners and they only needed to click ‘next’ to progress to the next page (or ‘back’ if something was missed), once they had read each bit of information on each page. Page 7 comprised a consent form for the experiment, which each listener ticked in order to show their consent (and then pressing ‘next’, as before), whereas pages 8-12 constituted certain open questions about the listeners – the questions on each page had to be answered in order to be able to progress (this was done to avoid any blank answers):



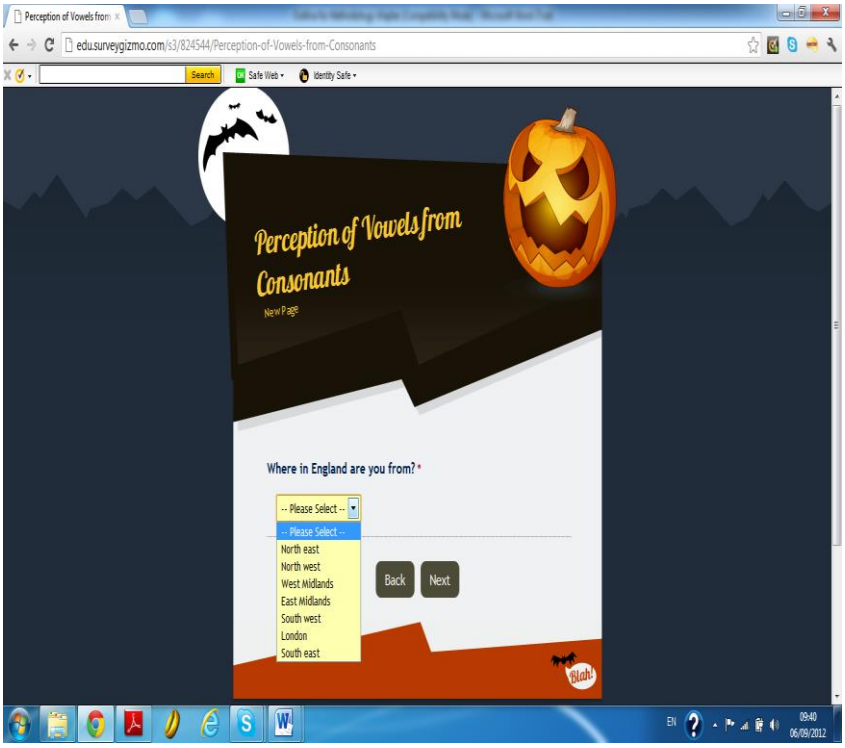
Page 7:



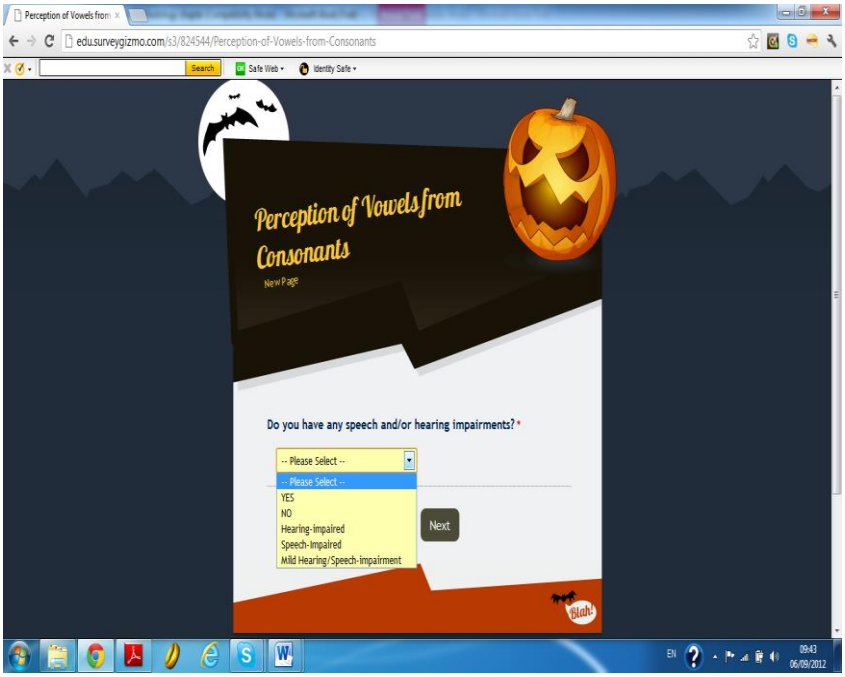
Page eight:



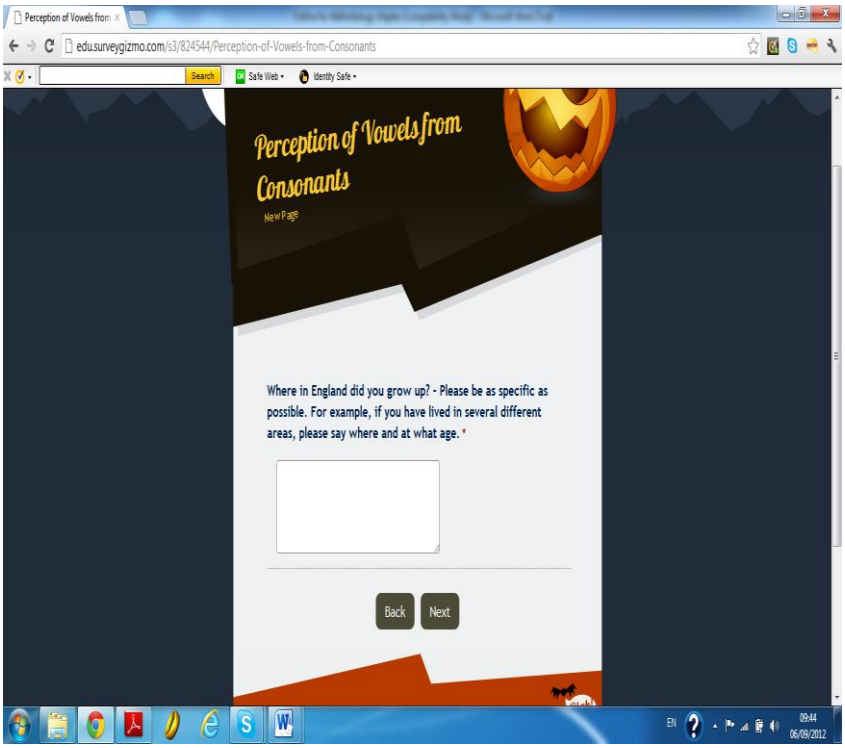
Page nine:



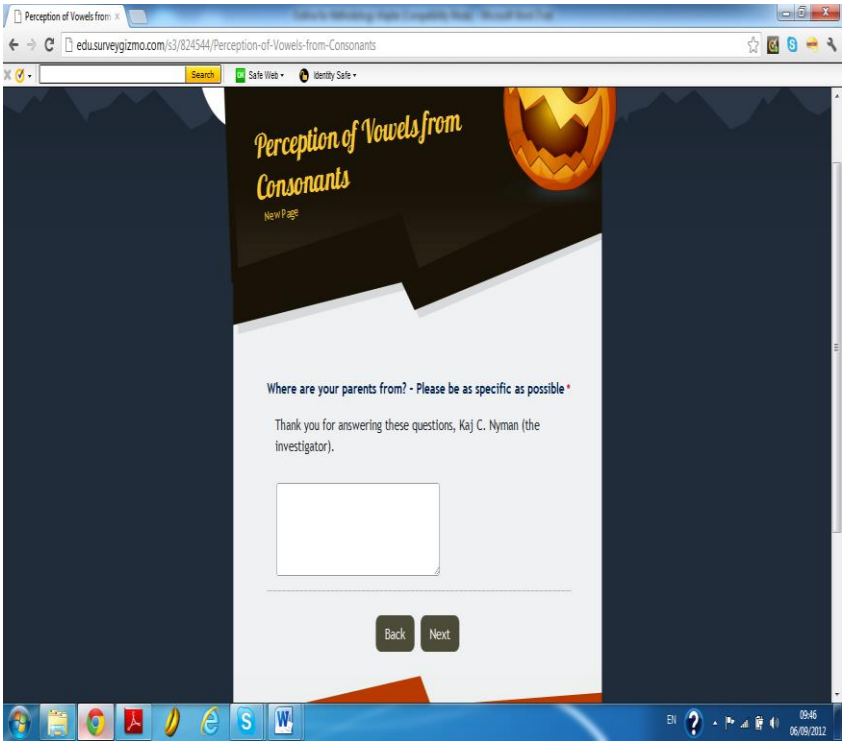
Page 10:



Page 11:

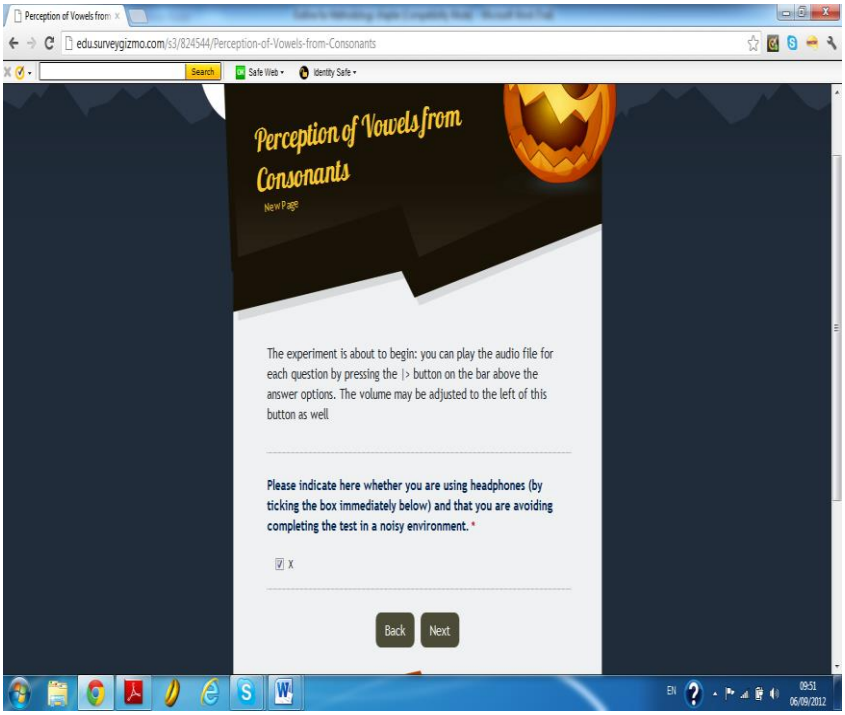


Page 12:



Page 13:

Page 14 constituted the ‘are you ready to begin the experiment?’ page, which the participants ticked when they were ready to begin the experiment. As can be seen by reading the bottom of page 14, the listeners were asked to re-confirm that they are wearing headphones when listening to the stimuli. This question also constituted the ‘point of no return’ for listeners, as they could not return to previous answers having confirmed their readiness togive answers to individual stimuli (also cf. 3.5.3).


Perception of Vowels from Consonants
New Page
The experiment is about to begin: you can play the audio file for each question by pressing the |> button on the bar above the answer options. The volume may be adjusted to the left of this button as well
Please indicate here whether you are using headphones (by ticking the box immediately below) and that you are avoiding completing the test in a noisy environment. *
x
Back
Next

Page 14:

Definitions

The following abbreviations appear in this work:

‘DP’ stands for ‘Declarative phonology’, a constraint based theory of phonology as advocated by John Coleman, Steven Bird, Jim Scobbie and colleagues.

‘FPA’ stands for ‘Firthian Prosodic Analysis’, a constraint based polysystemic theory of phonology as advocated by J.R. Firth.

‘FPD’ stands for fine phonetic detail, which denotes small distinctions between structurally identical utterances that reflect specific combinations of linguistic properties.

‘VISC’ stands for ‘vowel inherent spectral change’, a feature of phonetic exponency specific to vowels. VISC denotes the systematic variation undergone by the vowel formants through time.

‘AVP’ stands for ‘Auditory Vowel Path’ and refers to the way in which auditory nerves on the basilar membrane fire in response to the formant resonances.

