The Role of Prosodic Cues in Speech Intelligibility

Christine Binns

CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYDD

Thesis submitted to Cardiff University for the degree of
Doctor of Philosophy

February, 2007

CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my supervisor John Culling for his invaluable guidance and patience throughout my PhD, without which I would surely have been lost in some kind of acoustical maze. Thanks also for the continuous encouragement from my friends and family, the many coffee breaks and ever willing experimental subjects that they brought with them.

Thanks to my mum, dad and sister for their constant support; to Lise and Kate for listening where others would have stopped; to Lucy for being a friend; to Sam for occasionally letting me win; to James for giving me my bounce back; to Beth, Suze and Helen for those much needed time-outs; and last, but by no means least, to the rowing crowd for some fantastic days on the river, amazing nights out and of course, scrunting. Without you all I may have finished this a tad sooner (!), but it wouldn't have been half as much fun.

SUMMARY

Listeners are often required to attend to speech in background noise. Coherent prosodic structure has been found to facilitate speech processing (Cutler, Dahan & Donselaar, 1997). The aim of this thesis is to investigate to what extent these prosodic cues, in particular fundamental frequency (F0), aid speech intelligibility in noise. Experiments measured Speech Reception Thresholds for sentences with different manipulations of their F0 contour. These manipulations involved either a scaled reduction in F0 variation, or the complete inversion of the F0 contour. Experiments reported in Chapter 2 investigated the impact of these F0 manipulations against speech-shaped noise and single-talker interferers. Inverting the F0 contour was found to significantly degrade target speech intelligibility for both types of interferer, although a larger effect was observed with the single-talker interferer. No effect of altering the F0 contour of the interferer was found. Low-pass filtering the F0 contour (Chapter 3) showed that the most important frequencies lay at or below the syllable rate of speech, highlighting the importance of syllabic and suprasegmental fluctuations within the F0 contour. Experiments in Chapter 4 compared synthesised and natural targets. Synthesised speech was found to be considerably less intelligible than natural speech. No consistent effect of F0 inversion was noted for synthesised F0 contours. However neither natural F0 nor duration contours improved the intelligibility of the synthesised speech. Similar experiments using non-native English speakers (Chapter 5) showed a greater detriment to the perceived intelligibility of the speech with F0 manipulations than for native speakers. Results are explained in terms of F0 contours highlighting important content words. Intrinsic vowel pitch is also argued to contribute. Further study is required to determine why speech interferers caused listeners to rely more heavily on F0 cues, and to investigate the influence of other prosodic cues on speech intelligibility.

CHAPTER 1:

INTRODUCTION

Current Automatic Speech Recognition (ASR) systems perform well when used in quiet environments, but have difficulty in accurately recognizing speech in noise. For example, in one study word error rates (WERs) for continuous speech played in a background of babble noise at 0 dB Signal-to-Noise Ratio (SNR) were 16.5%, and rose to 62.4% at an SNR of -5 dB (Hazen, 2006). Contrary to human speech recognition (Festen & Plomp, 1990), ASR systems recognise speech in steady-state noise better than in competing speech, implying that humans use cues unavailable to the ASR system to enhance their performance. The question thus arises as to what these cues are and whether they can be used to aid ASR systems in noise.

This thesis is primarily concerned with the influence of F0 contours on the intelligibility of speech, particularly in background noise. Therefore, the introduction will discuss what is currently known about the role of prosodic cues in speech communication, with an emphasis on the F0 contour. This will be followed by a review of known factors affecting speech intelligibility. Finally it will examine current evidence for the importance of F0 contours for speech intelligibility.

1.1 Prosody

1.1.1 What is prosody?

Prosody is an important suprasegmental aspect of speech which can alter the perceived meaning of an utterance without changing its phonetic content. The three physical parameters most commonly referred to in discussion of prosody are intensity, duration and fundamental frequency. Intensity refers to changes in sound energy throughout the speech. Duration is the variation in segment length across utterances. Fundamental frequency (F0) is the measure of the rate of vibration of the vocal folds during voiced speech; its perceptual correlate, pitch, is a term sometimes used interchangeably in the literature with 'F0', although they are not synonymous in meaning.

1.1.1.1 Intensity

Intensity measures the amount of energy present in a sound or sequence of sounds. Prosodic variation in intensity is caused by variation in air pressure from the lungs. Intensity can be used to emphasise particular words or syllables within an utterance, that is, more intense speech will be more prominent. Intensity can indicate emotion. For instance, when angry, speakers may shout, thus increasing the intensity of the utterance. Intensity is also affected by phonemic factors. Open vowels, such as /i/, tend to be acoustically of greater intensity than closed vowels, such as /u/ (Cruttenden, 1986).

1.1.1.2 Duration

Duration, as mentioned above, is the measure of the length of speech segments. There are a number of factors which induce variation in duration of speech segments, which are discussed below.

As with intensity, duration can vary phonemically. Some vowels are intrinsically longer than others. Consider the words 'bit' and 'beat'. The [I] in 'bit' is shorter than the [i] in ‘beat’ despite being placed within the same context. Emphasis is achieved with intensity by increasing a syllable’s loudness. In the case of duration, however, emphasis is accomplished by lengthening the more prominent syllable. Indeed, Klatt (1976) noted that the average duration for a stressed vowel in connected speech is around 130 milliseconds, whereas an unstressed vowel is around 70 milliseconds long. Overall, speech duration obviously varies with the rate of speech. However, some speech sounds are more susceptible to timing changes than others. Steady sounds such as fricatives and vowels tend to vary more in duration than short-lived sounds such as the bursts of stop consonants. The position of phones and syllables also influences their respective duration. For example, syllable duration tends to increase at the end of a phrase or before a pause in speech. A word spoken in isolation has approximately the same duration as a word produced at the end of a phrase, but lasts twice as long as a sentence-initial word. Vowels before voiced consonants are generally longer than the same vowels produced before voiceless consonants, such as in the words ‘feet’ and ‘feed’, where the second production of [i] is longer. Also, in stressed-timed languages, such as English, an unstressed syllable tends to be shortened if it is between two stressed syllables. Finally, consonants in English tend to be longest in word-initial positions, shorter in word-final positions, and shorter still in word-medial positions.

1.1.1.3 Fundamental frequency (F0)

F0, as a prosodic component, can contribute to the meaning of an utterance both linguistically and non-linguistically. Non-linguistically, F0 can portray the speaker’s sex and age, as well as information such as the speaker’s emotional state. Linguistically, F0 can be described as either lexical or post-lexical. Lexical pitch refers to F0 that is included in the lexicon such that two phonetically-identical words with different pitch accents have different meanings. For instance, in Mandarin the word 'tʰaŋ' can mean 'soup' or 'sugar' depending on whether there is a level accent or a rising accent on the 'aŋ'. Lexical pitch is mostly linked with tonal languages, such as Mandarin, and is much less common in non-tonal languages, such as English. However, it does occur in some English noun/verb distinctions such as PERmit/perMIT. Post-lexical F0, often referred to as intonation, is where the meaning of the F0 contour is processed at the sentence level. For example in the case of a question/statement distinction, a question tends to be indicated with a rise in the F0 contour towards the end of the utterance, in contrast to a falling F0 contour in a statement. The following studies will use non-tonal languages; hence will be mostly concerned with post-lexical pitch.

There are some universal properties of F0, such as all languages use F0 to mark boundaries of syntactic units (Ladefoged, 1993). Utterances tend to fall in F0, a trend which is called 'declination' and is caused by falling subglottal pressure throughout the utterance (Lieberman, 1967). Incomplete utterances tend to have rising intonation, where the speaker is trying to indicate that there is information to follow. In most languages, raised pitch is used to contrast with preceding lower pitch to indicate a question, as opposed to a statement (Hirst & di Cristo, 1998). Vowels tend to have a higher fundamental frequency when preceded by consonants. The F0 peak tends to be at the beginning of a vowel following a voiceless consonant, but in the middle of the vowel after a voiced consonant. This peak is considerably higher for vowels after voiceless consonants than for vowels after voiced consonants. Different types of vowels have intrinsically higher or lower fundamental frequencies, an effect which has been found to hold true across languages (Whalen & Levitt, 1995). High vowels such as /i/ or /u/ tend to have a higher fundamental frequency than low vowels, such as /ɑ/, when used in the same phonetic context. However, the intrinsic F0 difference between vowels is relatively small. For instance, the difference between /i/ and /æ/ is approximately 20 Hz. Diehl (1991) claims that intrinsic F0 is used to enhance the speech signal, such that the speaker creates these intrinsic differences between the vowels to make the speech more salient for the listener. However, physiological explanations detailing the interaction between the vocal fold tension and subglottal pressure have also been proposed to account for this effect (Steele, 1986). Intrinsic F0 and its impact on speech perception will be discussed in more depth later in this chapter.

1.1.2. Uses of Prosody

1.1.2.1. Syntactic Disambiguation

As mentioned above, prosody can aid in syntactic disambiguation (Price, Ostendorf, Shattuck-Hufnagel & Fong, 1991). Price et al. (1991) presented listeners with pairs of phonetically similar sentences representing seven different types of structural ambiguity. Details of these are given below; example sentences are given in italics for two of the structural types.

1) parenthetical clauses versus non-parenthetical clauses, e.g.

    A) Don’t think you can hide your plans from them. *They know, you realise, your goals.*

    B) They think very highly of you and the way you get things done. *They know you realise your goals.*

2) appositions versus attached noun phrases, e.g.

A) Most of the women had forgotten the strange event by the next week. *Only one remembered, the lady in red.*

B) Most of the people forgot about the strange visitor. *Only one remembered the lady in red.*

3) main clauses linked by coordinating conjunctions versus a main clause with subordinate clause

4) tag questions versus attached noun phrases

5) far versus near attachment of final phrase

6) left versus right attachment of middle phrase

7) particles versus prepositions

Listeners were asked to indicate which structural context they felt applied to each utterance on an answer sheet which detailed both structural possibilities for the sentence. Overall, listeners were able to disambiguate the syntactic structure of the sentences using prosodic information. For example, with structural ambiguity (1) detailed above, sentences with major prosodic breaks surrounding the parenthetical 'you realise' were identified as version A 75% of the time, and sentences without the major prosodic breaks were identified as B 80% of the time. Prosodic breaks are units of speech defined by prosodic cues, such as the gradual decline in F0 and lengthening of vowels over the duration of the unit until the F0 and speed are reset to begin the next unit. Larger prosodic breaks tended to be associated with larger syntactic breaks; boundaries which divide up the grammatical structure of the utterance. Syntactic boundaries for clauses containing a sentence tend to coincide with major prosodic constituent boundaries such as syllable-final lengthening or pause. Price et al. (1991) did not, however, investigate which aspects of prosody caused this effect.

Streeter (1978) on the other hand, investigated the role of amplitude, duration and F0 in the perception of a form of syntactic break, called phrase boundaries. Expressions of the form 'A plus E times O' were recorded with different prosodic boundaries ('(A plus E) times O' and 'A plus (E times O)') as well as without parsing (A plus E times O). Amplitude, F0 and duration were manipulated for these utterances such that there were two levels for each cue, "neutral" and a value taken from one of the bracketed expressions. For example, using F0 as a cue, the 'neutral' or '-F0' condition consisted of a contour monotonised at 100 Hz and the '+F0' condition was the contour taken from one of the bracketed recorded expressions. All three variables were fully cross-manipulated giving a total of 8 manipulations for each expression. Listeners were asked to indicate where the prosodic boundaries for these utterances lay. Results showed that timing and F0 were reliable cues for marking the phrase boundary locations, with timing cues more strongly indicating phrase boundary position than F0. Amplitude was the least effective disambiguation cue. In a summary of work on durational cues and their applications, Klatt (1976) stated that duration tends to serve as a primary cue (as supported by the above data) for phrase boundary location by lengthening the phrase-final syllables in comparison to non-final ones.

1.1.2.2 Detecting Word Boundaries

In stress-timed languages such as English, speech rhythm is expressed with the alternation of strong and weak syllables. Strong syllables contain a full vowel and bear either primary or secondary stress. Weak syllables tend to be unstressed and consist of short, central vowels, such as schwa. Cutler and Carter (1987) found that over 90% of all English content words started with a strong syllable. On the basis of these results, Cutler and Butterfield (1992) investigated whether listeners use cues concerning syllable strength to aid them in speech segmentation. Listeners were required to listen to a set of unpredictable six syllable sequences with an alternating stress pattern of strong and weak syllables and write down what they heard, inserting a dash for any syllables they were unable to report. Sequences either consisted of only weak word-initial syllables (e.g. 'conduct ascents uphill'), only strong word-initial syllables (e.g. 'duty senseless drilling', or a mixture of the two (e.g. 'soon police were waiting'). Results showed that word boundary insertions were more common before strong syllables than weak syllables. Conversely, word boundary deletions were more common before weak syllables than strong syllables. Vroomen, van Zon and de Gelder (1996) found a similar pattern of results when they replicated the Cutler and Butterfield (1982) experiment on Dutch speakers listening to Dutch speech. Therefore it seems that speakers of stress-timed languages such as English and Dutch use information regarding syllable strength to segment speech.

Using nonsense words, Nakatani and Schaffer (1978) have investigated whether amplitude, duration or F0 are adequate prosodic cues for word boundary perception. In a preliminary study, participants were asked to state whether the reiterant phrase 'mamama' imitated a disyllabic adjective and monosyllabic noun (mama/ma) or monosyllabic adjective and disyllabic noun (ma/mama). All monosyllabic words had primary stress. Therefore, any phrases with secondary or null stress on the first or last syllable would indicate that these syllables were part of a disyllabic word. Some stimuli, however, had ambiguous stress patterns. For example, the phrases "tasty food" and "bold design" have different parsings, but they have the same stress pattern with primary stress on both the first and last syllable. Stimuli which mimicked phrases whose stress pattern uniquely determined the lexical breaks were more accurately parsed than those whose stress pattern was ambiguous. However, listeners were still able to parse the reiterant phrases with ambiguous stress patterns more successfully than one would expect simply by chance alone. This implies that there must be prosodic cues other than the stress pattern that can be used for word perception.

In light of these results, contributions of the duration, pitch and amplitude to word perception were studied by asking the listener to parse hybrid speech phrases. These phrases contained some prosodic information from one phrase and some from another. Hybrid speech phrases were produced by dynamically time-aligning two ambiguous phrases with the same stress pattern but different parsings, such that all prosodic cues could be transferred from one phrase to the other. For example, a hybrid speech sample could contain the pitch and amplitude cues from phrase 1, but the duration cues from phrase 2. All cues were parametrically-manipulated. Listeners were again asked to state whether the phrase should be grouped as 'mama/ma' or 'ma/mama'. The way in which the listener chose to parse the phrase therefore determined the more pertinent cues for word perception. Studying the above example, if listeners chose to parse the utterance in the same way as phrase 2, it could be stated that duration cues were more salient than pitch and amplitude cues for word boundary perception. Indeed, results showed that the only prosodic feature to influence the participants' parsing of the hybrid phrases was duration, contradicting research from Klatt (1976). Klatt stated that duration cues to word boundary location in English are too small and too variable to be of much perceptual use. Neither pitch nor amplitude affected the parsing. Therefore, it seems that in the absence of lexical content to aid the listener, duration contours can indicate lexical boundaries. The F0 contours of both reiterant phrases, however, were not significantly different. Given that the stress patterns were the same for both ambiguous phrases it seems that F0 is directly correlated with lexical stress when no syntactic information is portrayed.

Findings from research on the perception of lexical stress in polysyllabic words presented in isolation (Fry, 1958; Morton & Jassem, 1965) demonstrate that although intensity, duration and F0 all contribute to stress perception, intensity variation tends to be the weakest cue. Unlike the results from Nakatani and Schaffer's experiment which show duration to be the stronger prosodic cue, F0 tends to be the strongest cue in indicating lexical stress (Fry, 1958; Morton & Jassem, 1965).

1.1.2.3 Lexical Stress

A study by Fry (1958) investigated the importance of each prosodic cue for identifying stress in synthesised polysyllabic noun/verb word pairs such as SUBject/subJECT and CONtract/conTRACT. These were produced by pattern playback equipment at the Haskins laboratory by asking listeners to indicate the position of the stressed syllable in conditions where each of the three cues, F0, intensity and duration, had been manipulated. Results showed that changes to the vowel duration ratio adjusted the listener's perception of where the stressed syllable lay. If the first vowel was longer than the second, listeners tended to judge the word as a noun, with the initial syllable stressed. If the second was longer, it was judged as a verb with the second syllable marked as stressed. Intensity showed a similar effect, such that the more intense the syllable, the more stressed it was considered to be. However, intensity did not produce as strong an effect as duration. Manipulating the fundamental frequency showed that a higher frequency syllable was judged to be more stressed than a syllable at a low frequency. Unlike duration and intensity, the magnitude of the frequency change did not modulate the size of the effect. It appears that the occurrence of a frequency change is more important than the size of the change itself. A second experiment studied whether F0 cues outweigh other factors involved in stress judgements. This experiment again used the word-pair 'subject', but this time manipulated pitch change on individual syllables. Two types of F0 manipulation were employed. In the first scenario, F0 changed continuously throughout the vowel and in the second, the F0 change only took place over half the vowel duration. Duration was manipulated along with the F0. Results found that although duration cues, in the absence of frequency cues, did affect the stress judgement, sentence F0 contour was an over-riding factor in predicting sentence stress.

Morton and Jassem (1965) also indicated the importance of F0 in portraying lexical stress in experiments investigating the perception of stress in the nonsense words 'soso', 'sisi' and 'sasa'. These words were synthesised using PAT (Parametric Artificial Talker). Stimuli were chosen such that they would be phonetically simple to avoid distraction from lexical content. However, they would contain sounds and sound sequences which would be permissible in different languages. Amplitude, duration and F0 contours were manipulated within the synthesiser to produce stimuli either with one manipulated parameter on one syllable, two manipulated parameters on one syllable or one manipulated parameter on each of the syllables. Amplitude of the stimuli was manipulated by -3, -6 or -9 dB relative to the level of the standard stimuli produced by the synthesiser based on parameters extracted from actual utterances spoken by a phonetician. Duration was adjusted by 20 or 40 per cent on either side of the vowel length in the standard stimuli (210 milliseconds for the first vowel, 260 milliseconds for the second vowel in 'soso' and 280 milliseconds for 'sisi' and 'sasa'). With F0 manipulations, the standard F0 of 120 Hz was replaced by steady F0s of 76, 96, 110, 130, 151 and 190 Hz along with a slope up or down, to or from 120 Hz to the steady F0 that the standard F0 had been replaced with. Asking listeners to indicate upon which syllable they noticed syllable stress, they found that F0 changes were most effective in indicating where the stress lay. This confirmed Fry's belief that F0 cues potentially outweigh duration cues. All changes in F0 from the standard frequency resulted in the syllable with the altered F0 being marked as stressed. Raised F0 was found to be more effective than a lowered F0 in indicating a stress change, and in accordance with Fry's results a 58% change in F0 was no more effective than a 25% change. Thus, once again increasing the size of the F0 change did not increase the effect. Intensity and duration cues were much less consistent in marking lexical stress.

More recently, Kochanski, Grabe, Coleman and Rosner (2005) have shown using natural speech that amplitude and duration play a larger role in indicating prominent syllables than F0. Speech from the IViE (Grabe, Post & Nolan, 2001) corpus was used and in particular data from three styles of speech: 'sentences', 'read story' and 'retold story'. Syllables from this speech were marked for prominence by the listeners. Measures of loudness, aperiodicity, spectral slope, F0 and duration were then taken. This information was used to train a classifier to reproduce the listener's decision of whether a syllable was prominent or not. Examination of the classifier's performance on a test set showed how reliably listeners used each of the prosodic properties to mark prominent syllables. Classifiers based on loudness were approximately 50 per cent more accurate at prominence judgements than classifiers based on duration, and over twice as accurate as classifiers based on F0. Kochanski et al. (2005) proposed that in experiments using synthesised speech much larger F0 excursions were used than were found in the natural speech. They argued that this could account for F0 proving a more reliable cue to stress in previous studies. However, studies such as Fry (1958) used isolated words where the particular stress under focus was lexical stress. Kochanski et al. (2005) were investigating sentential prominence. Therefore, it could be reasoned that the difference in results is due to F0 being a more important perceptual cue for lexical stress than for sentential prominence.

1.1.2.4 Non-native use of prosody

Non-native listeners have been shown to segment speech using the rhythmic patterns learnt from their native speech (Otake, Hatano, Cutler & Mehler, 1993; Cutler, Murty & Otake, 2003). Otake et al. (1993) investigated English, French and Japanese listeners' segmentation of Japanese words. Japanese is a mora-timed language, whereas English is stress-timed and French syllable-timed. Listeners were presented with Japanese words of the form CVCVCV (e.g. tanishi) or CVNCV (e.g. tanshi), and asked to listen for a word beginning with a specific target sound of the form CV (e.g. ta-) or CVN (e.g. tan-). They found that non-native and native listeners segmented the words differently to each other. When speech is segmented using syllables, CV targets should be responded to more accurately in CVCVCV words and CVN targets more accurately in CVNCV words because these are the initial syllables of these words. Using mora, CV targets would be responded to with equal accuracy in all words because this is the initial mora of both words. Results showed that Japanese listeners used mora to segment the words. French speakers segmented the Japanese words using syllable structure, and English speakers did not seem to use either syllable or mora structure. Thus it seems that non-native listeners apply their native listening strategies to non-native languages. This fact has been confirmed by Murty et al. (2007) in experiments investigating target perception for Japanese listeners. When presented with words from Telegu, a language similar to Japanese in its syllable structure, despite being unfamiliar with the language, the Japanese listeners correctly segmented the words using their native mora structure. This is consistent with the idea that listeners extend their native language rules to non-native languages.

Non-native speakers have been shown to be unable to interpret and use native prosodic cues. Sanders, Neville and Woldorff (2002) found that despite non-native speakers being able to learn new segmentation cues, they were unable to exploit the syntactic information from these cues. Four groups of participants: Japanese late-learners of English, Spanish late-learners of English, near-monolingual Japanese speakers and near-monolingual Spanish speakers were compared to a monolingual group of English speakers on their ability to perceive a target word in English sentences. The sentences varied in the amount of lexical, syntactic and stress-information present. Stimuli were classed as either 'semantic', which referred to a normal English sentence, 'syntactic' in which all content words were replaced with non-words, and 'acoustic', in which every word was a non-word, thus retaining normal English prosody but with little meaning or syntactic information. Example sentences for each category are given below.

Semantic: The child stopped crying when a balloon was given to her.

Syntactic: The ferp trepped plawing when a barreal was kaffen to her.

Acoustic: Sa ferp trepp plawel ron i barreal hof kaffem gi wem.

Results showed that listeners were able to exploit lexical and semantic cues in the sentences to aid their retrieval of the target word. No group of non-native speakers, however, used syntactic information to the same extent as the native speakers. They also found that late non-native learners could rely on stress pattern information for segmentation cues when lexical and semantic information was absent, implying that the ability to learn stress pattern information is still present beyond the age of 12.

Specifically examining F0, Akker and Cutler (2003) applied the research of Cutler and Fodor (1979) to the use of F0 cues (see 1.3) by non-native speakers. The two language groups they chose were English and Dutch. Dutch is very similar to English in its prosodic structure, with the same rules for accent assignment and similar lexical stress assignment and intonation contour structure. The first set of experiments investigated listeners' use of F0 cues in their native language. Listeners heard a question followed by an answer. They were asked to attend closely to the sentence materials as well as identify the syllable-initial phoneme in the answer. The stimuli were constructed such that there were two possible target phonemes, one in the early part of the sentence and one in the latter part. The targets were always placed within stressed syllables and were word-initial. There were also two variants of prosodic context. The intonation contour in the first variant predicted an accented target; the second predicted a deaccented target. Two types of question were presented to the listener: one whose answer was in the first part of the sentence, and one whose answer was in the second. This meant that the question answer may or may not be the semantic focus of the sentence, depending on the syllable accent and the prosodic context. For instance, consider the following stimuli for the target /d/. In response to question 1, the target is both the semantic focus and accented in answer A, whereas although it retains the semantic focus of the sentence in B, it loses the accent. Question 2 makes the word 'Cuban' the focus of the sentence, thus making the /d/ target in A and B unfocussed, although it retains the accent in A.

1) Which bones were found by the archaeologist?

A) The bones of the DINOSAUR were found by the Cuban archaeologist.

B) The bones of the dinosaur were found by the CUBAN archaeologist.

2) Which archaeologist found the bones?

A) The bones of the DINOSAUR were found by the Cuban archaeologist.

B) The bones of the dinosaur were found by the CUBAN archaeologist.

A written test was administered after the experiment, which required the participants to decide in which of the sentences that had been presented to them the target had been early in the sentence and in which it had been late. A similar pattern of results was seen for both the English and Dutch native speakers in their own native language. Accented targets were responded to more quickly than deaccented, as were targets that were in a focussed position rather than unfocussed. However, results from non-native listening showed that although listeners can exploit focus and accent cues in their non-native language, they cannot equal native efficiency in mapping the accent to meaning. Thus when a different set of native Dutch speakers participated in the same experiment, using their non-native language, English, a significant effect of accent and focus was shown. In both native listening experiments, however, the effect of predicted accent was less than half when under focus than in an unfocussed position. In non-native listening the effect of accent under focus was not significantly reduced in comparison to those that were unfocussed. Therefore, non-native listeners did not exploit the cues as effectively as native listeners.

The inability of non-native listeners to match native listeners in the understanding of non-native prosody was also shown using two prosodically different languages by Atoye (2005) who investigated the perception and interpretation of English sentence intonation by Nigerian speakers of English. Subjects were native speakers of the tone language, Yoruba, who were all studying English at degree level. Five English sentence pairs were presented to each listener. The sentences differed only in the intonation placed on them. Subjects were asked to indicate whether there was a difference between the sentences, and then to paraphrase the distinct meaning of each sentence. For example, the sentence 'she dressed and fed the baby' could be interpreted as 'the baby was both dressed and fed' or as 'she dressed herself and then fed the baby' depending on the intonation of the utterance. In a forced choice task, subjects managed to correctly perceive the intonation differences 87.5% of the time. However, when the subjects were asked to paraphrase the sentences unambiguously, they only managed 25.7% of the time to correctly interpret the sentences. These results indicate that although non-native listeners can perceive differences in intonation, they are unable to interpret them accurately.

1.1.2.5 The Effect of Prosody on Speech Intelligibility

Prosody influences our perception of speech greatly. In general, intensity variations tend to be less influential than durational patterns and F0 contours (Fry, 1958; Morton & Jassem, 1965; Steeter, 1978; Holmes & Holmes, 2001). However, these studies do not show the effect that prosody has on overall speech intelligibility. One study by Cutler and Clifton (1984) claims that although deliberately mis-stressed words are responded to more slowly than correctly stressed ones, stress itself does not facilitate word recognition for English listeners. From this finding, one could infer that prosody is not overly important for speech intelligibility in English. Connine, Clifton and Cutler (1987) however, showed that with perceptually ambiguous information, lexical stress can be used to resolve ambiguity. Thus it seems that the importance of prosody increases with increasing ambiguity of the stimulus. This thesis is primarily concerned with one aspect of prosody, F0 and its impact on speech intelligibility. The following section will detail further aspects of F0.

1.1.3 Fundamental Frequency Contours

Evidence presented above indicates that F0 contours influence speech perception. It is therefore important to understand the structure of an F0 contour. Thus, this section will examine evidence from two approaches to understanding F0 contour structure: inferences from and predictions of natural F0 production. Inferences from production will cover the general trends in F0 contours and how they are interpreted. Predictive models for F0 contours, as used in speech synthesis, will then be discussed as a method of testing ideas concerning these F0 trends.

1.1.3.1 F0 Production

F0 is a strong cue for the perception of stress (Fry, 1958). Stress is defined by Hayes (1995) as the linguistic manifestation of rhythmic structure, which in stress languages such as English is a property of every utterance. In such languages, the syllable is the stress bearing unit (Hayes, 1995). However, no single F0 feature can be associated with stress. There are many different pitch accents that can be used to indicate stress. Autosegmental phonology deals with this distinction by representing stress separately from intonation, or tone (Gussenhoven, 2004). Within the autosegmental framework, Pierrehumbert (1980) proposed an intonational model based on a thorough phonetic analysis of American English intonation contours. Later models (e.g. Beckman & Pierrehumbert, 1986; Gussenhoven, 1983, 2004; Ladd, 1996) have built on Pierrehumbert's initial model to more adequately explain circumstances such as downstepping, the downward shift of tone between syllables or words of tonal languages. Full review of these models is beyond the scope of this thesis. Discussion will be limited to the general trends identified within Pierrehumbert's model.

In the model, Pierrehumbert represents pitch contours phonologically as sequences of discrete intonational events. These events correspond to pitch accents and edge tones in English. Pitch accents form the basis of pitch contours and are likely to involve a pitch change which in turn tends to cue the presence of prominent syllables. These pitch accents, when further analysed as tones, can consist of either a single tone, H (high) and L (low) or a combination of the two tones. The central tone of a pitch accent is labelled with an asterisk, H* or L*. Edge tones refer to either boundary or phrase tones. Boundary tones are single tones (H% or L%) associated with the end of an intonational phrase. Phrase tones are also single tones (H or L), but occur between the last pitch accent and the boundary tone. The phonetic realisation of these tones is considered separately. In Pierrehumbert's analysis, every intonational phrase consists of one or more pitch accents and ends with both a phrase tone and boundary tone.

In Cruttenden's (1986) descriptive account of intonation based on previous theoretical research (e.g. Pike, 1945; Crystal, 1969; Ladd, 1979) and his own analysis of F0 contours, he further discusses these intonation phrases, or groups. Within each of these groups, one syllable is said to bear the main accent or nucleus despite no specific nucleus (or 'head') being defined in Pierrehumbert's model. Pierrehumbert later argues this to be true on the basis of research showing that nuclear accents are not phonetically different to prenuclear accents (Silverman & Pierrehumbert, 1990). However, Ladd (1996) suggested that although nuclear and prenuclear accents may not be phonetically distinct, they may play different structural roles, therefore, the inclusion of prenucleus and nucleus accents in the model can provide an indication of the F0 contour structure.

Each intonation group must contain at least one accented syllable (Cruttenden, 1986) and there must be a pitch movement from or towards the accented syllable. The boundaries of intonation groups can be marked by pauses within the speech, which can be either unfilled (silence) or filled pauses. Filled pauses in English are often produced using [ə] or [m]. The length of the pauses tends to reflect the strength of the boundary (Cruttenden, 1986), with more major boundaries having longer pauses at the end. Pauses cannot be used as the only clue to boundary positions since pauses also occur due to hesitation within the speech. Other indicators of boundaries include the occurrence of an anacrusis, where one or more unstressed syllables occur at the beginning of a new piece of information, such as in the sentence 'I saw John yesterday/ and he was just off to London'. The words 'and he was' are likely to be unstressed, with 'just' bearing the first stress of the group, indicating the presence of an intonation group boundary. The final syllable in an intonation group also tends to be lengthened (Cruttenden, 1986). Another indication involves the pitch of unaccented syllables. Although changes in pitch level and accent most often occur on stressed syllables, they can also occur on unstressed syllables. When this happens, it is often an indication of the start of a new intonation group. In English, the nuclear tones can be classified on the basis of the initial pitch movement away from the nucleus (fall, rise or level), the start of this initial movement (high or low) and a second change of pitch direction after the nucleus, which produces complex tones such as rise-fall and fall-rise.

The intonation systems of languages differ. Speakers of English, Welsh and Romance languages will be used as subjects in these experiments and examples of cross-language differences will employ these languages.

In English, deaccenting of syllables may occur when repeating material or when lexical stress is moved from one syllable to another. For example, when using the '-teen' numbers within a sentence, the stress lies on the first syllable. However, during counting, the stress is moved to the final '-teen' syllable (Ladd, 1996). Romance languages resist deaccenting of syllables. Instead they employ a number of morphosyntactic strategies to achieve similar results. One such method is right-dislocation. This is where the constituent which would be deaccented in English is moved to the end of a sentence, leaving a pronoun in its place (Ladd, 1996). At the end of the sentence, its F0 will be lowered, achieving much the same effect as deaccenting. Consider the French example below. The noun 'passeport' (passport) has been moved to the end of the sentence, with the pronoun 'le' (it) replacing it, thus creating a similar effect to deaccenting the noun by placing it at the end of the phrase.

French: Mais tu l'as, ton passeport!
But you have it, your passport!

Another example of this from the Romance languages is with infinitive small clauses. English tends to place sentential stress on the final noun, such as 'I have a BOOK to read.' Italian, like Spanish, however places the sentential stress on the right-most content word, 'Ho un libro da LEGGERE' (Ladd, 1996).

Welsh is a Celtic language with a larger fundamental frequency range than English (Williams, 1999), hence it tends to sound more expressive than English. It has a characteristic intonation pattern where the second-to-last syllable on every word is accented. Welsh has no consistent lengthening of stressed vowels, or significantly increased intensity or higher F0 on these vowels, as in English (Williams, 1985). In terms of declarative sentences, Welsh sentences have a definite rise then fall at the end of the sentence (Cruttenden, 1986).

Despite differences between language intonation systems, similarities such as intrinsic vowel F0 occur. Intrinsic F0 refers to the consistent F0 difference between low and high vowels, such that low vowels have intrinsically lower pitch than high vowels. Whalen and Levitt (1985) surveyed 31 languages, from 11 out of the 29 main language families, covering all language types (tone, pitch accent and stress) and discovered that, despite differing vowel inventories across languages, intrinsic F0 did exist in all of them.

Although size of intrinsic F0 difference can vary in different circumstances, the existence of the difference remains consistent. Ladd & Silverman (1984) have shown that intrinsic pitch occurs within connected speech by comparing German high and low vowels in the fixed sentence 'Er hat '...' gesagt.', and three paragraphs of approximately 100 words each of read speech. They found that although the difference between the high and low vowels was significantly larger in the fixed vowel-carrier sentence, an effect of intrinsic pitch still existed in the read paragraph where the vowels occurred in a variety of prosodic contexts. Umeda (1981) also studied the F0 recordings of vowels in fluent speech, but found no effect of intrinsic pitch. In her study, the F0 of vowels in stressed syllables produced in two speaker's recordings of a 20 minute essay was measured. F0 was also measured at the beginning of any change of direction of the F0 contour. Although consistent effects of the preceding consonant were found, much as in non-fluent speech, the effect of intrinsic F0 for the vowels was not found. Ladd and Silverman (1984) claim that this lack of effect was due to Umeda's failure to control for the prosodic environment as she compared vowels from different positions within the intonation contours. The effect of position within the intonation contour on the vowel F0 is larger than that of intrinsic pitch, thus could drown out any differences in intrinsic F0 between the vowels. When Ladd and Silverman controlled for sentence position, nuclear stress and speaker range, an effect of intrinsic F0 was maintained in fluent speech as well as non-fluent.

Shadle (1985) also found that intrinsic vowel pitch occurs in both sentence context and outside of fluent speech. She made recordings of four speakers producing sentences containing a target word [rVd] with vowels [ɑ], [i] or [u]. F0 measurements of the vowels were recorded for four different positions of the target word in the sentence: initial, near-initial, near-final and final. Results showed that intrinsic F0 does exist in a sentence context, but in unaccented sentence-final positions, the difference between high ([u], [i]) and low vowels ([ɑ]) decreases. This replicated results from Ladd and Silverman (1984) where a larger difference in intrinsic F0 was found in the phrase-final position for stressed vowels than for unstressed ones.

Steele (1986) investigated the effect of prominence on the F0 of the high vowel [i] and low vowel [æ]. She used sentences in which the proper nouns could be readily exchanged without changing the phonetic environment or syntactic structure of the sentences.

'Peter baked a cake for Patty.'
'Patty baked a cake for Peter.'

Eight speakers, both male and female, recorded ten versions of each sentence and were instructed to speak with increasing emphasis on each repetition. Question versions of each sentence with different emphasis were also created. As the emphasis was increased for both the statements and questions, the intrinsic F0 difference increased between the two types of vowel leading Steele to conclude that prominence was an important factor affecting the magnitude of the intrinsic F0 difference.

Intrinsic F0 has, therefore, been shown to occur both outside of and within fluent speech. It appears that sentence position in addition to syllable prominence affects the size of this intrinsic difference but that in spite of this variance, intrinsic F0 still exists in all conditions.

1.1.3.2 F0 prediction

Synthetic speech aims to be both intelligible and natural sounding. Accurate synthesis of intonation is therefore highly important for synthesised speech in order to produce natural sounding synthesis that is linguistically meaningful. F0 models produce explicit F0 predictions based on trends such as those identified by Cruttenden (1986). Both the ToBI and Tilt models have been shown to produce acceptable F0 contours on informal listening tests (Dusterhoff & Black, 1997), although the models are derived from different approaches. ToBI has a linguistic approach to F0 synthesis based on Pierrehumbert's (1980) F0 model. It specifies a small number of distinct labels used to identify the intonational space of accents and tones. Tilt, on the other hand, adopts a more data-driven approach using an 'event detector' to produce a transcription of speech intonation from data input. It then creates algorithms for predicting F0 contours from these transcriptions.

The TOBI acronym is short for the 'TOnes and Break Indices' intonational model for speech synthesis (Silverman et al., 1992). Using Pierrehumbert's (1980) intonational model of speech, it marks stressed syllables with tones (high, H and low, L). Each intonational phrase is made up of a sequence of H and L tones. The height of any of the tones is determined by its relationship to the baseline frequency, the degree of prominence given to the syllable by the speaker and its relationship to preceding tones. The ToBI system itself does not define a mechanism to go from the tone labels to an F0 contour, or the reverse. The model is applied to synthesis using a time-varying F0 range coupled with rules determined to assign the F0 values to the stressed syllables.

The basic unit in the Tilt model (Taylor, 2000) is the intonational event. These events occur as separate instances with nothing between them. The two basic types of intonational event are pitch accents and boundary tones. Pitch accents give emphasis or focus to a particular syllable or word through a noticeable change in F0. Boundary tones in the Tilt model refer to a rise in F0 towards the end of an intonational phrase. A combination event occurs when a pitch accent and boundary happen so closely together that they are perceived as one pitch movement. Every event is linked to a syllable in the speech, but not every syllable is represented by an event. The sequence of events is called an intonational stream. This intonational stream is then applied to the sequence of phones, or segmental stream of the utterance. Unlike the ToBI system, which uses categories, a set of continuous parameters are employed within the Tilt model. These are called the 'Tilt parameters'. There are three Tilt parameters: duration, amplitude and the tilt. Duration refers to the sum of the rise and fall durations. Amplitude is the sum of the magnitudes of these changes. The tilt parameter expresses the overall shape of the event, independent to the duration or amplitude parameters.

1.2. Speech intelligibility

During conversation, a listener aims to parse the continuous speech waveform they hear into meaningful units of language. This is a complex task. Speech sounds are not discrete events. That is, co-articulation of speech sounds can occur such that neighbouring sounds affect the production of each other. This is an example of the effect that context can have on speech production, such that isolated sounds often sound different to those produced in fluent speech. Also, sentence position affects the articulation of speech sounds such that the same sound produced in a sentence-final position may be dissimilar to one produced in a sentence-medial position. Different speakers produce different articulations of these sounds and depending on their emotional state at the time of the utterance, additional variation may occur. The speech signal may also be masked by background noise, making it more difficult to determine the sounds. The relative intelligibility of speech is affected by the quality of its transmission channel. Speech spoken in background noise through a telephone (a band-limited channel) or in a room (a reverberant channel) will tend to be less intelligible than speech spoken into a close studio microphone with less noise interference.

Models, such as the Articulation Index (AI) and Speech Transmission Index (STI) have been developed to predict the intelligibility of speech sounds based on their physical parameters. The Articulation Index (AI) (French & Steinberg, 1946; Fletcher & Galt, 1950) is a measure of speech intelligibility determined by the intensity of speech received by the ear and takes into account the intensities of any unwanted sounds. It works on the assumption that the speech signal can be divided into frequency bands, each of which carries an independent contribution to the speech signal. The contribution from each band is then summed, producing a number from 0 to 1 indicating the overall gain of the system (Müsch, 2000). The Articulation Index predicts the intelligibility of linear speech transmission channels with additive noise.

The Speech Transmission Index (Steeneken & Houtgast, 1980), on the other hand, can account for non-linear distortions such as reverberation and echoes. It compares the speech signal received to an 'ideal' speech signal that the model produces. Any reduction in modulation depth between the ideal signal and the actual speech are taken as reductions in intelligibility.

Both the AI and STI are able to cope with simple speech input, and can explain continuous background noise and masking. However, neither model is particularly accurate in more complex environments such as when speech is distorted nor do they take into account the effects that different types of masker, such as speech as opposed to noise, can have on the speech signal. Both systems also need to be adjusted for the speech materials being used depending on factors such as their predictability.

The predictability of the speech materials affects the intelligibility of speech utterances. Boothroyd and Nittrouer (1988) demonstrated that phonemes placed in nonsense words of the form CVC were harder to recognise in noise than phonemes placed in real words in the same format. Listeners were more likely to correctly identify the phonemes in the words 'pass', 'time' and 'make' than in nonsense words such as /puk/, /teIt/ and /sig/. Random sequences of words (e.g. 'girls white car blink') were also less easily recognised than words in grammatically-correct sentences with few contextual cues (e.g. 'ducks eat old tape'). These were in turn harder to correctly identify than words placed in grammatically-correct sentences with contextual cues (e.g. 'most birds can fly'). It seems that listeners rely on contextual cues to help with sentence intelligibility. Therefore, the more contextual cues present, the better the phoneme, word or sentence recognition. In the following experiments, I hope to establish whether F0 cues can aid sentence intelligibility. Low-predictability sentences will be used to ensure than the listeners can make relatively little use of contextual cues to predict the words in the sentences, however grammatical and semantic structure will remain to which the F0 contour may be linked.