‘ASP’ defines ‘auditory space paths’ which refers to the function that needs to be introduced to understand how listeners derive representative vowel values from the inherent variability in VISC. This function has a particular domain reflecting specific vowel productions. The ASP enables speaker-listeners to integrate over the values corresponding to particular vowel paths in auditory vowel space

References

Ali, L., Gallagher, T., Goldstein, J. and Daniloff, R. (1971). Perception of coarticulated nasality. *Journal of the Acoustical Society of America*, 49 (2B), 538–540.

Anderson, S. (1985). *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations.* Chicago and London: Chicago University Press.

Arvaniti, A. and Garding, X. (2007). Dialectal variation in the rising accents of American English. In J. Cole and J. Hualde, (Eds), *Laboratory Phonology 9*. The Hague: Mouton de Gruyter, pp. 547–576.

Atterer, M. and Ladd, D. (2004). On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German. *Journal of Phonetics* 32, 177–197.

Beddor, P. and Krakow, R. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation, *Journal of the Acoustical Society of Amer*ica, 105(6), 2868–2887.

Blumstein, S. and Stevens, K. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of Amer*ica, 67, 648.

Boersma, P., and Wennink, D. (2010). Praat: Doing phonetics by computer (Version 5.1. 31)[Software]. Retrieved 10 Dec, 2014.

Browman, C. and Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219–252.

Caramazza, A. (1997). How many levels of processing are there in lexical access?. *Cognitive neuropsychology*, *14*(1), 177-208.

Carignan, C., Shosted, R., Shih, C. and Rong, P. (2011). Compensatory articulation in American English nasalized vowels. *Journal of Phonetics*, 39(4), 668–682.

Carlson, R., and Hawkins, S. (2007). When is fine phonetic detail a detail? *ICPhS XVI*, Saarbrücken, 6–10 August, 211–214.

Catford, J. (2001). *A Practical Introduction to Phonetics*, 2nd edition. Oxford and New York: Oxford University Press.

Chang, Y., Hsieh, F. and Hsieh, Y. (2011). Phonetic implementation of nasality in Taiwanese (and French): Aerodynamic case studies, *ICPhS XVII*, Hong Kong, 436–439.

Chapman, S. and Routledge, C. (2005). *Key Thinkers in Linguistics and the Philosophy of Language*. Oxford and New York: Oxford University Press.

Chomsky, N. and Halle, M. (1968). The sound pattern of English. New York, Harper & Row.

Clements, G. and Osu, S. (2002). Explosives, implosives and nonexplosives: the linguistic function of air pressure differences in stops. In C. Gussenhoven and N. Warner (Eds.) *Laboratory phonology, 7*, Berlin and New York: Mouton de Gruyter, pp. 299-350.

Clopper, C. and Pisoni, D. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207-238.

Cohn, A. (1990). Phonetic and phonological rules of nasalization. *Working Papers in Phonetics*, Department of Linguistics, UCLA, UC Los Angeles, 1–224.

Coleman, J. (1990). "Synthesis-by-rule" without segments or rewrite-rules. In G. Bailly, C. Benoit and T. R. Sawallis (Eds). *Talking Machines: Theories, Models, and Designs.* Amsterdam: North-Holland. pp. 43–60.

Coleman, J. (1998). *Phonological Representations*. Cambridge: Cambridge University Press.

Collins, B. and Mees, I. (2003). *Practical Phonetics and Phonology, a resource book for students*. Oxford, Routledge.

Cowan, N. and Morse, P. (1986). The use of auditory and phonetic memory in vowel discrimination. *Journal of the Acoustical Society of Amer*ica. 79, 500–507.

Cruttenden, A. (2014). *Gimson's pronunciation of English*. Oxford, Routledge.

Cullinan, W. and Tekieli, M. (1979). Perception of vowel features in temporally-segmented noise portions of stop-consonant CV syllables. *Journal of Speech and Hearing Research*, 22, 122–131.

Daniloff, R and Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research*, 2, 707–721.

Daniloff, R. and Moll, K. (1971). Investigation of the timing of velar movements during speech, *Journal of the Acoustical Society of Amer*ica., 50, 678.

Davis, S. and Summers, W. (1989). Vowel length and closure duration in word-medial VC sequences. *Journal of the Acoustical Society of Amer*ica, 85 (Suppl. 1), S28.