The experiments in this thesis will all be performed against competing noise. Masking the speech signal decreases the intelligibility significantly. However, using masking in experiments allows the assessment of speech intelligibility through analysis of error rates. Speech Reception Thresholds (SRTs) measure the signal-to-noise ratio (SNR) at which speech can be accurately identified for at least half the time (Plomp & Mimpen, 1979). Speech materials within this method consist of a set of simple everyday sentences. In Plomp and Mimpen's method, the first sentence is presented at a highly adverse SNR. This SNR is successively raised by 4 dB until the listener thinks that half the target sentence can be correctly heard. The SNR of subsequent sentences is adjusted by 2 dB using a one-up, one-down procedure depending on the accuracy with which the listener hears the target sentence (Levitt, 1971). If they achieve a criterion level of accuracy, SNR is reduced by 2 dB. If they are less accurate, the SNR is raised. Therefore, the resulting measurement is a SNR representing the target level at which they can achieve this specified level of accuracy. The effectiveness of the maskers depends on, amongst other things, the type of masker and the similarity of the masker to the target.

1.2.1 Masking

When the listener attempts to attend to a target signal against a competing signal, masking is said to occur. Masking is often divided into two categories: energetic and informational (e.g. Drullman & Bronkhorst, 2004). Energetic masking occurs when the interfering signal is more intense than the target signal, making the target speech inaudible in places. Informational masking is said to occur when the effect of masking cannot be attributed to energetic masking. This occurs, for example, when the linguistic content of the interfering speech becomes confused with that of the target speech, causing linguistic intrusions from the interferer. Informational masking tends to occur in impoverished listening conditions, with few grouping cues, such as when both voices originate from the same location or are within the same F0 range.

1.2.1.1 Masker type

Different types of maskers have been shown to produce different effects on speech intelligibility. Continuous white noise is a more effective masker than modulated noise (Miller & Licklider, 1950; Carhart, Tillman & Greetis, 1969; Festen & Plomp, 1990; Gustafsson & Arlinger, 1993). Carhart et al. (1969) compared the effect of white noise, modulated white noise and connected speech on listeners' spondee recognition. A spondee is a metrical unit with two long or stressed syllables. They found that modulating white noise four times a second reduced the amount of masking produced by 3.8 dB in comparison to unmodulated white noise. Therefore, the listener could perceive the target at a level 3.8 dB lower in modulated noise than in unmodulated noise. Connected speech produced the same reduction as the modulated noise. White noise and modulated noise differ as modulated noise, like connected speech, permits dip listening whereas white noise does not. Dip listening occurs when the listener is given glimpses of the target speech through the interfering noise due to the natural dips and pauses present in the interfering waveform. Noise that is modulated like speech therefore inherits these fluctuations, enabling the listener to dip listen. Miller and Licklider (1950) also found that speech interrupted by either regularly spaced or randomly-spaced bursts of noise was more intelligible than speech heard in the presence of continuous noise. They found that with as many as ten bursts of noise per second, listeners were still able to accurately record the test words 75% of the time. This is an extreme example of the opportunity for the listeners to dip-listen. However, it reflects the fact that modulated noise gives the listener glimpses of the target speech from which they are able to determine the words more clearly than in continuous noise.

Festen and Plomp (1990) measured SRTs for speech presented in steady-state noise, modulated noise and against a single-talker interferer. They found approximately 4 dB improvement in SRT for speech presented against modulated noise compared to a background of steady-state noise. Unlike the Cahart et al. (1969) experiment, however, a further 3 dB improvement in SRTs was seen with a single-talker interferer compared to modulated noise. Although modulated noise mimics the fluctuations from connected speech, allowing the listener to dip listen, it does not have an F0 like a single-talker interferer. Differences in F0 assist the listener (see below). It could be that this lack of F0 information increases the masking effect of the interferer.

Hawley, Litovsky and Culling (2003) further investigated the effect of different types of interferer on speech intelligibility. They measured SRTs for sentences presented against four types of interferer: same talker, time-reversed speech of the same talker, speech-spectrum shaped noise (white noise filtered by the spectrum of a speech sample) and speech-spectrum shaped noise modulated by the temporal envelope of the interferers. The number of interferers varied from one to three, as did their spatial locations. Results showed that where only one interferer was used, unmodulated noise produced the most masking. Modulated noise permitted dip-listening, and hence produced a smaller masking effect. The single-talker interferers produced the least masking. As the number of interferers increased, the masking from the speech interferers also increased. Results were explained in terms of an F0 cancellation mechanism (de Cheveigné, 1997) where although a listener is able to effectively cancel the F0 of a single-talker interferer, when multiple interferers and therefore multiple F0s, are presented to the listener, they are less capable of cancelling the interfering F0. Thus, the masking from multiple speech interferers is greater.

Indeed, Qin and Oxenham (2003) reported that the degree to which F0 cues are available determines the effectiveness of a speech masker. Stimuli for this experiment consisted of target sentences presented either against a single-talker interferer, modulated speech-shaped noise or steady-state speech-shaped noise. SRT measurements were taken for four processing conditions: 4, 8, 24 channels or unprocessed. Processing was applied to both the target and masker speech, with the aim of simulating a cochlear-implant, which reduces the F0 cues in speech. For the unprocessed speech, single-talker interferers were found to be less interfering than steady-state noise, replicating previous findings (e.g. Festen & Plomp, 1990; Peissig & Kollmeier, 1997). However, as F0 information was reduced through cochlear-implant simulation, by reducing the number of unprocessed channels, this advantage was lost. This confirms that normal-hearing listeners are able to exploit F0 cues to segregate the target and masker. When F0 cues are available to the listener, the target and masker can be more readily separated than without. The masking of the speech interferer increases with the loss of F0 cues, indicating that it is indeed F0 cues that differentiate modulated white noise from a single-talker interferer with respect to masking.

As the number of interfering talkers is increased, the amount of masking also increases (Peissig & Kollmeier, 1996; Drullman & Bronkhorst, 2004). The more background talkers present, the greater the chance of the dips in one talker's speech being filled by another talker and the less effect of F0 difference there is. Peissig and Kollmeier (1996) measured Speech Reception Thresholds (SRTs) for German sentences presented in noise. Babble sounding much like speech-shaped noise was generated by superimposing 524 words randomly. The speech background noise was produced using four different talkers, two female and two male. SRTs were 17 dB lower for target speech presented against interfering speech than against noise. However, as the number of interfering talkers increased, the SRTs rose. There was a 6 dB difference between the single-talker interferer and the two-talker interferer. With three interfering voices, the SRTs rose by a further 6.9 dB from the two talker condition reflecting a decrease in the number of pauses in the interfering sound with three voices overlapping that the listener could exploit to hear the target voice. Drullman and Bronkhorst (2004) confirmed these findings in an experiment measuring the SRTs for speech presented against one, two, four or eight-talker interferers compared to a noise masker. A single-talker interferer has the least masking effect, with SRTs approximately 8 dB lower than that of the two-, four- and eight-talker interferers. Speech-shaped noise proved to be a less effective masker, with a 3 dB difference between it and the multi-talker maskers. Therefore, increasing the number of interfering talkers reduces the listener's opportunity to dip-listen and increases the masking of the interfering speech.

The number of interfering voices is also thought to affect situations using 2 or 3 voice interferers (Carhart et al., 1969; Freyman, Balakrishnan & Helfer, 2004), and becomes less prominent as the number of voices increases above this amount. Freyman et al. (2004) showed that when interfering voices are perceived to come from the same location, the spatial difference advantage does not completely disappear, although masking decreases as the number of talkers increases from 3 to 4 to 6 to 10. The spatial difference advantage is the benefit gained from increasing the perceived distance between targets. Hence, with 1 or 2 interfering voices, recognition performance improves when the target and interferer appear to be from different locations, but as the number of interfering talkers from separate directions is increased, this cue no longer aids recognition performance. This effect is proposed to be due to the rise of informational masking as the number of voices is initially increased to 2 voices due to a potential increase in interfering linguistic information, and then a consequent fall as the number of voices increases to 3. As the number of talkers in the interferer increases, it is possible that they begin to mask each other, enabling the listener to pay more attention to the target voice.

1.2.1.2 Target-masker similarity

Decreasing target-masker similarity tends to reduce the informational masking effects of the stimulus (Durlach et al., 2003; Brungart, Simpson, Ericson & Scott, 2001). Brungart et al. (2001) used the Coordinated Response Measure (CRM) to investigate the effect of speaker similarity on interference. These sentences are of the form 'Ready (call sign) go to (colour) (number) now'. The listener is instructed to listen for their call sign and then indicate what the colour and number for their call sign is. Competing sentences are of exactly the same form, with different call signs. Studies relating the CRM task to the AI (Brungart, 2001) have shown that the CRM task is relatively insensitive to energetic masking by noise, which makes it an appropriate choice for use in experiments focussing on the effects of informational masking. Brungart et al. (2001) included conditions with two talkers, three talkers and four talkers. Using the same talker for the target as for the interferer generated more masking than using a different talker of the same sex, which in turn created a greater effect of masking than using a talker of the opposite sex. The single-talker interferer produced less masking than the conditions with more than one interfering voice, confirming discussion above that multi-talker interferers produce more masking than a single-talker interferer.

Darwin, Brungart and Simpson (2003) used the CRM corpus to investigate the effect of manipulating F0 and vocal tract length on discrimination of two simultaneous talkers. The first experiment manipulated the F0 alone, shifting the female talkers by -9, -3, -1, 0, 1 or 3 semitones, and the male talkers by -3, -1, 0, 1, 3, 9 semitones. Listeners were presented two sentences recorded by the same talker, but at varying F0 differences. Results found, on average, an increase in percent correct from 0 to 4 semitones difference, with a further increase from 6 to 12 semitones difference. A second experiment investigated vocal tract length alone and used 8 different manipulations of the vocal tract length. Increasing the length of the vocal tract produces a more masculine sounding voice, whereas reducing it produces a more feminine voice. Again, the larger the difference in vocal tract length, the more successful the listeners were at segregating the two voices. A small difference in vocal tract length (8%) produced little effect; however, a change of 13% increased performance from 50% to around 55% and a 38% change in vocal tract length produced 67% correct responses. A third experiment manipulated both the F0 and vocal tract length simultaneously. Combining the two cues produced a larger increase in performance than manipulating only one variable. These results once again demonstrate that decreasing target-masker similarity increases the listener's ability to track the target. In addition, by shifting the F0 and vocal tract length of the speakers, the gender of the speaker is being 'changed'. This may account for results reported by Brungart et al. (2001) which investigated the effect of having different sex target and maskers.

1.2.1.3. Segregation by F0

Differences in F0 can also influence the effectiveness of a masker. The identification of target speech is better when the F0 differs from that of the interfering speech, as seen above in the Darwin et al. (2003) study. Previous experiments have shown that listeners are able to better segregate two speech sources if they differ in mean F0 (Brokx & Nooteboom, 1982; Bird & Darwin, 1998; Assmann, 1999; Drullman & Bronkhorst, 2004). Using monotonous target and interfering speech, Brokx and Nooteboom (1982) found that word recognition rates increased from approximately 40% correct at 0 semitones difference to roughly 60% correct at 3 semitones difference. A further measurement taken at 12 semitones difference showed a decline in the F0 segregation advantage, with approximately 50% correct word recognition. This reduction in performance at the octave presumably occurs because the harmonics of one voice coincide with those of the other. Also using monotonised speech, Bird and Darwin (1998) have shown that the segregation advantage increases beyond the 3 semitone difference found by Brokx and Nooteboom up to 8 semitones, with participants achieving approximately 60% correct recognition at a 3 semitone difference, and 80% at 8 semitone difference. However, this does not reflect a steady increase in recognition scores; no improvement was seen from 3 to 6 semitones difference.

Using both intonated and monotonous speech, Assmann (1999) showed that a gradual increase in speech intelligibility occurs from 0 semitones difference up to 8 semitones difference in mean F0 between the target and interferer. Drullman and Bronkhorst (2004) reported that listeners' performance progressively improves with increasing F0 difference up to 12 semitones when the speech retains naturally-modulated F0s, as opposed to the speech with fixed F0s used by Brokx and Nooteboom. At 0 dB Signal-to-Noise Ratio (SNR), listeners correctly identified 46% of target words at a 0 semitone difference. This increased to 76% correct recognition at 4 semitones difference, 97% accuracy at 8 semitones difference and 99% at 12 semitones difference. In naturally-modulated speech, the F0 fluctuations within speech mean that although the average F0 difference between the target and interferer is fixed, the difference at any point in time fluctuates around this average. For the octave condition, this effect prevents the coincidence of harmonics that occurs if the speech is monotonous. Therefore, it seems that in the case of naturally-intonated speech, a greater difference in mean F0 between the target and interferer produces a monotonically larger segregation advantage for the listener. This is because there is less chance that the target and interferer F0 values will overlap. However, the difference between 8 and 12 semitones is not particularly large.

A mechanism for F0 segregation was proposed by de Cheveigné (1993, 1997). In this neural-cancellation model, a harmonic sound could be segregated by isolating the F0 of the interfering sounds and cancelling it. According to the model, the listener is able to identify target speech more successfully when it differs in F0 from the interfering speech because of a cancellation process that perceptually removes the interfering speech on a particular F0. Using this model, in principle there is no limit to the number of interfering harmonic sounds at different F0s that could be perceptually cancelled. However, Culling, Linsmith and Caller (2005) reported that when target speech was played against two interfering speech sources, the listener was best able to identify the target speech when the interfering voices were monotonised at a different F0 to the target speech, but the same F0 as each other. When the interfering voices were on different F0s to each other and to the target speech, only a minimal advantage was seen compared to when all voices were monotonised on the same F0. This implies that a cancellation mechanism exists but that it can only remove sounds at one F0, not multiple sounds at different F0s. A single-talker interferer will therefore produce less masking than multiple speech interferers on different F0s.

1.2.2 Effect of Masking on Non-Native Speakers

Several studies have examined the enhanced effect of interfering noise on non-native listeners. Mayo, Florentine and Buus (1997) compared four groups of listeners: monolingual, bilingual-since-birth, bilingual-since-toddler and bilingual-post-puberty, on both high- and low-predictability sentences presented in both quiet and competing babble. Signal-to-noise ratios were measured for the sentences presented in babble. All listeners performed with near-ceiling, native-like proficiency in quiet using high-predictability sentences. For the majority of speakers, context aided the intelligibility of the utterances. An effect of context was also demonstrated in quiet for non-native speakers. This effect was not present for native speakers. However, sentence predictability had no effect on speech intelligibility in noise in the bilingual-post-puberty group. This suggests that late second language learning hinders the listeners' ability to use contextual cues when speech is presented in noise. The results indicate that learning a second language at an early age is important for understanding speech in noise. Listeners who became bilingual after puberty performed worse than the early bilingual listeners. However, noise had a larger detrimental effect on early bilinguals than on monolingual speakers with low-predictability sentences. Mayo et al. (1997) proposed that this effect could be due to competition between the sound systems of the listener's first and second language, contributing to the added difficulty of non-native speech perception in noise.

Lecumberri and Cooke (2006) have also noted an increased susceptibility to noise for non-native listeners. They tested English and Spanish groups on both native and non-native perception of intervocalic consonants in quiet and in four different noise conditions. Listeners were asked to identify the consonant in the VCV stimuli presented in quiet, 8-talker babble, speech-shaped noise and competing English and Spanish speech. All listeners performed best when speech was presented in quiet. Performance deteriorated when speech was presented against a single-talker interferer (English and Spanish speech), was worse against speech-shaped noise and was poorest in the 8-talker babble. Spanish listeners were affected to a larger extent by the presence of speech-shaped noise than English listeners. English listeners were better able to tune out the Spanish speech interferer than the English speech interferer. The Spanish listeners were equally affected by both English and Spanish. This is likely to be because the Spanish speakers were familiar with both languages.

Cutler, Weber, Smits and Cooper (2004) reasoned that the disadvantage for non-native speakers in noisy conditions is not associated with a disproportionate decrease in phoneme identification. Cutler et al. (2004) investigated the phoneme identification performance of native American and non-native Dutch listeners in 6-talker babble at signal-to-noise rations of 0, 8 and 16 dB. Results showed that non-native listeners performed consistently worse than the native speakers. However, no interaction between the effects of native language and background noise was demonstrated in their phoneme identification task; the effect of the background noise was not significantly different for non-native than native speakers. Increasing the signal-to-noise ratio increased the phoneme identification accuracy for both groups at a similar rate. Therefore, Cutler et al. (2004) reason from the results of these experiments that phoneme identification is not solely responsible for the increased effect of noise for non-native speakers. Factors such as a lack of understanding of prosodic cues leading to poor segmentation and misinterpretation of sentence syntax could also contribute.

1. 3 Role of F0 in Sentence Intelligibility

Details of the impact of prosody, masking and sentence context on speech intelligibility have been discussed above. I am primarily interested in the role that fundamental frequency plays in speech intelligibility. If F0 is indeed important to speech intelligibility, manipulating the F0 contours and therefore altering the F0 cues, will detrimentally affect the intelligibility. Experiments reviewed in this section of the introduction will therefore detail studies where the F0 contour has been manipulated and the effects of these manipulations on speech intelligibility have been reported.

Wingfield, Lombardi and Sokol (1984) studied the intelligibility of speech presented in quiet at different speech rates. They found that when the speech rate of an utterance was increased, the advantage of having a sentence with 'normal' prosody, as opposed to a monotonised F0 contour or list-read speech, also increased. List-read speech consisted of a passage of speech where each word was read as though it were a word in a list. Each word had equal stress and there was no timing, amplitude or pitch variation across the words in the passage. At the speaker's average speech rate of 229 words per minute, list-read speech was found to be more intelligible than both normal-prosody and monotonised speech. Wingfield et al. (1984) argue that this is because in list read speech there is no effect of co-articulation. Each word is spoken clearly and with equal stress. However, as the speech rate increased to 460 words per minute, speech with normal prosody was more accurately reported than list-read and monotonised speech. In fact, even monotonised speech was correctly transcribed more consistently than list-read speech at the highest speech rates. This effect was accounted for by the fact that despite all F0 variation cues having been removed in the monotonised speech, the list-read speech had none of the remaining prosodic cues that the monotonised speech had, such as timing and amplitude.

In addition, Laures and Weismer (1999) studied the effect of having normal intonation in sentences on speech intelligibility. In this study, two male speakers read nine sentences selected from the Speech Perception in Noise (SPIN) test using a pronounced prosody. Each of these sentences were resynthesised using the LPC resynthesis technique. The F0 contour of the voiced segments was monotonised by adjusting their F0 value in Hertz to the arithmetic mean of the sentence. Ten listeners were presented with both normally-intonated and monotonous sentences in a background of white noise and asked to transcribe them. On a second presentation of the stimuli, listeners were asked to rate the sentences for intelligibility on a scale of 1-7; 1 being unintelligible and 7 being highly intelligible. A significant difference between the normally-intonated and monotonous sentences was found for both the number of words transcribed correctly and rated intelligibility. One explanation proposed for these findings is the fact that the rise and fall of the F0 contour directs the listener's attention to the content words of the utterance (Cutler, 1976; Cutler & Fodor, 1979); without these cues, the intelligibility of the utterance is lowered.

Using a phoneme-monitoring task, Cutler (1976) found that an integral part of processing sentences involves actively searching for the sentence accents. To demonstrate this effect, she recorded twenty sentences in three conditions: one where the target-bearing word was heavily-stressed, one where the word had very reduced stress and another where the target word bore neutral stress. The monosyllabic target word began with one of the target phonemes, /b/, /d/ or /k/. Slightly different sentence endings were used in each of the three conditions to ensure that the intonation patterns sounded natural. Examples of these sentences are shown below using the target word 'dirt'.

High stress on target: She managed to remove the dirt from the rug, but not the berry stains.

Low stress on target: She managed to remove the dirt from the rug, but not from their clothes.

Neutral stress on target: She managed to remove the dirt from the rug.

The stimuli were constructed by removing the target word from both the high- and low-stress conditions, and replacing it with the version from the neutral stress condition. Therefore, two versions for each target word were created; one with a surrounding intonation contour predicting high-stress and one predicting low-stress. Participants were primed to press a button as soon as they heard a word in the sentence beginning with the target phoneme. Responses to unaccented target words placed in the high-stress sentence were faster than those in the low-stress sentence. Responses to the unmanipulated sentences were faster for the high-stress sentences than the low-stress sentences. The response times for the unmanipulated high-stress sentences were also faster than for the unaccented word placed in the high-stress sentences. Thus, it appears that listeners used the surrounding intonation contour to guide them to the target word, whether it was accented or not. Having an accent on the target word did reduce response times, indicating an advantage of target accent. However, the surrounding F0 contour seems to be an important cue which listeners use to find the target word.

Cutler and Fodor (1979), also using a phoneme-monitoring task, extended this research to demonstrate that in searching for the sentence accent, listeners are looking for the semantic focus of the sentence. For the experiment, the semantic focus of the sentence was manipulated by presenting the listener with a question prior to the target sentence. For example, consider the following three sentences.

1) The woman with the bag went into the dentist's office.

2) Which woman went into the dentist's office?

3) Which office was it that the woman went into?

If sentence (1) is preceded by question (2), the semantic focus or answer to that question would be 'the woman with the bag'. Question (3) would instead make 'dentist' the focus of the sentence. The accented target position was varied as in the Cutler (1976) experiment. There were four different conditions: two where the semantic target and focus were aligned either early in the sentence or late, and two where they were in different places. Listeners were again instructed to respond to the target word beginning with /b/, /d/ or /k/. Results showed that reaction times to the semantic target were faster when the focus position was the same as that of the target. This result, in addition to findings by Cutler (1976), suggests that in actively searching for accented words, listeners are in fact looking for the semantic focus of the sentence. New or important information that the speaker wishes to convey tends to be highlighted to make it more prominent to the listener. These experiments show that the intonation contour is an important part of creating the sentence focus for native speakers.

Another contributing factor to the effect of F0 manipulation on speech intelligibility may be the presence of F0 movement. F0 modulation within a vowel relates to short-term changes of F0 direction in the intonation contour. The effect of frequency modulation on the identification of vowel sounds presented concurrently with interfering vowels was investigated by both McAdams (1989) and Culling and Summerfield (1995). McAdams (1989) asked listeners to judge the perceived prominence of vowels /a/, /i/ and /o/ in four conditions in which each of them was always presented. Each vowel was synthesised on three different mean F0s (130.8, 174.6, 233.1) either with or without a mixture of sinusoidal and random frequency modulation. In the first condition, none of the vowels were modulated, in the second all three vowels were modulated coherently, in the third one vowel was modulated and the other two remained steady. Finally, in the fourth condition one vowel was modulated independently (at random phases of modulation chosen from a uniform distribution spanning 360°) of the remaining two vowels which were modulated coherently together. Figure 1.1 depicts examples of these conditions.

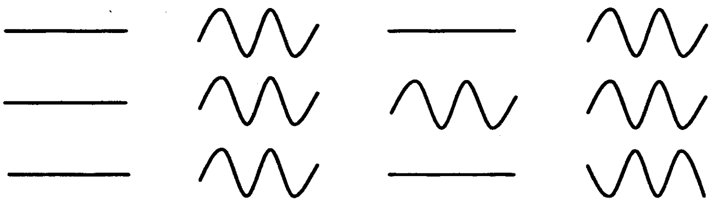

Fig. 1.1: Conditions for the McAdams (1969) experiment. The first condition is where all three vowels are steady. The second depicts all three vowels modulated coherently. The third condition is where one vowel is modulated, but the remaining two are steady. The final condition is where one vowel is modulated independently of the other two modulated vowels.

Results showed that F0 modulation increases the perceptual salience and identification of vowels against the background of other unmodulated vowels. These results suggest that frequency modulation can enhance the target by making it more salient than the competing sound. Culling and Summerfield (1995) later showed that this effect carries over to an identification task and is mediated by the movement of individual harmonics; even independent modulation of each component of a vowel produces a similar enhancement.

Our studies aim to further investigate the effects of F0 manipulation on speech intelligibility by decreasing the amount of F0 variation gradually. Experiments will include conditions where the contour retains only half or a quarter of the total variation present in the normally-intonated contour, along with an inverted F0 contour condition. An inverted F0 contour retains the variation present in the original F0 contour, but these variations are reversed so that where there would be a rise in the normally-intonated contour, there is now a fall, and vice versa. In experiments investigating the effect of reverberation on listeners' ability to separate two competing voices, inverted F0 contours have been found to detrimentally affect the intelligibility of an utterance (Culling, Hodder & Toh, 2003). Speech-Reception Thresholds (SRTs) were measured for sentences played against a single-talker interferer in both anechoic and reverberant conditions in two different spatial configurations (collocated and separate) for three different forms of intonation (original, monotonised and inverted). The result of interest here is that in all conditions, the inverted F0 speech was less intelligible than the monotonised speech, which in the anechoic conditions was less intelligible than the original speech. This result indicates that it is not simply the movement *per se* of the F0 contour that aids the listener in their comprehension of the utterance. This movement is still available in the inverted F0 contour; however it could be that the inverted contour itself is misleading. In a normally-intonated sentence, content words are highlighted by a number of cues, including their F0. Since this F0 cue is reversed in an inverted contour, it is possible that the listener's expectations of which words are important will be confused and hence sentence intelligibility may be compromised.

Hillenbrand (2003) has also studied monotonised and inverted F0 contours. He measured the intelligibility of synthetic speech presented in quiet. When the F0 contour was monotonised, there was a reduction in intelligibility, but when it was inverted there was no further effect. However, when the sentences were low-pass filtered at 2 kHz to reduce their overall intelligibility before manipulating the F0 contours, he found that the effect of manipulating the contours increased and the monotonous sentences became more intelligible than the inverted sentences. These results imply that the F0 contour's contribution to speech intelligibility increases in adverse conditions. This conclusion is supported in part by Wingfield et al.'s experiments with list-read speech (1984). Their study reported that the advantage of speech with normal prosody compared to monotonous and list-read speech increased with the speech rate. Therefore, it seems that cues which are not particularly important in easy listening conditions become more significant for intelligibility in difficult listening environments. In our experiments, adverse listening conditions were created through the presence of background noise.

In a similar set of experiments to those conducted by Laures and Weismer (1999), Laures and Bunton (2003) used two different types of interferer: white noise and multi-talker babble. The significant difference found in the Laures and Weismer (1999) paper was replicated both in the percentage of words transcribed correctly and in the rated intelligibility. However, no significant difference was found between the two types of interferer. The conclusion was that fundamental frequency variation is an important acoustic cue in noisy listening environments, regardless of whether the background noise is speech-like or not. However, the lack of difference between the two types of interferer could be because it is difficult to follow the meaningful speech of any particular talker and that the more talkers included in the babble, the less chance the listener has for dip-listening. Twelve voices were used in the multi-speaker babble here, which would severely reduce the listener's ability to exploit any dips or pauses in the speech. Against fewer competing voices, the results may be different.

1.4 Thesis overview

This chapter has outlined how prosody and in particular fundamental frequency contribute to speech perception. In addition, other factors that are known to influence speech intelligibility have been discussed. In later chapters, experiments will be reported which examine the effect of F0 on speech intelligibility. The first set of experiments (Chapter 2) focus on the effect of manipulating the F0 contour for speech presented in speech-shaped noise and single-talker interferers. Chapter 3 details an experiment in which the F0 contour of speech is low-pass filtered in order to determine where the important modulation frequencies within the F0 contour lie. Synthesised speech is studied in comparison to natural speech in Chapter 4. The F0 and duration contours of the synthesised speech are compared to those of natural speech in an attempt to discover the reason for the inferior intelligibility of synthesised speech. The above studies were conducted with native English speakers. Chapter 5 investigates whether F0 manipulation has an effect on speech intelligibility in non-native speakers in view of their reduced understanding of prosodic cues in a non-native language (Sanders et al., 2002).

CHAPTER 2:

MANIPULATING THE TARGET F0 CONTOUR

2.1 Introduction

Previous experiments have explored the role of intonation in speech recognition by monotonising the F0 contour and comparing the intelligibility of these utterances to normally-intonated ones in a quiet listening environment, with a background of white noise, against a single-talker interferer or multi-speaker babble (Wingfield et al., 1984; Assmann, 1999; Laures & Weismer, 1999; Laures & Bunton, 2003). These studies have found a detrimental effect of monotonising, hence removing F0 variation from the contour, on speech intelligibility. F0 inversion has been found to further decrease the intelligibility of speech (Culling et al., 2003; Hillenbrand, 2003). Inverting the F0 contour retains the F0 variation of a normally-intonated contour, but it is misplaced such that incorrect cues are placed on syllables within the speech. Thus it seems that having incorrect F0 cues on speech is more detrimental to its intelligibility than simply reducing them.

The following experiments measure the impact of F0 cues in the intelligibility of speech against interfering noise. Speech-shaped noise and a single-talker will be used as interferers. Speech-shaped noise approximates the long-term spectrum of speech and has a similar masking effect to that of multi-speaker babble, creating a more natural speech-like background noise. A single-talker interferer, unlike speech-shaped noise, contains F0 information which can be used to segregate the target and interfering speech (Festen & Plomp, 1990; de Cheveigné, 1997; Hawley et al., 2003).

One concern with these experiments is that having a large number of contextual cues in an utterance could potentially mask any suprasegmental cues to sentence intelligibility. Indeed, in Assmann's 1999 experiment no significant effect on sentence intelligibility was found of removing F0 variation. In this experiment a set of 48 high predictability sentences from the SPIN test were used. In the monotone condition, sentences were monotonised at 100 Hz using the STRAIGHT speech analysis-synthesis program. Sentences were presented in pairs to the listeners twice. One sentence in each pair had a mean F0 of 100 Hz, whilst the other was 0-8 semitones higher depending on the F0 difference condition required. On first presentation, listeners were instructed to transcribe as many of the words as possible from either sentence. A second presentation allowed the listener another attempt at the transcription. The number of keywords correctly transcribed from either sentence was taken as the measurement of identification accuracy. Although the mean transcription scores were approximately 8% higher for intoned sentences than monotone sentences in the 0 and 1 semitone average F0 difference, the difference between the intoned and monotone sentences was not statistically significant. The 8% difference between the monotone and intoned sentences was reduced as the difference in average F0 increased up to 8 semitones, implying that as it became easier to segregate the two sentences, the importance of the F0 variation decreased. It could be that the lack of a significant effect is due to the use of high-predictability sentences. It is possible that the listener does not need to rely on the F0 contour of the speech when they are able to use semantic cues to predict the following words. For this reason, low-predictability sentences will be used in this experiment to ensure that the listeners can make relatively little use of contextual cues to predict the words in the sentences.

Reducing the F0 variation incrementally from the variation present in a normally-intonated contour to a monotonous contour, with no variation, will enable us to see whether the more variation in the contour, the more intelligible the speech or whether it is the case that just a small amount of F0 variation can give the listener enough cues to produce the same intelligibility. Based on previous studies (e.g. Culling et al., 2003, Hillenbrand, 2003) indicating that the listener extracts information from the F0 contour to enhance speech intelligibility, it was expected that placing inverted F0 contours on the speech should mislead the listener because the F0 cues will now be incorrect.

2.2. Experiment 1

2.2.1. Method

2.2.1.1. Listeners

20 paid participants were recruited from Cardiff University Participation Panel. All were native speakers of English and had normal hearing. None of the listeners were familiar with the sentences used in the study.

2.2.1.2. Stimuli

The set of 100 target sentences used in this experiment was from the Harvard IEEE corpus (Rothauser et al., 1969). The recordings made at M.I.T. of male voice DA were digitised at 20 kHz sampling rate with 16-bit quantization. All the sentences were of low predictability and included five nominated keywords. For example,

"a WISP of CLOUD HUNG in the BLUE AIR"

Each sentence was manipulated using the Praat PSOLA speech analysis and resynthesis package. The F0 contour of each sentence was manipulated using equation (1); *m* is the coefficient for the particular manipulation; $\overline{F0}$ is the mean fundamental frequency. The mean F0 of the target speaker was 107 Hz. The average pitch range of the talker was 120 Hz. By inverting the F0 contour, there were times when the F0 reached below the acceptable range for Praat. Praat cannot synthesise F0 values lower than about 70 Hz, hence when inversion produced values below this limit, Praat incorrectly synthesised the F0 at a higher value. For this reason, the F0 contour of each sentence was multiplied by 1.5 in order to raise the mean F0 so that the F0 contour could be inverted without falling below the accepted F0 range for the Praat software package. The F0 range of different voices is approximately equal on a logarithmic scale (Graddol, 1986; Traunmüller & Branderud, 1989; Nolan, 2003), hence this was preferred over a linear scale for manipulating the F0 contour across sentences. After this, each of the sentences was resynthesised.

$$F0' = \left[1.5\,\overline{F0}\,\exp(\,m\,\ln(\,F0\,/\,\overline{F0}))\right] \tag{1}$$

Five different manipulation coefficients (*m*) of the F0 contour were applied (1, 0.5, 0.25, 0, -1), corresponding to the five conditions (original, half, quarter, monotone, inverse), see fig. 2.1. The original contour refers to the normal intonation placed upon that sentence by the speaker. The half and quarter conditions follow the same general shape of the original contour, but the amount of variation in both is reduced; in the half condition this variation is half that of the original contour and in the quarter, it is quarter. The monotone condition refers to a monotonised F0 contour, and the inverse to an inverted contour.

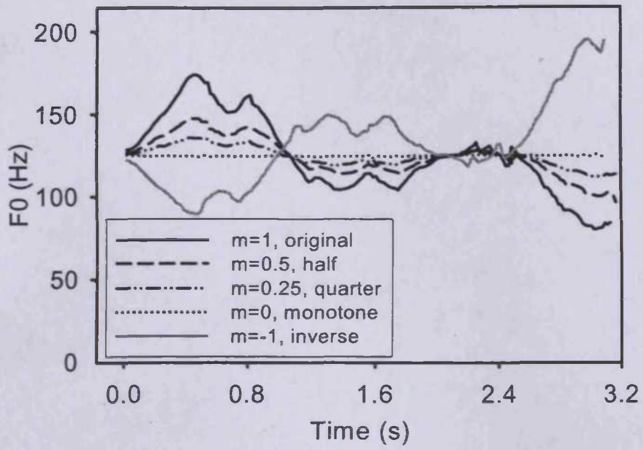

Fig. 2.1: Manipulations (*m*) of the F0 contour for an example sentence ('All sat frozen and watched the screen.') from Experiment 1. Manipulations are *m* = 1, 0.5, 0.25, 0 and -1, corresponding to the five conditions of the experiment: original, half, quarter, monotone and inverse, respectively.

Speech-shaped noise was created by processing white noise with a 512-point digital FIR filter whose magnitude response was equal to the long-term average spectrum of the entire set of unmanipulated target sentences. This noise was used as the interferer and edited to be, on average, 0.1 seconds longer than the speech targets so that no part of the speech target was presented in silence at any stage.

2.2.1.3. Procedure

Participants were seated in an IAC single-walled sound-attenuating booth in front of a computer screen visible through the booth window. They listened to stimuli over Sennheiser HD-590 headphones and responded to them using a keyboard placed inside the booth.

Speech Reception Thresholds (SRTs) were measured for each condition in the experiments (Plomp & Mimpen, 1979; Culling & Colburn, 2000). In the first phase of an SRT measurement, the sound level of the target started low. The listener's instructions were to attend to the voice and ignore the interfering noise. The listener was allowed to hear the same target sentence and interferer repeatedly at the start of each run by pressing the return key on the keyboard. Each time this key was pressed, the level of the target sentence was increased by 4 dB and the stimulus repeated. Once the listener believed that they could hear more than half the sentence correctly, they typed out their transcript. The correct sentence was then displayed underneath the listener's transcript, containing five capitalised keywords. The listener compared these words to those in their own sentence and entered the number of keywords that they had heard correctly (0-5). The second phase of the measurement then began. The listener could now only hear the sentence once and an SRT was measured using a one-up/one-down adaptive SRT technique. That is, if the listener reported three or more words correctly, the level of the target speech was decreased by 2 dB; otherwise it was increased by 2 dB. The SRT measurement was complete once the listener had heard and written out their transcript for all ten target sentences. The measured SRT was the average signal-to-noise ratio for sentences three to ten.

The session began with two practice SRT measurements with normally-intonated sentences played against speech-shaped noise interferers. This practice familiarised the listener with the task. Ten further SRT measurements were made. Two SRT measurements were made for each condition (original, half, quarter, monotone, inverse). The order of the presentation of the conditions was rotated for each listener, while the sentence material stayed in the same order. This ensured that any effects of order or materials were counterbalanced.

2.2.2. Results and Discussion

Figure 2.2 shows that there was a slight increase in SRTs as *m* was reduced, from the original targets (*m*=1) to the monotonous (*m*=0) targets, with a greater increase for targets with an inverted F0 contour (*m*=-1).

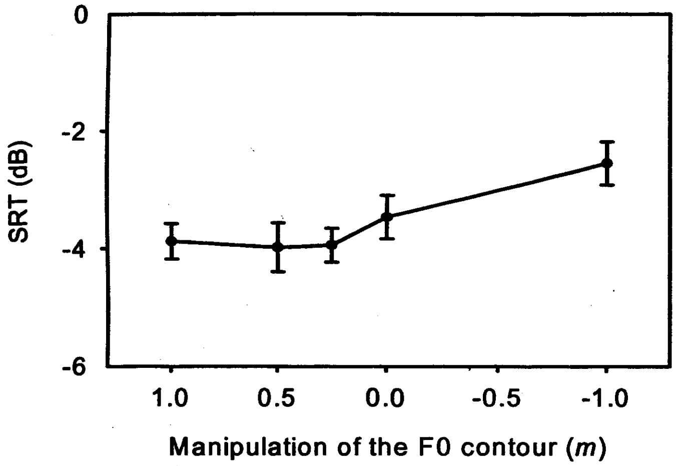

Fig. 2.2: Mean SRT measurements (dB) across all listeners for the different manipulations of the F0 contour (*m*) in Experiment 1. Manipulations *m* = 1, 0.5, 0.25, 0 and -1 correspond to conditions original, half, quarter, monotone and inverse. Error bars represent ± 1 standard error.

The difference between speech with an original contour (*m*=1) and monotonous speech (*m*=0) was 0.41 dB and was 1.33 dB between original speech (*m*=1) and speech with an inverted F0 contour (*m*=-1). A repeated measures ANOVA found a significant main effect of the F0 contour (F (4, 76) =3.525; p<0.02). Post-hoc Tukey HSD tests showed there to be a significant difference between the original, half and quarter conditions (*m*=1, 0.5, 0.25) and the inverted condition (*m*=-1) at the 0.05 significance level. No other comparisons were significant.

The reduction in SRT when the F0 contour was monotonised was non-significant. Thus the results did not reflect those of earlier studies where monotonising the F0 contour had proved significantly detrimental to speech intelligibility (e.g. Assmann, 1999; Laures & Weismer, 1999; Laures & Bunton, 2003, Culling et al., 2003). Inverting the contour, on the other hand, significantly reduced the intelligibility of the utterance, similarly to the Culling et al. (2003) results. The fact that the quarter and half conditions did not substantially reduce the sentence intelligibility, whereas inverting the contour did, implies that as long as there is only a small amount of F0 modulation in the 'correct' direction, the intelligibility of the utterance is almost unaffected.

The inverted condition produced a statistically significant decrement but still one that was smaller than observed by Culling et al (2003). One difference between the current experiment and that of Culling et al. (2003) was that a single-talker interferer was used in their experiment instead of speech-shaped noise. Single-talker maskers have been shown to produce less masking than speech-shaped noise, which is thought to be due to the listener's ability to exploit the dips in the temporal envelope of the speech and the exploitation of F0 differences between voices (Festen & Plomp, 1990; Peissig & Kollmeier, 1997). Although it is clear from the Culling et al. (2003) experiment that the interferer F0 is important in some way due to the increased F0 inversion effect compared to that found in Experiment 1, it is not obvious why this would be the case. Culling et al. (2003) inverted both the target and interferer F0s simultaneously, but did not perform any other manipulations. For these reasons, a factorial manipulation of both the interferer and target contours was performed in order to investigate what role the interferer F0 plays in speech intelligibility.

2.3. Experiment 2

2.3.1. Method

2.3.1.1. Listeners

18 undergraduate Psychology students were used as participants in this experiment. All were normally-hearing native speakers of English. None of the participants were familiar with the sentences used, nor had they taken part in the previous experiment.

2.3.1.2. Stimuli

The set of sentences used here was selected from the same corpus as in the previous experiment. However, the recordings of another male voice, CW, were used instead of DA. This voice was also digitised at 20 kHz with 16-bit quantisation. The sentences were once again manipulated using Praat PSOLA. The following formula was applied to the F0 contour of each sentence before resynthesis, shifting the average F0 of all the target sentences to 125 Hz. This shift enabled a consistent difference in mean F0 of 9 semitones to be set between the target and interfering sentences.

$$F0' = \left[125 * \exp\left(m \ln(F0 \,/\, \overline{F0})\right)\right] \quad (2)$$

Only three values of *m* were used this time (*m*=1, 0, -1) corresponding to the three conditions (original, monotone, inverse).

A 9 semitone F0 difference was used to avoid potentially confounding effects of the F0 difference (Assmann, 1999; Brokx & Nooteboom, 1982; Culling & Darwin, 1994; Bird & Darwin, 1998). For instance, if both the target and interferer were monotonised with the same F0, there would be no F0 differences to exploit, but if the target had a normal F0 contour and the interferer was monotonous, the differences in instantaneous F0 could be used to differentiate between the two at the points in time where the normal F0 contour deviates from the mean. By introducing a constant difference in mean F0, we sought to minimise differences in mean instantaneous F0 difference between the various conditions of contour manipulation. Bird and Darwin (1998) showed there to be a monotonic increase in sentence identification from 0 to 8 semitones F0 difference between the target and masker. In the experiment by Brokx and Nooteboom (1982), the advantage of the F0 difference was found to decrease at one octave (12 semitones). However, no research has indicated what happens between 8 and 12 semitones. The 9 semitone difference used here was felt to be as large a difference as could be employed without encountering the decline in the F0 difference advantage found by Brokx and Nooteboom (1982), whilst at the same time being large enough to reduce the chance of overlap of the F0 contours. This ensured that any effect found from manipulating the contour was due to the contour change itself and not the difference in F0 between the target and interferer.