Delattre, P. (1965). Change as a correlate of the vowel-consonant distinction. *Studia Linguistica*, 18, 12–25.

Docherty, G. (1992). *The Timing of Voicing in British English Obstruents*. Berlin & New York: Foris Publications.

Docherty, G. and Foulkes, P. (Eds). (1999). *Urban Voices: Accent studies in the British Isles*. London: Arnold Publishers.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice* (Vol. 373). Springer.

Field, A. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll).*. Thousand Oaks, California, Sage Publications.

Firth, J. (1934c).The word "phoneme", *Le Maître phonétique*. Reprinted in Firth 1957a: 1–2.

Foulkes, P. and Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34(4), 409–438.

Fowler, C. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics, 68*(2), 161-177.

Fowler, C. and Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, 36, 171–195.

Fry, D. (1947). *Acoustic Phonetics*. London, Cambridge, New York: Cambridge University Press.

Goffman, L., Smith, A., Heisler, L. and Ho, M. (2008). The breadth of coarticulatory units in children and adults. *Journal of Speech, Language and Hearing Research*, 51, 1424–1437.

Goldinger, S. and Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics, 31*(3), 305-320.

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic analysis, 2*, 23-68.
Goldstein, B. (2013), Sensation and Perception, ninth edition. California, USA, Wadsworth Cengage Learning.

Grosjean, F. (1980). Spoken word recognition process and the gating paradigm. *Perception & Psychophysics*, 28(4), 267–283.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes*, 11(6), 597–604.

Gussenhoven, C. (2007). A vowel height split explained: compensatory listening and speaker control. In J. Cole and J. Hualde (Eds). *Laboratory Phonology 9*. Berlin and New York: Mouton de Gruyter. pp. 145–172.

Günther, F. (2003). Neural control of speech movements. In N. Schiller and A. Meyer (Eds.) *Phonetics and phonology in language comprehension and production: Differences and similarities*, Berlin: Mouton de Gruyter, pp. 209-239.

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, *106*(2), 1064-1077.

Hardcastle, W. and Hewlett, N. (1999). *Coarticulation: theory, data, and techniques in speech production*. Cambridge and New York: Cambridge University Press.

Harrington, J., Fletcher, J. and Beckman, M. (1999). Lip and jaw coarticulation. In W.J. Hardcastle and N. Hewlett (Eds.) *Coarticulation,*. Cambridge: Cambridge University Press, pp. 144–175

Harris, J. and Lindsey, G. (1995). The elements of phonological representation. In J. Durand and F. Katamba (Eds). *Frontiers of phonology: atoms, structures, derivations*. Harlow, Essex: Longman, pp. 34–79.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding, *Journal of Phonetics*, 31, 373–405.

Hawkins S. (2010a). Phonological features, auditory objects, and illusions. *Journal of Phonetics*, 38, 60–89.

Hawkins, S. (2010b). Phonetic variation as communicative system: perception of the particular and the abstract. In C. Fougeron. M. D'Imperio and B. Kühnert, (Eds). *Laboratory Phonology X*. The Hague: Mouton de Gruyter, pp 479-510.

Hawkins, S. and Nguyen, N. (2001). Perception of coda voicing from properties of the onset and nucleus of *led* and *let*, *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001.

Hawkins, S., and Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics, 32*(2), 199-231.

Hawkins, S. and Slater, A. (1994). Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech. *ICSLP 94 Proceedings of the 1994 International Conference on Spoken Language Processing* 1, 57–60.

Hawkins, S, and Smith, R. (2001). Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics – Rivista di Linguistica*, 13(1), 99–188.

Hawkins, S. and Stevens, K. (1985). Acoustic correlates of the non-nasal-nasal distinction for vowels. *Journal of the Acoustical Society of America.*, 77, 1560–1575.

Heid, S. and Hawkins, S. (2000). An acoustical study of long-domain /r/ and /l/ coarticulation. *Proceedings of the 5th Seminar on Speech Production: Models and Data.* (ISCA). Kloster Seeon: Bavaria, Germany, 77–80.

Henke, W. (1966) Dynamic articulatory model of speech production using computer simulation, PhD dissertation, MIT.