Speech interferers were used instead of noise and were created using ten sentences from the Harvard IEEE corpus. The recorded voice used was DA. The F0 contour of the interferers was manipulated using Praat PSOLA as for the target sentences. Eq. (3), which results in a fixed mean F0 of 210.25 Hz, a 9 semitone difference between the target and interferers, was applied to the sentences before resynthesis.

$$F0' = \left[210.25 * \exp\left(m \ln(F0 / \overline{F0})\right)\right] \tag{3}$$

Nine conditions were set up, using each possible combination (*m*=1, 0, -1) of both the target and interfering F0 contour. Hence the target F0 contour was manipulated in three ways to create a normally-intonated, monotonous and inverse condition for the target.

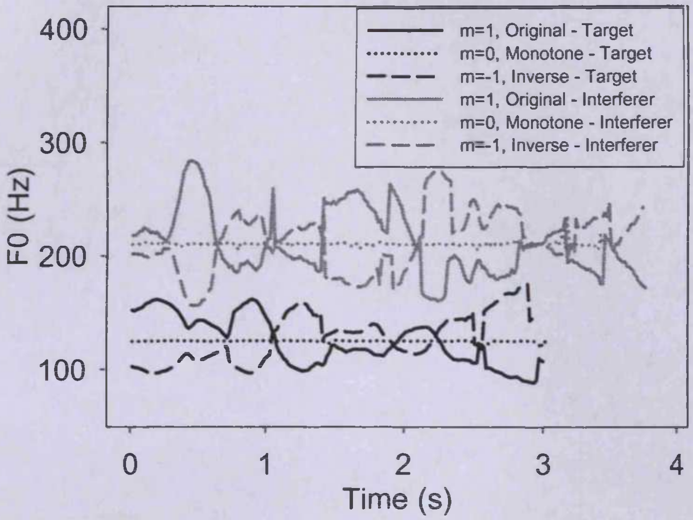

Fig. 2.3: Manipulations (*m*) of the F0 contour for both the target and interferer sentences in Experiment 2. The target speech is 'Where were they when the noise started?' and the interfering sentence is 'A ridge on a smooth surface is a bump or flaw'. The interfering sentence in this case is 0.6 seconds longer than the target sentence. The mean F0 of the target is set to 125 Hz, with the mean of the interferer being 9 semitones higher, at 210.25 Hz. Manipulations are *m*= 1, 0 and -1 for both the target and interfering contours, corresponding to the original, monotone and inverse conditions respectively.

Similarly, the interferer F0 contour was manipulated to be normally-intonated, monotonous or inverse. Figure 2.3 shows the different manipulations of both the target and interferer contours, around their F0 means.

2.3.1.3. Procedure

The procedure was similar to that in Experiment 1, measuring SRTs for each of the conditions within the experiment. Each experiment therefore consisted of nine runs, each with ten sentences, along with a practice run at the start of the experiment to familiarise the listener with the task.

As in the previous experiment, at the start of each run, the target level would be set low, relative to the interferer. The interfering sentence remained the same in both content and manipulation throughout each run. Each of the ten runs used a different interferer. In order to differentiate the target from the interferer, the listener was instructed to pay attention to the quieter sentence (target) at the start of the run and ignore the louder sentence (interferer). The listener typed out their transcript, as in Experiment 1, as soon as they judged they could hear more than half of the target sentence and scored the number of words they heard correctly. SRTs were recorded for each condition. Once again, the order of the conditions was rotated for each listener, while the sentence material stayed in the same order.

2.3.2. Results and Discussion

Figure 2.4 shows that there was a larger difference in SRT between the normally-intonated, monotonous and inverted target speech than in Experiment 1, where the interferer was speech-shaped noise. The difference between normally-intonated speech and monotonous speech was 2.0 dB, and was 3.8 dB between normally-intonated speech and speech with an inverted F0 contour. The difference between the conditions in this experiment closely resemble the results from Culling et al. (2003) where an approximately 2 dB difference between natural and monotone conditions and a 3 dB difference between natural and inverted conditions was observed.

A repeated measures ANOVA found a significant main effect of the target F0 contour ($F(2, 24) = 18.288$; $p<0.001$). Post-hoc Tukey HSD tests showed there to be a significant difference between the original (m=1) target F0 condition and inverse (m=-1) condition (p<0.01). The difference between the monotone (m=0) and inverse (m=-1) conditions, and the monotone (m=0) and original (m=1) conditions was also significant (p<0.05).

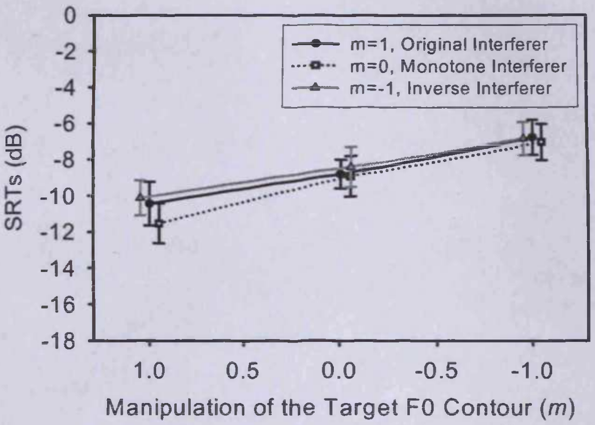

Fig. 2.4: Mean SRT measurements across all listeners for each of the nine conditions in Experiment 2. The three lines represent the manipulations (m=1, 0, -1; original, monotone, inverse) of the F0 contour for the interferer, with the x-axis representing manipulations (m=1, 0, -1; original, monotone, inverse) of the target F0 contour. Error bars represent ± 1 standard error.

There was no effect of varying the F0 contour of the interferer, indicating that when the interferer is speech, it does not matter what manipulation is performed on its F0 contour. The monotone interferer is slightly less disruptive than the other interferers, which is consistent with increased target salience when it alone carries F0 modulation (McAdams, 1989, Culling and Summerfield, 1995). However, this effect is non-significant in this experiment, which indicates that the effect is of little importance for the understanding of running speech.

Analyses of the listener's transcripts showed no errors from intrusions from the interfering speech. Instead, errors were unrelated errors in the target sentences, such as misheard and missing words. The implication of this is that the listeners were better able to follow and interpret the target speech where they had an increased amount of correct F0 information, rather than becoming confused by the content of the interfering speech.

The larger F0 inversion effect observed in this experiment and in the results from Culling et al. (2003) implies that there is something about a speech background which requires the listener to rely more heavily on the F0 contour of the target sentence in order to understand it. Mattys (2004) has also found that different cues are used in different circumstances. He found that coarticulation cues were more prominent than stress cues in quiet, whereas stress cues became more important than coarticulation cues when the speech was presented in a white noise background. Hence, it could be the case that cues that do not normally influence speech intelligibility greatly in certain conditions become more important as the conditions change.

Having established that a robust effect of the F0 contour is observed when the target is manipulated and presented against a target interferer, the next experiment returned to the manipulations performed in the first experiment, where the F0 variation was reduced systematically, hence it included the half and quarter conditions used previously. These conditions were repeated, but this time against interfering speech.

2.4. Experiment 3

2.4.1. Method

2.4.1.1. Listeners

10 paid participants were recruited from Cardiff University Participation Panel. All were normally-hearing native speakers of English. None of the listeners were familiar with the sentences used in the study nor had they taken part in either of the previous experiments.

2.4.1.2. Stimuli

F0 modulations as used in Experiment 1 (see fig. 2.1) were employed for the target speech (m=1, 0.5, 0.25, 0, -1) to represent the five different conditions (original, half, quarter, monotone, inverse). The single-talker interferer remained normally-intonated (m=1) since altering the F0 contour of the interferer did not make a difference in Experiment 2. All other aspects of the stimuli and procedure were identical to Experiment 2.

2.4.2. Results and Discussion

Figure 2.5 shows that there were again larger differences between each of the conditions than in Experiment 1. This confirms the results from Experiment 2 that using speech as the interferer increases the importance of the F0 contour in understanding the target speech.

There was a 1.4 dB difference between the half and the quarter conditions, a further 0.2 dB difference between the quarter and the monotone conditions, and a 0.9 dB difference between the monotone and inverted conditions. A repeated-measures

ANOVA showed a significant main effect of the F0 contour (F (4, 36) =3.722; p<0.05). Post-hoc Tukey HSD tests showed a significant difference (p<0.05) between the original (*m*=1) and inverse condition (*m*=-1), and the half (*m*=0.5) and inverse condition (*m*=-1).

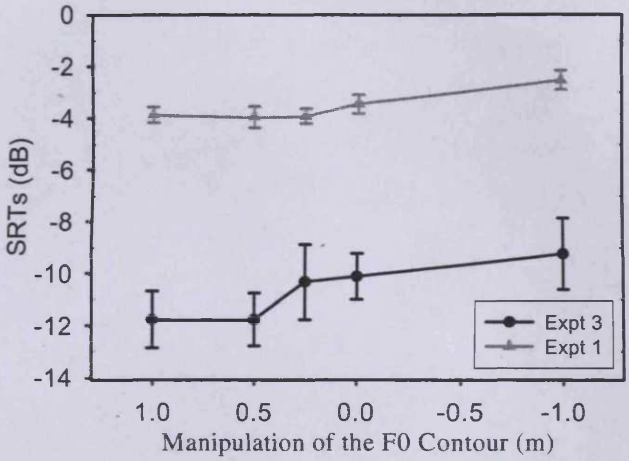

Fig. 2.5: Mean SRT measurements (dB) for the different manipulations (*m*=1, 0.5, 0.25, 0, -1; original, half, quarter, monotone, inverse) of the F0 contour in both Experiments 1 and 3. Error bars represent ± 1 standard error.

The significant effect of F0 modulation implies that it is easier for the listener to separate and understand the target sentence from the single-talker interferer if the target F0 contour follows an appropriate F0 pattern.

2.5. General Discussion

A reduction of F0 modulation was found to have a detrimental effect on speech intelligibility in adverse listening conditions, confirming previous findings (Wingfield et al., 1984; Laures & Weismer, 1999; Assmann, 1999; Laures & Bunton, 2003; Culling et al., 2003). The F0 inversion effect increased with a single-talker interferer masker compared to speech-shaped noise. It is not clear why a speech interferer would accentuate these effects. However, the fact that an inverted F0 contour caused a greater deficit than a monotonous contour implies that it perhaps contains false information which depresses the contour's intelligibility. There are two related ways to account for these effects: the incorrect cuing of content words and the F0 contour shape in general.

2.5.1. Disrupted Cuing of Accented Words

In a normally-intonated sentence, important content words will be accented, tending to be above the average F0 of the sentence. As noted by Cutler and Butterfield (1992) listeners of English assume that strong syllables denote the onset of content words. This factor, along with the content words generally being louder and articulated more slowly contributes to making these words acoustically clearer than the surrounding words (Lehiste, 1970). In a monotonous sentence, none of the words in the sentence are accented and all are at the same F0; therefore there are fewer cues as to the whereabouts of the content words. In a sentence with an inverted F0 contour, however, the accented content words will now be accented in the opposite direction, for instance, a rise will become a fall and vice versa. Also, words which were previously above the average F0 will now be below. Therefore, whereas in a monotonous sentence important content words are not highlighted by their F0, in a sentence with an inverted contour, these F0 cues will be misleading and highlight words which are not particularly important to the meaning of the sentence.

2.5.2. Importance of the Contour Shape

As mentioned above, the focus (accented word) of the sentence is searched for by the subject in order to quickly gain an understanding of the utterance. An F0 contour in the 'right' direction enables the listener to parse the sentence more quickly since it directs the listener to syntactic and phrase boundaries within the utterance (Streeter, 1978; Ladefoged, 1993). The contours surrounding focussed words guide the listener's attention to those particular words (Cutler, 1976; Cutler & Fodor, 1979). Therefore when there is an F0 contour in the 'right' direction, the focussed words may remain focussed, even if only slightly, and the surrounding contour still guides the listener towards those words more readily than in a monotone condition where the variation has been removed. For this reason, the small decrease in F0 variation from the half (m=0.5) to the quarter (m=0.25) conditions in the third experiment reduces the amount of guidance the listener receives to the whereabouts of the content words. Hence, there are two cues present in a contour in the normally-intonated condition, the accent and the surrounding contour guiding the listener to the content words, which are not present in the monotonous condition and are misleading in the inverted condition. The present experiments quantify the contribution that these cues make and show them to be more important than the effect of F0 movement on salience.

CHAPTER 3:

LOW-PASS FILTERING THE F0 CONTOUR

3.1 Introduction

The experiments in the previous chapter have shown that removing variation in the F0 contour, and inverting the F0 contour cause detrimental effects to sentence intelligibility. What has not been shown by these experiments is where the important modulation frequencies for this effect lie. Low-pass filtering the contour removes the high frequency components of the F0 contour whilst retaining the low-frequency components. By progressively removing the higher frequencies, the frequencies most important for intelligibility can be revealed. Low-pass filtering the amplitude modulation spectrum of speech has shown the important frequencies for the amplitude modulation of the speech signal as a whole lie between 1 Hz and 16 Hz (Drullman, Festen & Plomp, 1994a, b; Arai, Pavel, Hermansky & Avendano, 1996; Kanedera, Arai, Hermansky & Pavel, 1999), with the most important frequency being 4 Hz, the syllable rate of speech.

Both of the Drullman et al. studies (1994a, b) required the participants to listen to low-/high-pass filtered sentences in noise and in quiet, and reproduce them as accurately as possible. In the low-pass filter experiment, the modulation spectrum of sentences was low-pass filtered at a range of frequencies from 0 Hz (where no modulations remain) up to 64 Hz. A control condition in which there were no modifications to the amplitude envelope was also included. Sentences low-pass filtered at frequencies less than 1 Hz were presented in quiet due to their already highly-degraded intelligibility, whereas those low-pass filtered at 2 Hz and above were presented in noise. The results showed that speech intelligibility increases progressively with low-pass cut-off frequencies up to a value of 16 Hz. Modulation frequencies above 16 Hz were not found to contribute much to sentence understanding. In the high-pass filter experiments, the sentences were filtered at values ranging from 0 Hz (control condition with all modulations intact) up to 128 Hz. A condition with no modulations left intact, '∞ Hz', was also included. Again, the lowest intelligibility conditions where the sentences had been high-pass filtered at 64 Hz and above were presented in quiet, with the remaining conditions being presented in noise. The intelligibility of the sentences with high-pass filtered envelopes remained at the same level as unprocessed speech up to the 4 Hz cut-off frequency. The implication from this was that envelope modulations below 4 Hz did not aid speech intelligibility as long as higher modulations remain intact.

Arai et al. (1996) found similar results for filtering the modulation spectrum of Japanese syllables. For the low- and high-pass experiments, 12 cut-off frequencies ranging from 0 Hz to '∞ Hz' were used, giving 13 conditions including the clean speech. For the band-pass experiment, there were 8 lower cut-off frequencies ranging from 0 Hz to '∞ Hz', and 8 higher cut-off frequencies within the same range, with 29 conditions in total including clean speech. Subjects were presented with the stimuli and asked to identify each syllable as it was played to them. 16 listeners participated in the low- and high-pass experiments. One of these listeners was then selected to participate in the band-pass filtered experiment. The results showed that speech intelligibility is not severely impaired when the modulation spectrum is high-pass filtered at 1 Hz or low-pass filtered at 24 Hz, and when band-pass filtered between 1 Hz and 16 Hz. The curves representing the low- and high-pass filtered results intersect between 3 and 4 Hz, around the syllable rate of speech, much as in the Drullman et al. studies.

The relative importance of various amplitude modulation frequencies in both quiet and noise have also been studied in ASR systems (Kanedera et al., 1999). In this study, the time-trajectories of the spectral envelopes of 216 phonemically-balanced Japanese words produced by five native male speakers were band-pass filtered and the corresponding recognition accuracies measured for the ASR system in quiet. The results showed again that the most useful linguistic information lay in modulation frequencies between 1 and 16 Hz, with the dominant frequency being 4 Hz. For the speech-in-noise experiments, the ASR system was trained on clean English and Japanese speech. The test data was then degraded by spectral filtering and additive noise. In this case, the range below 2 Hz and above 8 Hz was found to be less important to speech intelligibility than when the speech had been presented in quiet. The range below 1 Hz was even found to significantly degrade the recognition accuracy in some noisy environments.

These experiments seem to indicate that the syllable rate of speech (4 Hz) is an important cue in speech perception. Greenberg (1996) argues that the syllable is the primary unit of speech perception due to that fact that most co-articulation effects occur within a syllable, hence it is easier to recognise the syllable as a whole rather than individual co-articulated phonemes. Cutler (1997) argues that, although the syllable is a highly important unit in French and other syllable-based languages, in English, it is a less important unit. Indeed, Norris and Cutler (1988) have shown that the phoneme is an important unit for speech perception in English since they argue that if listeners were processing speech syllable-by-syllable, they would need to hear the entire syllable before responding to a syllable target. Norris and Cutler measured reaction times for listeners on both phonemes and syllables. They used 'foils' to force listeners to analyse the whole target rather than respond prematurely. Foils are items which are distinct from the target in only one aspect. For example, the phonemes /d/ and /t/ differ only in voicing (+/-) and therefore /t/ could act as a foil for the phonemic target /d/. Syllabic examples of targets and foil pairs are /pid/ and /pit/, or /bid/ and /pid/. Norris and Cutler argued that if syllables were perceived as units rather than a set of phonemes, listeners should have the same reaction time to both syllables with phoneme initial foils and phoneme final foils. If however, they relied on a phonemic analysis, then the phoneme initial foils should behave in a similar way to the actual phoneme target foils and only syllables with phoneme final foils should force a complete analysis of the syllable. Listeners responded prematurely to the syllable targets, indicating that they were processing the speech using a unit smaller than the syllable, such as the phoneme.

There is much controversy over what is the basic unit of speech perception, with the majority of studies favouring the syllable. In terms of the F0 contour, it could be that both the syllable and phoneme play an important role. Given that English is a stress-timed language (Cruttenden, 1986) involving rhythmic alternations of stressed and unstressed syllables, the F0 contour itself is based around the accents placed on these syllables. However, there are smaller variations that are influenced by the phonetic detail of the speech which influence the shape of the contour. For instance, when voicing restarts after a voiceless consonant, the F0 value tends to be a little higher than it would be after a voiced consonant. Also, some vowels have intrinsically higher pitch than others due to muscular interactions between the articulators and the larynx. Since the shape of the F0 contour is highly dependent on the accents placed on the syllables, it is possible that the modulation frequencies around the syllable rate of speech would be important to its intelligibility. On the other hand, it could be that the smaller fluctuations around the phonetic content of the speech contribute to the effect. The following experiment will test this prediction through low-pass filtering the F0 contour of the speech.

3.2. Experiment 4

3.2.1. Method

3.2.1.1. Listeners

20 paid participants were recruited from Cardiff University Participation Panel. All were normally-hearing native speakers of English. None of the listeners were familiar with the sentences used in this study nor had they taken part in Experiments 1-3.

3.2.1.2. Stimuli

In order to low-pass filter an F0 contour, it needs to be an uninterrupted waveform. For the F0 contour to be uninterrupted, the speech used needs to be continuously voiced. For this reason, 100 continuously voiced sentences were created. The words were selected from the MRC psycholinguistic database to contain only vowels, nasals, liquids and voiced fricatives, and on the basis of their Kucera and Francis (1982) frequency being under 1000, in order to ensure that the sentences were, like the Harvard IEEE sentences, of low-predictability. The sentences were written in the style of the IEEE sentences in that they each contained five keywords by which the listeners would rate their responses. Due to the limitations on the sentence construction for these stimuli (that they need to be continuously voiced) the sentences tended to be less contextually coherent than the Harvard sentences. The fewer contextual cues present, the harder it is for the listener to follow the sentences (Boothroyd & Nittrouer, 1988), hence SRT values were expected to be higher for these stimuli, although this was not tested. The list of continuously-voiced sentences can be found in Appendix 1.

The sentences were then recorded by a British female speaker using a Sennheiser K6 microphone with an ME62 omnidirectional capsule. The signals were conditioned and digitised using Tucker Davis Technologies equipment (an MA1 microphone amplifier and a DD1 digital-to-analogue converter) at 20 kHz with 16-bit quantization. The F0 contours of the sentences were extracted and low-pass filtered at cut-off frequencies of 1, 2 and 4 Hz (see Figure 3.1), and, if required, inverted, before resynthesis using the Praat PSOLA speech analysis and resynthesis package as in the previous experiments.

The formula below was applied before resynthesis, setting all the sentences at an average F0 of 210 Hz.

$$F0' = \left[210 \exp\left(m \ln(F0 / \overline{F0})\right)\right] \tag{4}$$

In this experiment *m* was set to either 1 or -1, so that both the original and inverse conditions could be compared when low-pass filtered.

Both the syllable rate and stressed syllable rate of the continuously-voiced stimuli were measured. The average syllable rate was 5.9 Hz, which is higher than the 4 Hz figure cited in previous studies. The increased syllable rate could be due to the nature of the continuously-voice sentences. The average stressed syllable rate was 2.9 Hz.

Speech interferers were used once again and were created using ten sentences from the Harvard IEEE corpus. The recorded voice used was DA. The formula below was applied to the interferers, setting their average F0 at 125 Hz, hence allowing a 9 semitone difference between the target and interferer contours, as in both Experiments 2 and 3.

$$F0' = \left[125 \exp(m \ln(F0 / \overline{F0}) \right] \qquad (5)$$

The variable *m* was set to 1 to allow for a normally-intonated contour.

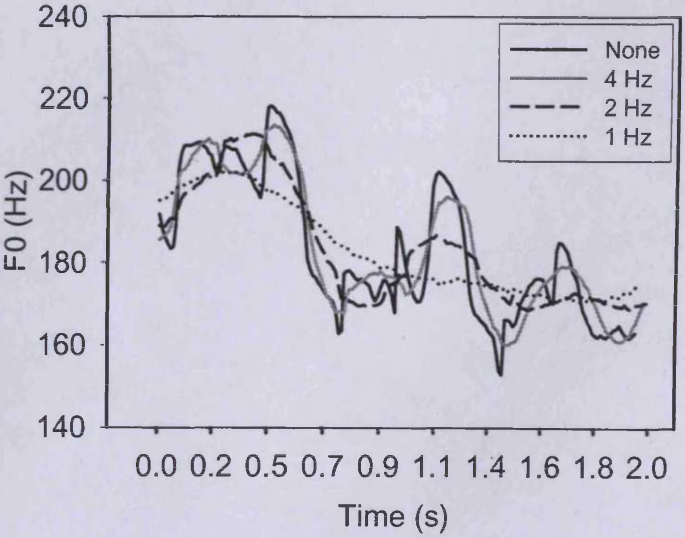

Fig. 3.1: Low-pass filtered F0 contours compared to an un-filtered F0 contour for the same continuously voiced sentence from Experiment 4. The sentence is 'The raven rose over the rim of the ravine'. The contour is filtered at 1, 2 and 4 Hz.

Both normally-intonated (*m*=1) and inverse (*m*=-1) conditions were compared for the target speech, in each of the five low-pass-filtering conditions (0, 1, 2, 4 Hz and unfiltered), giving 10 conditions in total. Low-pass filtering the speech at 0 Hz should produce the same result for both *m*=1 and -1 because it removes all F0 variation from the utterances. These two conditions were included as controls to ensure that the effect of F0 inversion found in previous experiments was not due to some artefact of the experimental manipulation. The same procedure was followed as in Experiments 2 and 3.

3.2.2. Results and Discussion

An informal comparison between the continuously-voiced sentences and stimuli from previous experiments indicates that SRTs were higher for the continuously-voiced sentences. This is potentially due to the reduced number of contextual cues in the continuously-voiced utterances. However, the use of a different talker in this experiment limits the conclusions that can be drawn concerning the intelligibility of continuously-voiced sentences.

Figure 3.2 shows that low-pass filtering detrimentally affected the intelligibility of the normally-intonated condition (*m*=1). The most important frequencies for the normally-intonated contour seem to lie between 2 and 4 Hz, with a 2.5 dB difference in SRT between these two conditions. However, enough information is retained in the 1 Hz filtered condition to improve the intelligibility of the utterance relative to the monotone condition, with a further 2 dB difference between these conditions.

The detrimental effect of inverting the F0 contour detailed in experiments 1-3 was replicated in this experiment (F(1, 19)=51.063, p<0.001), supporting the conclusions drawn that having the correct F0 cues on the appropriate syllables aids speech intelligibility against an interfering voice.

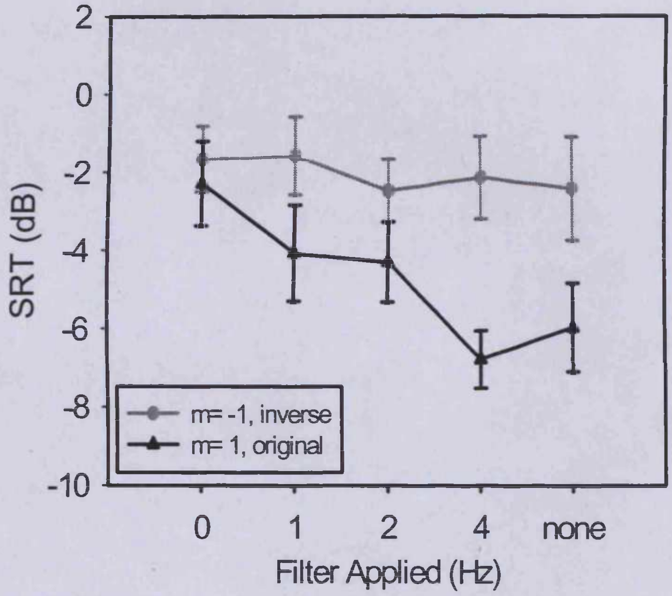

Fig. 3.2: Mean SRT measurements (dB) across all listeners in Experiment 4. The black line corresponds to results for the manipulated normally-intonated (*m*=1) sentences, whereas the grey line represents results for the sentences with an inverted F0 contour (*m*=-1). The low-pass filtering applied to the sentences (1, 2 and 4 Hz), along with monotone (*m*=0, labelled '0 Hz') and original (*m*=1, labelled 'none') F0 manipulations are on the x-axis. Error bars represent ± 1 standard error.

Low-pass filtering the F0 contour was found to degrade the intelligibility of the speech (F(4, 76)=4.740, p<0.005). Figure 3.2 shows that low-pass filtering the F0 contour produced a greater increase in SRT for the original F0 contour than for the inverse F0 contour. A significant within-subjects linear contrast was noted between the F0 manipulation and low-pass filtering conditions (F(1, 19)=4.609, p<0.05). This contrast was caused by low-pass filtering detrimentally affecting the normally-intonated contour (Wilks' Lambda=0.75, F=13.467, p<0.001), but not having a significant effect on the inverse F0 contour (Wilks' Lambda=0.232, F=1.337, p>0.05).

The strongest effect of filtering the F0 contour was found when the frequencies between 2 and 4 Hz were removed. This is below the syllable rate, 5.9 Hz, of the stimuli but encompasses the stressed syllable rate, 2.9 Hz. As mentioned above, the stress placed on syllables is actively searched for by listeners in order to segment speech and pinpoint the key words. Indeed, deliberately mis-stressing syllables within English words has been reported to decrease response time in comparison to correctly stressed words (Cutler. & Clifton, 1984). Hence these results reinforce the idea that the clarity of the accents placed on the syllables is important for the listener and show that the search for these accents by the listener makes a measurable difference to the intelligibility of the speech, since as the prominence of the accent stress is reduced, the intelligibility of the speech is also reduced.

In reference to the synthesis of speech intonation, the syllable is a highly important unit. Generally F0 synthesis models operate in two stages. The first stage involves generating an abstract description of the intonation contour, such as indicating where pitch accents and boundary tones lie. The second stage is where this abstract description is converted into a sequence of F0 values. In the ToBI model (Silverman et al., 1992) individual syllables in the speech are tagged with labels indicating the type of accent that needs to be assigned, thus describing intonation in terms of pitch accents and boundary tones. Likewise the Tilt model (Taylor, 2000) is made up of intonational events which correspond to the pitch accents and boundary tones. Every event is linked to a syllable. Thus, speech synthesis models reflect the importance of the syllable in the F0 contour for both natural and synthetic speech.

By reducing the amount of information in the inverse contour, it was expected that the intelligibility of the utterance would improve since less misleading pitch information would be supplied to the listener. This prediction followed from the previous three experiments where the inverted F0 contour caused a greater detriment to speech intelligibility than a monotone F0 contour. As the inverted F0 contour is increasingly low-pass filtered, its shape tends towards a flattened or monotone F0 contour. Previous results would infer that its effect on speech intelligibility would therefore decrease. However, this did not happen.

An interesting point to note is that the difference between the monotone and original conditions was larger with the continuously voiced sentences than in the previous experiment. On the other hand, the difference between the monotone and inverse conditions is no longer significant. There are two ways to interpret this result. The effect of the inverse F0 contour may have disappeared using continuously voiced materials. It could be that part of the inversion effect is mediated by stop closures; the presence of onset and offset cues may serve to confuse the listener further when presented with the incorrect F0 cues in inverse contour, and hence once these cues are removed, the inverse contour causes no greater effect than the monotonous contour. Alternatively, the effect of inversion may be masked by some interaction between combined use of continuous voicing and complete monotonisation of the F0 contour, which elevates thresholds in this particular condition. The latter suggestion is supported by the fact that the difference between the inverse and normally-intonated contours remained the same as in previous experiments (at around 4 dB). It could be that the lack of onset and offset cues from stop closures combined with the lack of variation in the F0 contour in the monotone condition is particularly detrimental to understanding and yields SRTs that are similar to those produced by inversion.

It is clear that the continuously voiced sentences are harder to follow than the previous sentences used, with a 4 dB difference in SRTs between the original condition for both Experiments 3 and 4, and this effect has been noted in previous research (Bird & Darwin, 1998; Stubbs & Summerfield, 1990). However, it is unclear as to why the difference between the monotone and inverse conditions decreased when using these different materials. It is interesting to note that in the Hillenbrand (2003) experiment, where synthesised speech was used, a small difference was found between the monotone and inverse conditions only when the stimuli had been low-pass filtered to impair their intelligibility, suggesting that the choice of materials may influence observation of the effect.

CHAPTER 4:

COMPARISON OF NATURAL AND SYNTHESISED SPEECH

4.1. Introduction

Results from experiments in the two previous chapters have shown a detrimental effect of both monotonising and inverting the F0 contour of natural speech. Inversion has a more harmful effect on speech intelligibility than monotonisation, indicating that having incorrect F0 cues affects the listener's ability to hear the speech accurately more than having no F0 variation within the utterance. This finding prompted questions about the importance of generating accurate F0 contours for use in speech synthesis.

Modelling the prosodic contour, including intensity, timing and F0, correctly is an important issue for speech synthesis systems which aim to recreate natural sounding speech that is acceptable to human listeners. Prosody is an important aspect of natural and intelligible speech. As mentioned in the introduction, intensity is mainly determined by phone identity, although it does vary with stress. Intensity variations tend to be less influential on prosody than timing patterns or fundamental frequency (Fry, 1958; Morton & Jassem, 1965; Streeter, 1978). In speech synthesis, timing patterns are created through allocating a particular duration to each speech segment determined by a set of rules based on phonetic knowledge (Holmes & Holmes, 2001). There are a number of rules that govern the duration of speech segments. Synthesis systems work by applying these rules to each segment of speech to determine the length of these segments. F0 synthesis models tend to work in two stages (Holmes and Holmes, 2001). The first stage creates an abstract description of the F0 contour from a set of rules used within the model. The second stage converts this abstract contour into actual F0 values. Experiments in this chapter will focus on the timing patterns and F0 of both synthesised and natural speech in order to see how successful synthesis systems are at modelling prosody using speech reception thresholds as a metric of synthesis quality.

A multi-lingual speech synthesis system, Festival (CSTR), developed at the University of Edinburgh will be used to create the synthesised speech. This is a diphone-based concatenative speech-synthesis system. Intonation and duration modules are included within the system and can be manipulated by the user to adjust the synthesis output. A number of voices are also provided with the system. The following experiments used two of the male American voices. In Festival, intonation for these voices is predicted by a ToBI-like system using decision trees to predict accent and end tone positions. Actual F0 values are predicted at three points in each syllable using linear regression trained from the Boston University FM radio database, which is then mapped onto the speaker's pitch range. Rules specifying the duration patterns for these voices are also calculated using the Boston University FM radio database.

The first experiment described here directly compares the effect on SRT of manipulating the F0 contours of natural and synthesised speech. In the past, tests of F0 synthesis quality have concentrated on the fidelity with which synthetic F0 contours mimic natural ones, either objectively or subjectively. Since naturally-produced F0 contours may differ from one speaker to another, the definition of the optimal contour cannot be precisely defined. The subjective methods therefore centre on acceptability or perceptual similarity to a natural model, while the objective ones test how accurately a naturally generated F0 contour can be reproduced. The question therefore arises as to how the two types of measurement are related.

Clark and Dusterhoff (1998) contrasted three objective methods of measuring the similarity of two fundamental frequency contours with a perceptual test. In the perceptual task, listeners were asked to rate the similarity of the intonation patterns between pairs of utterances on a five-point scale. The first objective method involved calculating F0 differences using Root Mean Squared Error (RMSE), which measured the distance between two F0 contours on the time axis. Correlation coefficients were then used to discover how closely the synthetic F0 contour followed the direction of the original F0 contour. The second method, the tangential estimation method, treats one contour as the reference contour and measures the perpendicular distance from this contour to the contour being compared to it. Finally, the warping method measures the distance between contour points of the two F0 contours within corresponding pitch events. Results showed that listeners are able to distinguish between natural and synthetic contours, but not able to tell which synthetic contours are closer to the natural one, despite objective methods registering some differences. This is most likely due to the small improvements shown by objective methods being masked by the overall inaccuracy of the synthetic F0 contour. Therefore it seems that although objectively the F0 contour may seem more accurate, perceptually larger improvements need to be made to provide a noticeable difference. None of these tests, however, determine the relative intelligibility of these F0 improvements. Given our previous results, it is possible that although the listeners may be unable to pinpoint differences between the synthesised and natural F0 contours, the intelligibility of the utterances may be different.

This experiment will determine whether the F0 contour of the synthetic speech contributes to speech intelligibility to the same extent as natural F0 contours through comparing the effect of F0 inversion for both types of speech. Inverting the F0 contour detrimentally affects the intelligibility of natural speech. If synthetic F0 contours aid speech intelligibility in the same manner as the F0 contours of natural speech, then a similar effect of inversion should be seen with the synthesised speech. If, however, inverting the F0 contour does not cause a detrimental effect on the intelligibility of the synthesised speech then it can be assumed that the F0 models do not accurately capture the contribution to intelligibility from natural speech F0 contours.

4.2. Experiment 5

4.2.1. Method

4.2.1.1. Listeners

24 paid participants were recruited from Cardiff University Participation Panel. All were normally-hearing native speakers of English. None of the listeners were familiar with the sentences used in this study nor had they taken part in Experiments 1-4. 12 of the participants completed the task with synthetic speech targets and 12 had normal speech targets.

4.2.1.2. Stimuli

60 sentences from the Harvard IEEE corpus (Rothauser et al., 1969) were selected for this experiment. Male voice, CW, was used as the target in the natural speech conditions. The 60 synthesised speech sentences were produced using the Festival Speech Synthesis System, version 1.4.3 (Black, Taylor & Caley, 2003). Male voice 'Kal', provided with the software, was used for the synthetic target speech. All speech was manipulated using Praat PSOLA, as in the previous experiments. The following formula was applied to the F0 contour of each target sentence, setting the geometric mean F0 to 110 Hz.

$$F0' = \left[110 \exp(m \ln(F0 / \overline{F0}))\right] \qquad (6)$$

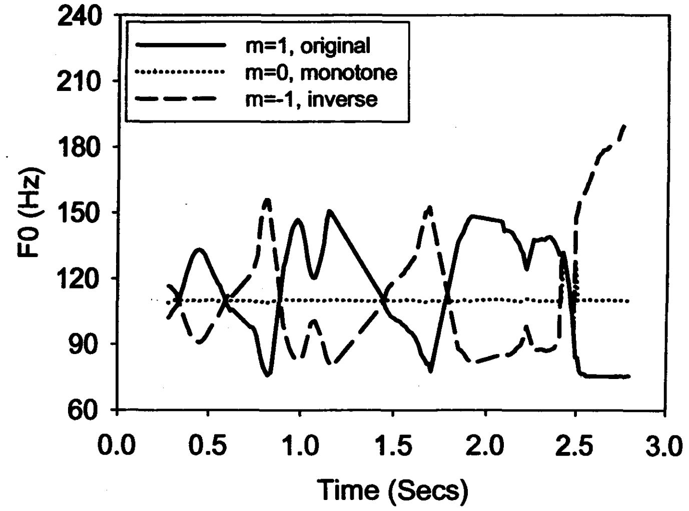

Fig. 4.1: Manipulations (*m*) of the F0 contour for an example of the natural target speech for Experiment 5. The sentence corresponding to the F0 contour is "The girl at the booth sold fifty bonds". Manipulations are *m*=1, 0 and -1, corresponding to the three conditions: original, monotone and inverse, respectively.

In this experiment, *m* was set to 1, 0 or -1, corresponding to normally-intonated, monotonous and inverted F0 conditions. See figures 4.1 and 4.2 for comparison of the F0 contour for both natural and synthetic speech for the same sentence.

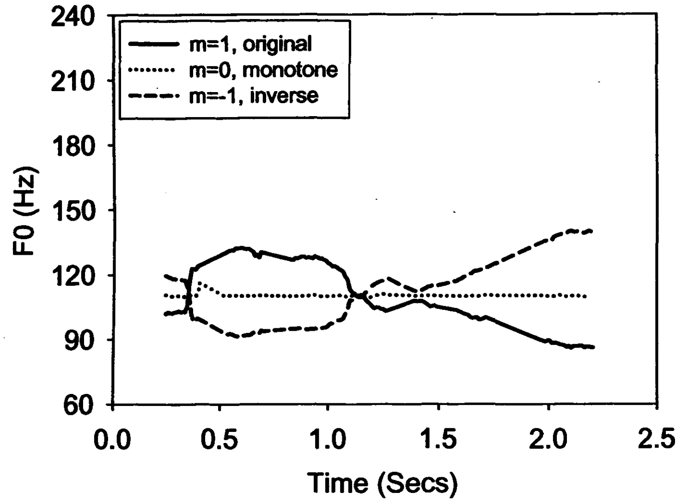

Fig. 4.2: Manipulations (*m*) of the F0 contour for an example of the synthetic target speech for Experiment 5. The sentence corresponding to the F0 contour is, as in figure 4.1, "The girl at the booth sold fifty bonds". Manipulations are *m*=1, 0 and -1, corresponding to the three conditions: original, monotone and inverse, respectively.

Interfering speech was once again set at 9 semitones apart from the target speech to reduce the chance of overlap between the target and interferer F0 contour. Six interferer sentences were chosen from the Harvard IEEE corpus. Voice DA was used for the six natural speech interferers. The six synthesised interferers were produced using male voice 'Ked', also provided with the Festival Speech Synthesis software. The following formula was applied to the interferers, setting their geometric mean F0 to 210 Hz.

$$F0' = \left[210 \exp\left(m \ln(F0 / \overline{F0})\right)\right] \tag{7}$$

The variable *m* was set to 1 to allow for a normally-intonated F0 contour.

4.2.1.3. Procedure

The same procedure was followed as in previous experiments, except that this time half the participants were presented with synthetic speech targets and half with natural speech targets. In pilot studies, where listeners were presented with both synthetic and natural speech targets in the same block of trials, the consistent F0 inversion effect found in previous experiments was not replicated. It seemed that the large difference in intelligibility between the synthetic and natural target speech affected the results of F0 manipulation. Listeners simply found natural speech much more intelligible in comparison to the synthesised speech. For this reason, it was felt best to manipulate the target speech between subjects.

Normally-intonated (m=1), monotonous (m=0) and inverted (m=-1) F0 targets were compared against both types of interferer (synthetic, natural), giving 6 conditions within-subjects (3 F0 conditions x 2 interferer types). The two types of target speech (synthetic, natural) were manipulated between subjects. Each session began with two practice SRT measurements with normally-intonated sentences being played against a single-talker interferer. One practice used natural speech as the target, and one used synthetic speech. Both practices used a natural speech interferer.

4.2.2. Results and Discussion

A repeated measures ANOVA showed that the overall effect of the type of interferer was not significant (F(1,22)=3.478; p>0.05). A large difference in SRTs (9-15 dB depending on the F0 manipulation, see fig. 4.3) between the synthetic and natural speech targets was found to be significant (F(1, 22)= 136.448; p<0.00001).

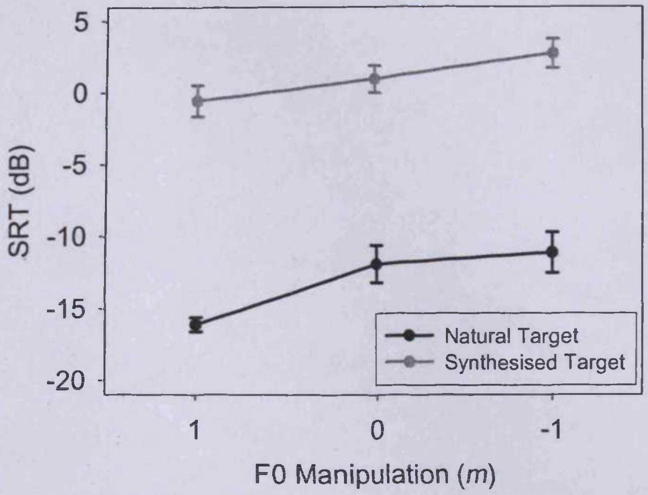

Fig. 4.3: Mean SRT measurements (dB) averaged across all listeners and both types of interferer in the two target conditions for the different manipulations of the F0 contour (*m*) in Experiment 5. Manipulations *m*=1, 0 and -1 correspond to conditions original, monotone and inverse. Error bars represent ±1 standard error.