Hillenbrand, J., Getty, L., Clark, M. and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of Amer*ica, 97, 3099

Hillenbrand, J., Clark, M. and Nearey, T. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of Amer*ica, 109, 748.

Howell, P. (1981). Identification of vowels in and out of context, *Journal of the Acoustical Society of Amer*ica, 70(5), 1256–1260.

Janse, E. (2003). Word perception in fast speech: artificially

time-compressed vs. naturally produced fast speech. *Speech Communication*, 42, 155–173.

Johnson, K. (1997), Acoustic and Auditory Phonetics, 1st edition. Oxford, Blackwell.

Johnson, K. (2011), Acoustic and Auditory Phonetics, 3rd edition. Malden, MA. Wiley-Blackwell.

Kelly J. and Local J. (1986). Long-domain resonance patterns in English. *Proceedings of the International Conference on Speech Input/Output*, IEE, 304–308.

Kelly, J. and Local, J. (1989). *Doing Phonology*. Manchester, UK and New York: Manchester University Press.

Kent, R. and Read, C. (2002). Acoustic Analysis of Speech, 2nd edition, USA, Delmar, Cengage Learning.

Klatt, D. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America.*, 59, 1208–1221.

Kohler, K. (2005). Timing and communicative functions of pitch contours. *Phonetica* 62, 88–105.

Krakow, R. (1994). Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate, *Haskins Laboratories Status Report on Speech Research*, SR-117/118, 31–48.

LaRiviere, C., Winitz, H. and Herriman, E. (1975). Vocalic transitions in the perception of voiceless initial stops *Journal of the Acoustical Society of Amer*ica, 5(2), 470–475.

Laver, J. (1994). *Principles of phonetics*. Cambridge: Cambridge University Press.

Liberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English, *Language*, 33(1), 42–49.

Local, J. (2003). Variable domains and variable relevance: interpreting phonetic exponents. *Journal of Phonetics*, *31*(3), 321-339.

Local, J. and Ogden, R. (1997). A model of timing for

nonsegmental phonological structure. In J. Van-Santen, J. Olive, R. Sproat and J. Hirschberg (Eds). *Progress in speech synthesis*, pp. 109–121.

Macmillan, N., Kingston, J., Thorburn, R., Walsh Dickey, L. and Bartels, C. (1999). Integrality of nasalization and F1. II. Basic sensitivity and phonetic labeling measure distinct sensory and decision-rule interactions. *Journal of the Acoustical Society of Amer*ica, 106(5), 2913–2932.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1-71.

Matisoff, J. (1975). Rhinoglottophilia: the mysterious connection between nasality and glottality. In C. Ferguson, L. Hyman and J. Ohala (Eds). *Nasálfest: Papers from a symposium on nasals and nasalization.* Stanford: Language Universals Project, Stanford University, 265–287.

Modarresi, G., Sussman, H., Lindblom, B. and Burlingame, E. (2004). An acoustic analysis of the bidirectionality of coarticulation in VCV utterances. *Journal of phonetics*, *32*(3), 291-312.

Moore, B. (2008). *The Psychology of Hearing* (5th edition). Bingley, UK: Emerald Publishing.

Nearey, T. and Assmann, P. (1986). Modelling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of Amer*ica, 80(5), 1297–1308.

Nittrouer, S. (2007). Dynamic spectral structure specifies vowels for children and adults. *Journal of the Acoustical Society of Amer*ica, 122(4), 2328–2339.

Nyman, K. (2010). Cues to vowels in English plosives. *BAAP 2010*, Proceedings of the British Association of Academic Phoneticians conference, p. 62.

Ogden, R, (1992). Parametric interpretation in York talk. *York Papers in Linguistics*, 16, 81–89.

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics*, 38, 1752–1775.

Ohala, J. (1975). Phonetic explanations for nasal sound patterns. In C. Ferguson, L. Hyman and J. Ohala (Eds). *Nasálfest: papers from a Symposium on Nasals and*

Nasalization. Stanford: Language Universals Project, Stanford University, 289–316.

Ostreicher, H. and Sharf, D. (1976). Effects of coarticulation on the identification of deleted consonant and vowel sounds. *Journal of Phonetics*, 4, 285–301.