The significant effect of manipulating F0 contour found in previous experiments was replicated (F(2, 44)=13.300; p<0.0001). A test of Wilks' Lambda showed the effect of F0 manipulations to be present for both natural target speech (F=13.192; p<0.0001) and for synthetic target speech (F=5.186; p<0.02). For the natural target speech, a significant effect was shown through pairwise comparisons with Bonferroni corrections between the original and monotone condition (p<0.01) and the original and inverted conditions (p<0.0001). A significant difference was also seen between the normally-intonated and inverted condition for the synthetic target speech (p<0.02), although this was not as strong as with the natural target speech. Averaged across the two interferers, the difference between the original and inverted F0 contours was smaller for synthesised speech (3.3 dB) than for natural speech (5.0 dB), but the interaction between the F0 manipulation and type of target speech was not significant (F(2,44)=16.176, p=0.25). Therefore, the effect of F0 inversion for synthesised speech shows that synthetic F0 contours do aid speech intelligibility. However, the trend shown by the inversion effect being numerically smaller with synthesised speech suggests that the synthetic F0 contour may contribute less to intelligibility than natural F0 contours.

The interaction between the interferer and type of target speech was also found to be just non-significant (F(1, 22)=3.789; p=0.06). Figure 4.4 illustrates this near-significant interaction. For natural target speech, a 3 dB difference in SRTs was noted between the synthesised and natural interferers. Only a 0.07 dB different was found between the two types of interferer for synthesised target speech.

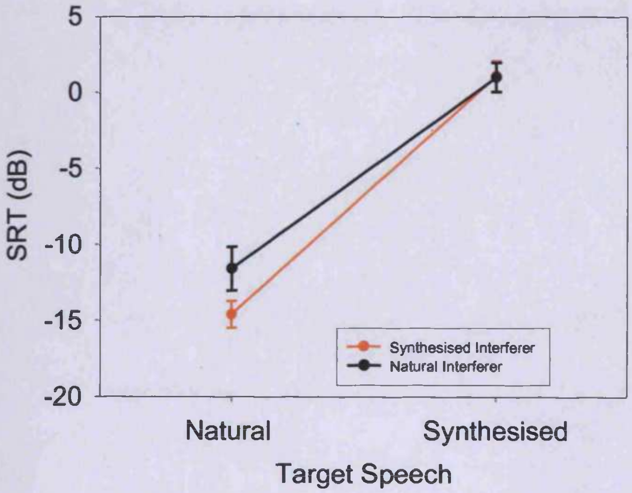

Fig. 4.4: Mean SRT measurements (dB) across listeners for both natural and synthesised target speech for both types of interferer, natural and synthesised. Error bars show ±1 standard error.

The large difference in intelligibility between the natural and synthesised speech indicates that there is something highly unnatural about the synthesised speech that is causing such a detrimental effect. However, it is possible that a confound of speech rate existed in this experiment. Synthesised speech tended to be faster than natural speech. Hence it could be that the increased rate of the synthetic speech in comparison to natural speech contributed to the decreased intelligibility of the synthetic speech.

The following experiment investigates the intelligibility of the synthesised and natural F0 contours further. Switching the F0 contours from the synthesised speech onto the natural speech will enable us to see if the synthetic F0 contour has a detrimental effect on the intelligibility of the natural speech. Likewise, placing the F0 contours from natural speech onto synthesised speech will show if the natural F0 contour aids synthesised speech intelligibility.

4.3. Experiment 6

A previous study conducted by AT&T Research Labs (Jilka, Syrdal, Conkie & Kapilow, 2003) has looked into placing natural F0 contours onto synthesised speech. Five test sentences were chosen which were representative of the four main discourse situations: declarative ('Oak is the type of wood Dennis likes best.'), continuation rise ('It snowed, rained and hailed the same morning.'), wh-question ('How may I help you?') and yes/no questions ('Do you want to make a collect call?'). Prosodic patterns for these sentences produced by six native speakers of four varieties of English (US English, UK English, Australian English, Indian-accented English) were placed on a synthetic US English voice. Listeners rated the utterances on a 5-point scale for quality. The experiment found that natural prosodic variations are not particularly influential in the perceived quality of synthesised speech. This is potentially due to poor phonetic quality of the speech, hindering any perceived improvement to the speech quality from more natural prosodic information. As mentioned earlier, the listener's perception of the speech quality does not necessarily reflect the speech intelligibility, which is to be investigated here.

As discussed in Chapter 1, in terms of prosody, both F0 and durational cues tend to outweigh intensity cues (Fry, 1958; Morton & Jassem, 1965; Streeter, 1978). Duration cues have been found to be particularly important for indicating word boundaries (Nakatani et al., 1978) and prosodic boundaries (Streeter, 1978). For this reason, these experiments will also compare the synthetic durational contours to those of natural speech in order to determine any intelligibility differences between the two.

SRT measurements were taken for the synthesised speech with natural F0 and timing contours in order to give some indication of the effect on speech intelligibility of replacing the contours. Synthesised contours will also be placed onto natural speech to investigate whether speech intelligibility deteriorates. If it is the case that the synthesised speech does not show any improvement from the natural contours, but that the natural speech is degraded by the synthesised contours, the suggestion that the phonetic quality of the synthesised speech is too poor to utilise the natural prosodic cues would be supported.

4.3.1. Method

4.3.1.1. Listeners

24 paid participants were recruited from Cardiff University Participation Panel. All were normally-hearing native speakers of English. None of the listeners were familiar with the sentences used in this study nor had they taken part in Experiments 1-5. As

in the previous study, 12 of the participants completed the task with synthetic-speech targets and 12 had natural-speech targets.

4.3.1.2. Design

The type of target speech (natural, synthesised) was manipulated between-subjects. Type of F0 contour, type of duration contour and F0 manipulation were within-subjects factors. Thus, within-subjects it was a 3x2x2 design: F0 manipulation (*m*=1, 0, -1: original, monotonous, inverse), F0 contour (natural, synthesised), duration contour (natural, synthesised).

4.3.1.3. Stimuli

120 sentences from the Harvard IEEE corpus (Rothauser et al., 1969) were selected for this experiment. As in the previous experiment, male voice CW was used as the target in the natural speech conditions and the synthesised speech was produced using the Festival Speech Synthesis System's voice Kal. All speech was manipulated using Praat PSOLA.

Prior to any manipulation of the F0 contour, synthetic speech was dynamically time warped to match the duration contour of the natural speech, and vice versa. Dynamic time warping matches two utterances such that each segment within each utterance is time aligned to its corresponding section of speech in the other utterance. Thus, one sentence spoken by two different people can be time aligned and compared directly (see fig. 4.5 for an example).

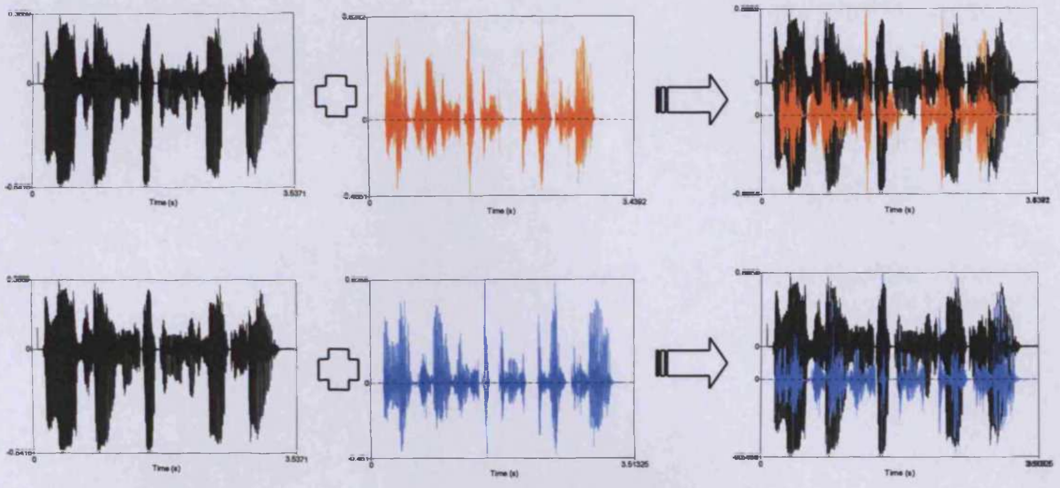

Fig. 4.5: Spectrograms of two utterances of the same sentence 'A large size in stockings is hard to sell'. The top set of spectrograms depicts the utterances before any time warping has taken place. The bottom set displays the two utterances once the second has been time-aligned to the first utterance.

The synthetic speech, on the whole, tended to be faster than the natural speech. This meant that when the synthetic speech was time-aligned to the natural speech, it tended to be lengthened, whereas natural speech was shortened when manipulated to match the synthetic speech. Time-aligned speech represented one condition for this experiment. This condition enabled us to test whether the duration contours of each type of target speech had an effect on the intelligibility. Thus, if the contour from the synthetic speech was less intelligible than that of natural speech, by time-aligning the natural speech to match the synthetic speech, the intelligibility of this speech should be reduced. The total duration of the two types of speech when they were not time-aligned was not equalised, however, which potentially left a confounding variable of speech rate in this condition. This was taken into account when considering the results.

Once the speech was time-aligned, the F0 contours of both types of target could be swapped. In one condition, the F0 contour was the only change made to the target. In another, both the F0 contour and the duration contour of the target speech were from the synthetic target speech. In the condition where the F0 contour was the sole change, the speech from which the F0 contour was being taken was time- aligned to match the speech that the contour was being placed onto. This ensured that the temporal pattern of the target speech remained the same. In the condition where both the F0 contour and the timing of the target speech were replaced, the target speech was time-aligned to the speech giving the contours. See figures 4.6 and 4.7 to compare the F0 contours for all four manipulations for both natural and synthetic speech.

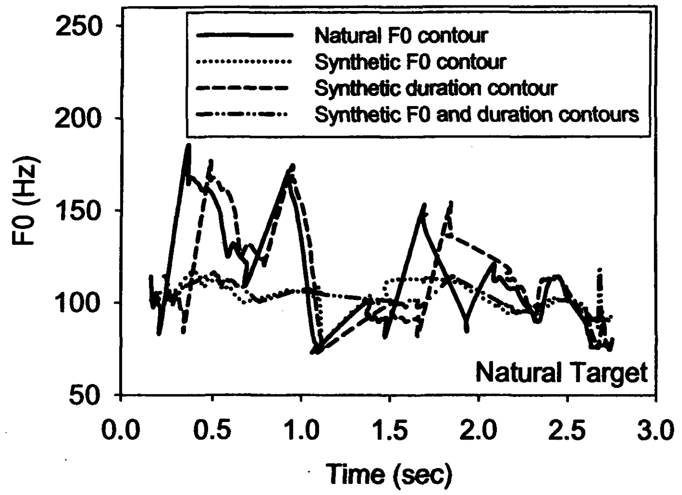

Fig. 4.6: F0 contours for an example sentence ('The fin was sharp and cut the clear water.') for natural speech with F0 and duration contours from synthesised speech. Speech with synthetic duration contours has been dynamically time warped to match the temporal dynamics of the synthesised speech.

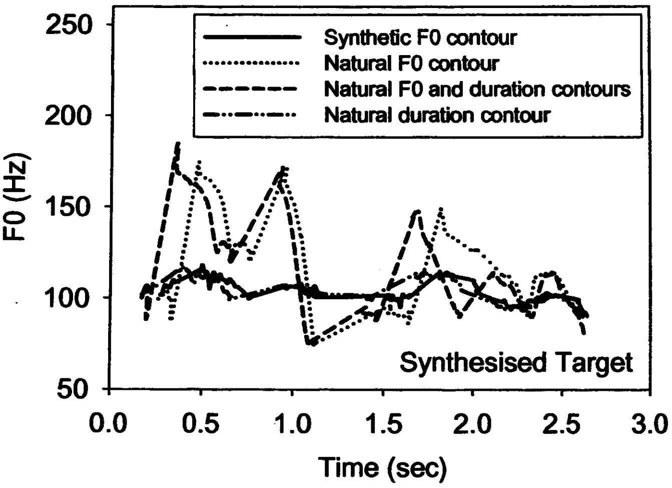

Fig. 4.7: F0 contours for the same example sentence ('The fin was sharp and cut the clear water.') for synthesised speech with F0 and duration contours from natural speech. Speech with natural duration contours has been dynamically time warped to match the temporal dynamics of the natural speech.

The F0 contour in each condition produced within this 2x2 design: (F0 contour type (original, switched) x duration contour type (original, switched), was then manipulated again within Praat. The following formula was applied to the F0 contour of each target sentence, setting the geometric mean F0 to 110 Hz.

$$F0' = \left[110 \exp(m \ln(F0 / \overline{F0})\right] \quad (8)$$

In this experiment, m was set to 1, 0 or -1, corresponding to normally-intonated, monotonous and inverted F0 conditions. Hence, there were 12 conditions in total in this experiment with the 4 variations of the time/F0 contours and the 3 F0 manipulations.

Given the lack of effect of interferer type in the previous experiment, only natural speech interferers were used in this experiment. Interfering speech was once again set at 9 semitones apart from the target speech to reduce the chance of overlap between the target and interferer F0 contour. Voice DA was used for the twelve speech interferers. The following formula was applied to the interferers, setting the geometric mean F0 to 210 Hz.

$$F0' = \left[210 \exp\left(m \ln(F0 / \overline{F0})\right)\right] \qquad (9)$$

The variable *m* was set to 1 to allow for a normally-intonated F0 contour.

4.3.1.4. Procedure

The same procedure was followed as in Experiment 5, with half the participants being presented manipulated synthetic speech and half manipulated natural speech.

4.3.2. Results and Discussion

A mixed-ANOVA was run, with the type of target speech (synthesised/natural) as the between subjects factor and a 3x2x2 within-subjects design (F0 manipulation x F0 contour type x duration contour type).

Synthesised speech was found to be significantly less intelligible than natural speech (F(1,22)=65.813, p<0.0001), supporting the results from Experiment 5. A main effect of manipulating the F0 contour was also replicated in this experiment (F(2, 44)=11.265, p<0.0001). Pairwise comparisons showed that there was a significant difference between original (*m*=1) and inverse (*m*=–1) conditions (p<0.005), with the inverted F0 contour being less intelligible than the original contour. Thus, for both synthesised and natural target speech, an inverted F0 contour negatively impacts speech intelligibility.

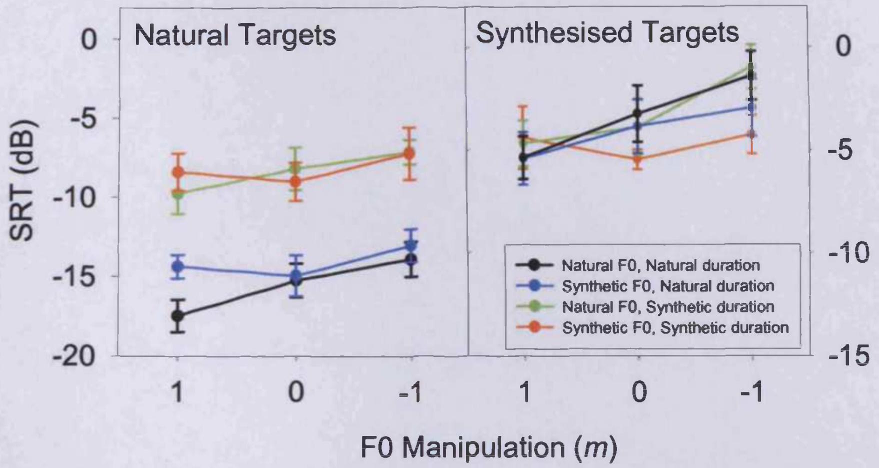

Fig. 4.8: Mean SRT measurements (dB) across all listeners for the different manipulations of the F0 contour (*m*) for both the synthesised and natural target speech. Manipulations *m*=1, 0 and -1 correspond to conditions original, monotone and inverse. Error bars represent ± 1 standard error.

The synthesised duration contour was found to significantly increase SRTs compared to the natural duration contour (F(1, 22)=54.038, p<0.0001). A significant interaction between the duration contour type and target speech (F(1, 22)=62.588, p<0.0001) reflected the result that the natural duration contour produced lower SRTs than the synthesised duration contour for natural speech (p<0.0001), although no significant improvement in SRT occurred when a natural duration contour was placed on synthesised speech. This result argues against the confound of speech rate discussed earlier. It was thought that the increased speech rate of the synthesised speech could cause the increase in SRT levels for the natural speech. However, since replacing the synthetic duration contour with a natural duration contour on the synthetic target speech did not alter SRTs, it would seem that decreasing the synthetic speech rate with the natural duration contour does not impact its intelligibility. Therefore, it could be argued that the difference in speech rate between the two types of target speech did not greatly impact the speech intelligibility. Although the exact cause of this effect from switching the duration contours is not clear, it seems that the phonetic quality of the synthesised speech is too poor to benefit from either the slower speech rates or the duration contour of the natural speech.

Unlike synthesised duration contours, synthetic F0 contours did not significantly affect SRTs compared to natural F0 contours ($F(1, 22)=0.257$, $p=0.62$). Thus, synthetic F0 contours do not significantly degrade the intelligibility of the target speech. However, there was an interaction between the target speech and F0 contour type ($F(1, 22)=7.940$, $p<0.05$). Pairwise comparisons showed this to be due to a significantly greater effect of F0 manipulation for the natural F0 contour than for the synthetic F0 contour (see fig. 4.9). The natural F0 contour generated higher SRTs in both the monotonous and inverted conditions than the synthesised F0 contour.

A significant linear contrast between the F0 manipulation and the type of F0 contour ($F(1, 22)=6.094$, $p<0.05$) highlights the difference between the natural and synthetic F0 contours further. Pairwise comparisons showed a significant effect of F0 inversion for the natural F0 contour ($p<0.0001$), but no significant effect for the synthetic F0 contour ($p>0.05$). For natural target speech, the difference between the original ($m=1$) and inverse ($m=-1$) conditions decreased from 3.6 dB with the natural F0 contour to

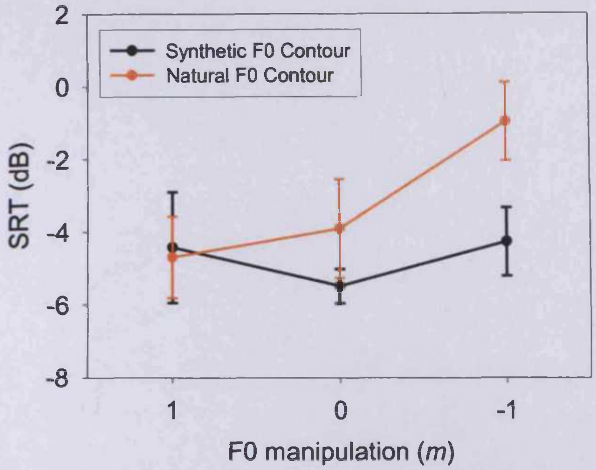

Fig. 4.9: Mean SRT measurements (dB) across all listeners for the different manipulations of the F0 contour (*m*) for the synthesised target speech with both a synthetic and natural F0 contour. Manipulations *m*=1, 0 and -1 correspond to conditions original, monotone and inverse. Error bars represent ± 1 standard error.

1.3 dB with the synthesised F0 contour. Likewise, with the synthesised target speech the difference increased from 0.2 dB with a synthesised F0 contour to 3.4 dB with the natural F0 contour. The increased SRTs in the inverted condition for the natural F0 contour compared to the synthetic contour supports the idea that an inverted natural F0 contour is misleading to the listener. The lack of inversion effect with the synthesised F0 contours implies that the F0 cues listeners exploit to aid speech intelligibility in the natural F0 contours are either absent or insufficiently prominent in the synthetic speech. This supports a conclusion drawn by Kochanski et al. (2005). They found that the larger the F0 excursions within an utterance, the greater the effect of indicating prominence. The natural F0 contours used in this experiment featured much larger F0 excursions than the synthetic F0 contours. Thus, it could be that the smaller F0 movements within the synthetic F0 contours do not clearly mark prominent syllables, reducing the effect of manipulating the F0 contour.

Overall, the results showed that although manipulating the natural speech to become more similar to synthetic speech decreased its intelligibility substantially, manipulating the synthetic speech to be more similar to natural speech did not greatly increase its intelligibility. The duration contour of the synthesised speech significantly reduces the intelligibility of the natural speech, although this is potentially influenced by the increased speech rate. The synthesised F0 contour does not particularly degrade the intelligibility of the natural speech, but the lack of an effect of F0 manipulation for the synthesised F0 contour indicates that the synthesised F0 contour does not contain accurate F0 information. That is, although it does not degrade speech intelligibility significantly, inverting the synthetic F0 contour has less effect than inverting the natural F0 contour. Hence, information usually conveyed by the F0 contour, which would add to the meaning of the utterance is not accurately portrayed by the synthesised F0 contour. Duration cues have been shown to mark phrase boundary (Streeter, 1978) and word boundary location (Nakatani & Schaffer, 1978) more strongly than F0 cues. This greater salience of duration in the perception of phrase and word boundaries compared to F0 accounts for the large difference in intelligibility between the synthetic duration and F0 contours. If so, it would suggest that locating phrase boundaries is very important to intelligibility in noise. The lack of improvement in the synthesised speech intelligibility contrasted to the degradation of the natural speech shows that the phonetic quality of the synthesised speech is too poor to be able to exploit the additional cues given in the natural F0 and duration contours, supporting results from Jilka et al. (2003).