Öhman, S. (1966). Coarticulation in VCV utterances: spectrographic measurements, *Journal of the Acoustical Society of Amer*ica, 39, 151–168.

Parnell, M. and Amerman, J. (1978). Maturational influences on perception of coarticulatory effects. *Journal of Speech and Hearing Research*, 21(4), 682.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation, Massachusetts Institute of Technology).

Plug, L. (2005). In Local, J and Wells, B (Eds) Seventy years of Firthian phonology: prospect and retrospect. *York Papers in Linguistics*, 2(4), pp. 15–48.

Port, R. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, *32*(2), 141-152.

Ranbom, L. and Connine, C. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, *57*(2), 273-298.

Recasens, D. (2002). An EMA study of VCV coarticulatory direction. *Journal of the Acoustical Society of Amer*ica, 111, 2828–2841.

Repp, B. and Crowder, R. (1990). Stimulus order effects in vowel discrimination. *Journal of the Acoustical Society of Amer*ica, 88(5), 2080–2090.

Rosner, B. and Pickering, J. (1994). *Vowel perception and production*. Oxford: Oxford University Press.

Schourup, L. (1973). A cross-language study of vowel Nasalization. *Ohio State Working Papers in Linguistics* 15, 190–221.

Simpson, A. (2005). From a grammatical angle: congruence in Eileen Whitley's phonology of English. *York Papers in Linguistics*, 2(4), 49–90.

Shockey, L. (2003). *Sound patterns of spoken English.* Oxford, Blackwell.

Shulman, R., Rothman, D., Behar, K, and Hyder, F. (2004). Energetic basis of brain activity: implications for neuroimaging. *Trends in neurosciences*, *27*(8), 489-495.

Smith, R., Baker, R. and Hawkins, S. (2012). Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *Journal of Phonetics*, 40(5), 689–705.

Sprigg, R. (2005). The short-quantity piece in English lexical items, and its vowel systems. *York Papers in Linguistics,* 2(4), 157–188.

Stevens, K. (1998). *Acoustic Phonetics*. Cambridge, MA. and London, England: MIT.

Stevens, K. and Blumstein, S. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358-1368.

Tekieli, M. and Cullinan, W. (1979). The perception of temporally segmented vowels and consonant-vowel syllables. *Journal of Speech and Hearing Research,* 22, 103–121.

Temple, R. (2009). (t,d): The variable status of a variable rule. Oxford University Working Papers in Linguistics, Philology and Phonetics. In O. Parker-Jones and E. Payne (Eds). *Papers in Phonetics and Computational Linguistics*, 12, 145–170.

Tyler, L. and Wessels, J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception & Psychophysics*. 38 (3), 217–222.

Waldstein, R. and Baum, S. (1994). Perception of coarticulatory cues in the speech of children with profound hearing loss and children with normal hearing. *Journal of Speech, Language and Hearing Research,* 37, 952–959.

Watt, D., and Allen, W. (2003). Tyneside English. *Journal of the International Phonetic Association*, *33*(02), 267-271.

Wells, J. (1982). *Accents of English: the British Isles.* (Vol. 2), Cambridge: Cambridge University Press.

West, P. (1999b). Perception of distributed coarticulatory properties of English /l/ and /r/. *Journal of Phonetics* 27(4), 405–426.

Whalen, D. (1990). Coarticulation is largely planned. *Haskins Laboratories Status Report on Speech Research,* 102, 149–176.

Winitz, H., Scheib, M. and Reeds, J. (1972). Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of Amer*ica, 51(4B), 1309–1317.

Wright, J. (1975). Effects of vowel nasalization on the perception of vowel height. In C. Ferguson, L. Hyman and J. Ohala (Eds). *Nasálfest: Papers from a symposium on nasals and nasalization,* Stanford University, Language Universals Project, pp. 373–388.

Wright, J. (1986). Nasalized vowels in the perceptual vowel space. In J. Ohala and J. Jaeger (Eds). *Experimental phonology,* Orlando: Academic Press, pp. 45–67.

Xu, Y. (2009). Timing and coordination in tone and intonation - an articulatory-functional perspective. *Lingua*, 119, 906–927.