CHAPTER 5:

NON-NATIVE USE OF F0 CUES

5.1 Introduction

We have shown that having a correct F0 contour, with appropriate F0 cues on relevant syllables, is important to speech intelligibility for native English speakers. Native English speakers are able to exploit the F0 cues and contours to enhance the intelligibility of the speech when it is presented against background noise, particularly competing speech. We have attributed this effect in part to listeners actively searching for the content words of the utterances in order to process the import of what the speaker has said (Cutler & Fodor, 1979). These content words are highlighted by F0 accents and the F0 contour surrounding the content words guide the listener to them. Disrupting these cues disrupts the native listener's perception of the speech. The question therefore arises as to whether a similar pattern of results would be seen for non-native speakers of English.

Languages differ in their prosodic structure, and native prosodic expectations have been shown to influence a listener's ability to correctly segment speech (Otake et al., 1993; Cutler et al., 2003); interpret paralinguistic meaning from pitch contours (Chen, Gussenhoven & Rietveld, 2004; Atoye, 2005); and exploit syntactic information (Sanders et al., 2002). Native listeners' performance on speech in noise tasks also exceeds that of non-natives (Mayo et al., 1997; Lecumberri and Cooke, 2006), even for a bilingual-since-infancy group (Mayo et al., 1997). Finally, Cutler et al. (2004) showed that this decrease in performance of non-native speakers in noise compared to native speakers is not due to a disproportionate amount of phoneme misidentification.

These results leave at least three possibilities. First, the cues that native speakers use to distinguish speech from background noise may not be efficiently exploited by non-native speakers. In this case, it could be that non-native speakers are less able to track the F0 contour of the target or of the interfering noise in order to perceptually separate the speech and noise. Second, non-native listeners may fail to exploit information about the sentence content that is carried by the F0 contour. Third, their difficulties in background noise could be the result of less efficient lexical access in a second language.

Mennen (1999) stated that the more similar the intonation patterns between two languages, the harder it is for the non-native speaker to produce the new intonation, since they are unable to form distinct categories for each language. For perception, however, it could be easier for the non-native speakers if the intonation patterns are similar, since they will be able to follow the contour more readily. Non-native speakers in this experiment will be from two groups, native Romance language speakers and native Welsh speakers, both of which are different language groups to English, which is a Germanic language.

Non-native listeners are able to notice differences in intonation contours and exploit F0 cues in their second language to aid speech intelligibility (Akker & Cutler, 2003). This finding implies that inverting the F0 contour could affect the non-native listeners in the same way as native listeners, producing a raised SRT. On the other hand, if non-native listeners are unable to correctly interpret these F0 cues (implied for example in Chen et al., 2004), it could be that the inverted F0 contour will not be as unintelligible for non-native speakers as it is for native speakers, thus inducing a smaller effect of F0 inversion. The third possibility is that non-native speakers rely more heavily on the F0 contour than native speakers due to a poorer understanding of the lexical content. With the increased difficulty of the background noise, it could be that the non-native speakers use the F0 cues to follow the speech more than native speakers due to slower and less accurate word recognition overall. It is possible that Welsh speakers, with the larger frequency range and more lyrical intonation might find the F0 accents and fluctuations in the F0 contour not as salient as the Romance language speakers.

5.2. Experiment 7

5.2.1. Method

5.2.1.1. Listeners

20 paid participants were recruited from Cardiff University Participation Panel. 10 were normally-hearing non-native speakers of English with a Romance first language (French, Spanish, Portuguese or Italian). Half of these participants began to learn English under the age of 14, and half learnt English at or above 14 years old. The remaining 10 participants were native Welsh speakers, all of whom had begun to learn English as infants. None of the listeners were familiar with the sentences used in this study.

5.2.1.2. Stimuli

This experiment was similar to Experiment 3 in which target speech was presented against a single-talker interferer. The F0 contour of the target speech was manipulated, but the interferer F0 contour was not. 50 target sentences from Experiment 3, using speaker CW, were used. The interferer sentences were spoken by DA. The same equations were applied to both the target and interferer voices as in Experiment 3, achieving the different mean F0s of 125 Hz and 210.25 Hz, respectively. The interferer remained normally intonated (*m*=1), as in Experiment 3. The same F0 modulations as used in Experiments 1 and 3 were employed for the target speech (*m*=1, 0.5, 0.25, 0, -1) to represent the five different conditions (original, half, quarter, monotone and inverse). The same procedure was followed as in Experiment 3.

5.2.2. Results and Discussion

Native Romance language speakers found the task much harder than Native Welsh or Native English speakers from Experiment 3, with mean SRTs 8 dB higher than SRTs in the original (*m*=1) condition (see Fig. 5.1). This outcome is consistent with previous reports of non-native speakers' performance, in speech-in-noise tasks. Even listeners who perform as well as native listeners in quiet have much more difficulty in noise (Mayo et al., 1997; Cutler et al., 2004; Lecumberri and Cooke, 2006). The native Welsh speakers used here differ from the native Romance language speakers in that they will have been taught in both English and Welsh throughout school, having much more consistent exposure to English from an early age, hence the difference between the two groups is not particularly surprising. There was also much larger variability among the native Romance speakers' SRTs compared to SRTs from other listeners, reflecting the greater range of ability amongst these listeners.

The native Welsh speakers showed a similar pattern of results to the native English speakers. There was a slight difference of approximately 2 dB in overall SRTs between the two sets of results, but this was not significant. As mentioned above, Welsh has a very distinctive intonation pattern with a large frequency range and an

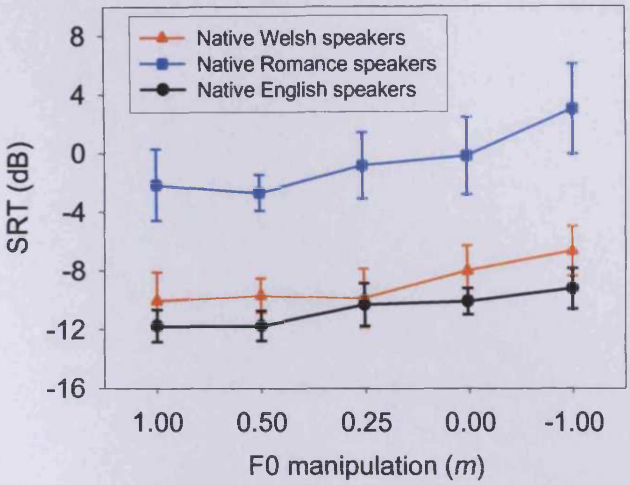

Fig. 5.1 : Mean SRT measurements (dB) for the different manipulations (*m*=1, 0.5, 0.25, 0, -1; original, half, quarter, monotone, inverse) of the F0 contour in both Experiments 3 and 7. Both experiments used the same stimuli, but experiment 3 was run on native English speakers, whereas Experiment 7 was run on non-native English speakers. Error bars represent ± 1 standard error.

accent on the second-to-last syllable of a word. It was thought that this might influence the native Welsh speakers' perception of English intonation such that differences in F0 within the English intonation contour may not be as noticeable for a Welsh speaker as for an English speaker since they would be comparatively small. However, the Welsh speakers were able to both perceive F0 differences and utilise the F0 cues within the contour. The similarity between the results for the English and Welsh speakers could be due to the fact that the native Welsh speakers used here would be have equally immersed in both languages from a young age, and potentially use their English as much as, if not more than Welsh in their adult life.

For the native Romance language speakers, a 2.2 dB difference was found between the original (*m*=1) and monotonous (*m*=0) conditions, with a further 3.3 dB difference between the monotonous (*m*=0) and inverse (*m*=-1) conditions. A repeated measures ANOVA found a significant main effect of the target F0 contour (F(4, 32) = 3.453, p<0.02) for these listeners. Pairwise comparisons with Bonferroni corrections showed there to be a significant difference between the original (*m*=1) and the inverse (*m*=-1) conditions (p<0.05). No other significant differences were found. Thus, it seems that the native Romance listeners do exploit F0 cues and accents in order to aid them in following the target speech, which is a similar result to that found in the Akker and Cutler (2003) experiment with Germanic speakers. This result also supports evidence suggesting the vowel intrinsic F0 is universal (Whalen & Levitt, 1985). Manipulating the F0 contours interferes with this intrinsic difference in F0 between high and low vowels. It is possible that interrupting this cue to vowel identification affects the speech intelligibility. Given that both native and non-native speakers experience an effect of F0 inversion, it is possible that intrinsic vowel F0 is an important cue to vowel perception, present in all languages tested here.

A larger difference was found between the original (*m*=1) and inverse (*m*=-1) conditions for the Romance non-native speakers (5.5 dB) than for native speakers (2.9 dB). This difference suggests that the Romance non-native listeners are able to learn the F0 cues of their second language or that the important cues are those that are shared between many languages (Sanders et al., 2001). With poorer lexical recognition than native speakers and increased difficulty in background noise the Romance non-native listeners may rely more heavily on these cues than native speakers to direct them to the content words of the sentence explaining the somewhat larger effect.

A still larger effect of inversion was observed among listeners who learnt English relatively late. The Romance language speakers were separated into two groups: those who had learnt English before the age of 14 and those who had learnt English at or later than the age of 14. Looking at these two groups separately, a large difference of F0 manipulation was found for those who had learnt English after the age of 14 (6.9 dB between *m*=1 and *m*=-1), with a much smaller difference for those who had learnt English before the age of 14 (1.8 dB between *m*=1 and *m*=-1).

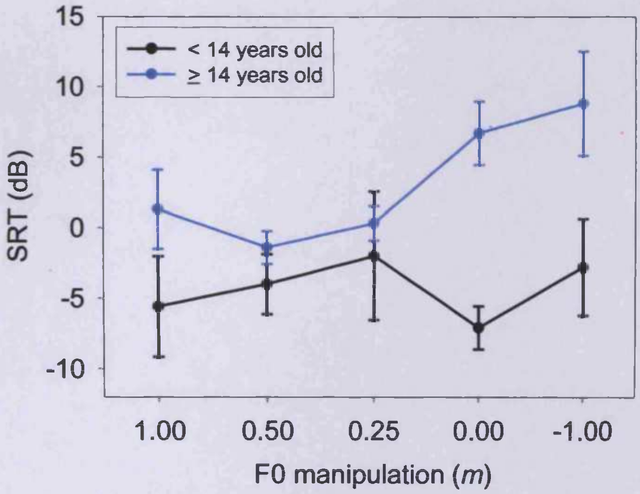

Fig. 5.2: Mean SRT measurements (dB) for the different manipulations (*m*=1, 0.5, 0.25, 0, -1; original, half, quarter, monotone, inverse) of the F0 contour in Experiment 7 comparing participants who started to learn English below the age of 14 and those that started to learn at or above 14 years of age. Error bars represent ±1 standard error.

A significant interaction was found between the F0 manipulation and the age of learning (F(4, 32) = 4.658, p<0.005). Simple main effects showed that there was no significant effect of F0 manipulation for those that learnt English below the age of 14

(F(4,32)=1.475, $p>0.05$), but a significant effect of F0 manipulation was seen for those who learnt English at or above the age of 14 (F(4,32)=5.226, $p<0.05$). Pairwise comparisons with Bonferroni corrections showed there to be a significant difference between the original (m=1) and inverse (m=-1) conditions ($p<0.04$) and between the half (m=0.5) and monotone (m=0) conditions ($p<0.02$) for those who learnt English at or over 14. A significant difference was also found using pairwise comparisons between the two different ages of learning in the monotone condition (m=0), ($p<0.0002$) and the inverse condition (m=-1), ($p<0.05$).

Mayo et al. (1997) found that highly-proficient non-native listeners who learnt English after puberty perform worse in noise than those who learnt as infants or toddlers. The results here show that those who learnt after the age of 14 had an increased effect of F0 inversion compared to those who learnt before the age of 14. This implies that those listeners who are less familiar with the language rely even more heavily on the F0 contours to increase the intelligibility of the speech. It is possible that these listeners are not as proficient as early bilinguals in recognising non-native speech in noise; hence they require the F0 contour to give them cues to where the important content words lie.

Interestingly, unlike in the Mayo et al. (1997) experiment, it is only in the monotonous (m=0) and inverse (m=-1) conditions where the two groups differ significantly in performance, implying that when the F0 contour was intact both groups were able to perform to a similar standard. Only when the F0 cues were removed did the late learners find the task more difficult than the early learners. Mayo et al. (1997) found that bilingual post-puberty learners tolerated less background noise even with a normally-intonated contour than speakers who learnt the language earlier. Mayo et al. compared high and low predictability target sentences. The average difference between the non-native groups was larger for the high predictability sentences than the low predictability sentences. Late learners were less able than early learners to make use of sentence context, with the bilingual post-puberty group showing no effect of sentence context. It is possible that this result skewed the overall difference between the two groups of learners. The experiment reported here used only low-predictability sentences and therefore neither group had the advantage of being able to exploit sentence context.

The difference found between late and early learners is consistent with results shown in speech production experiments where late bilinguals are less able to produce native-like pronunciation than early bilinguals (Flege, Yeni-Komshian & Liu, 1999; Piske, Flege, Mackay & Meador, 2001). Piske et al. (2002) showed that frequent second language (L2) use improves pronunciation for both early and late learners. All participants in these experiments used their L2 frequently; hence no comparisons of this kind could be made. The length of residence in an English-speaking country was also investigated. No significant difference was seen between the group that had lived in an English-speaking country for longer than 2 years, and those that had lived in an English-speaking country for less than 2 years.

Thus, it seems that non-native speakers can learn and use the F0 cues of their second language. In fact, the larger difference in SRT between the original (m=1) and inverse (m=–1) conditions for non-native speakers indicates that they perhaps rely on the F0 contour more heavily than native speakers, potentially due to poorer lexical recognition than the native speakers. The results also imply that it is not a lack of ability to use the F0 contour in the non-native speech that causes the non-native speakers to perform so much worse than native speakers when the speech is presented in background noise. Coupled with Cutler et al.'s argument (2004) that it is not poorer recognition of phonemes that causes the increased impact of masking on speech intelligibility scores for non-native listeners, it seems that there must be some other factor causing the deterioration in speech intelligibility when played in background noise such as poorer lexical access.

CHAPTER 6:

DISCUSSION

The experiments described in this thesis have primarily been concerned with the effect of Fundamental Frequency contour on speech intelligibility. In particular, SRTs for manipulations of the F0 contour have been compared for natural and synthetic speech in studies involving both native and non-native participants. The duration contour of synthesised speech was also studied and compared to that of natural speech. SRTs produce an objective record of the SNR at which listeners can achieve 50% correct recognition accuracy. This is measured in decibels. Therefore, it can give an objective measure of the quality of speech materials in terms of human perception. Different forms of manipulation can easily be compared using a common metric.

The salient results of this research are summarised in section 6.1 below. Section 6.2 discusses explanations for these results and Section 6.3 proposes topics suitable for further research following on from the results of this thesis. Finally, a brief summary is presented in 6.4.

6.1. Summary of Results

The most important results from Experiments 1 to 7 are summarised below:

1. Reducing the amount of variation within the F0 contour so that it is monotonous, reduces the intelligibility of an utterance in the presence of interfering sound. Inverting the F0 contour, and therefore misplacing the

natural F0 cues but maintaining variation, causes further reduction in speech intelligibility (experiments 1-3).

2. The effect of F0 manipulation is more marked when speech is played against a single-talker interferer as compared to target speech in a background of speech-shaped noise (experiments 1-3).
3. Manipulation of the interferer's F0 contour does not affect the intelligibility of the target speech (experiment 2).
4. Low-pass filtering the F0 contour showed that the most important frequencies for the F0 manipulation effect lie between 2 and 4 Hz. The rate of stressed syllables in the target speech lies between these values (experiment 4).
5. Synthesised speech is much less intelligible than natural speech when played against interfering speech. However, this result may have been affected by the potential confound of speech rate (experiments 5 and 6).
6. Replacing the natural duration contour with a synthesised duration contour significantly raises SRTs for natural target speech. This may also be associated with the increased speech rate of the synthesised speech (experiment 6).
7. No improvement in SRTs is found when synthetic F0 and duration contours are replaced with natural contours. Therefore, the increased speech rate of the synthesised speech may not be the cause of the effects mentioned above in 5) and 6) as decreasing the speech rate with the natural duration contour does not increase its intelligibility. This may occur because the phonetic quality of the synthesised target speech is too poor to allow listeners to exploit the cues present in the natural F0 and duration contours (experiment 6).

8. Although replacing the natural F0 contour with a synthesised F0 contour does not detrimentally affect the overall intelligibility of the natural speech, the effect of F0 inversion is not replicated with the synthesised F0 contour. This suggests that the synthetic F0 contour does not carry the same information as the natural F0 contour (experiment 6).
9. F0 inversion also affects non-native listeners. A larger difference between the standard (m=1) and inverse (m=-1) conditions was observed for the Romance speakers than for native English and Welsh speakers (experiment 7).

6.2. Explanations

6.2.1. Effect of F0 inversion

The effects caused by F0 inversion described in these experiments may be due to several processes: stress, F0 contour shape and intrinsic vowel pitch. The most important frequencies for this effect lie around the stressed syllable rate of speech.

6.2.1.1. Stress and F0 contour shape

The results of the experiments described in Chapters 2 and 3 are consistent with previous findings which showed that F0 stress cues direct the listener to the important content words of the utterance (Cutler & Fodor, 1979). F0 cues also aid the listener in parsing speech as they are a valuable cue to syntactic (Ladefoged, 1993), phrase (Streeter, 1978) and content word boundaries (Cutler & Butterfield, 1992). Incorrect parsing of the speech may reduce the ease with which listeners segment speech and locate the focus of the sentence. In a normally-intonated sentence, the content words of utterances tend to be 'highlighted' using prosodic cues including F0 information. F0 is therefore an important cue to the accented words (Fry, 1958; Morton & Jassem, 1965; Nakatani & Schaffer, 1978). Even when the stress is removed from the words, as seen in experiments by Cutler (1976) and Cutler and Fodor (1979), the surrounding F0 contour guides the listener to the content words of the utterance. In a monotonised sentence, all F0 cues directing the listener to the focus of the sentence have been removed. Without these cues, listeners must rely on cues such as context or prosodic information other than F0 to direct them to the sentence focus. The reduced intelligibility of the monotonised sentences compared to the normally-intonated sentences indicates that removing F0 as a cue to sentence focus has a measurable effect on intelligibility.

Inverting the F0 contour misplaces any F0 cues on focussed words in addition to distorting the F0 contour shape. Although there are F0 cues within the utterance, they are misleading. F0 inversion is more detrimental to speech intelligibility than simply removing the F0 cues. This adds further support to the argument that listeners actively search for the sentence focus. When cues are misleading, listeners may be directed to an incorrect focus. This disrupts their ability to quickly identify the important words within an utterance. This appears to contribute to the reduced intelligibility of the speech.

6.2.1.2. Vowel Intrinsic pitch

Vowels have different intrinsic pitches. Low vowels, such as /ɑ/ tend to have intrinsically lower pitch than high vowels such as /i/. Diehl (1991) states that this is an example of 'auditory enhancement' such that a speaker produces this intrinsic F0 difference to enhance the speech signal for the listener. It could, however, simply be a physiological property of producing such sounds. Regardless of the reason for the differences in the vowel F0, producing consistent differences between the high and low vowels may enable the listener to more clearly identify each vowel than if they were produced at the same F0. Vowel intrinsic pitch may therefore be a property of vowels that aids their recognition. Monotonisation removes the F0 differences, and reduces any benefit of auditory enhancement concerning these differences. Moreover, inverting the F0 contour may cause further confusion because a normally high F0 on a high vowel would be replaced with a low F0, and vice versa. Hence typically low vowels would have higher F0 values than high vowels. The universality of vowel intrinsic pitch (Whalen & Levitt, 1995) could explain why both native and non-native speakers experienced an effect of F0 manipulation.

6.2.1.3. Syllable/phoneme debate

The results from experiment 4 demonstrate that the most important frequencies for the F0 inversion effect lie between 2 and 4 Hz. As previously described, this is below the syllable rate and around the stressed syllable rate of the speech used. This highlights the importance of both the accents placed on the syllables and the contours of the larger prosodic units surrounding these focussed syllables to F0 perception. Whether the syllable or a smaller unit, such as the phoneme, is the basic unit of speech perception is a much debated topic. Mehler et al. (1991), Segui (1984) and Greenberg (1996), for instance, argue that the syllable is the primary unit of speech perception. Alternatively, Cutler and Norris (1988) have demonstrated the importance of the phoneme as a unit of speech perception. The experiments described in this thesis have determined that the stressed syllable is a primary unit affecting the intelligibility of the F0 contour. Earlier it was noted that listeners actively search for syllable stress in order to determine the important content words of the speech (Cutler

& Fodor, 1979) and use syllable stress to aid in speech segmentation (Cutler & Butterfield, 1992). It would therefore seem correct to assume that the stressed syllable is an important unit in the perception of F0 contours. In terms of slower fluctuations regarding larger prosodic units, results from Experiment 4 support the conclusion that listeners use the F0 contour surrounding the accented syllables to aid them in interpreting the content words of the utterance because lower frequencies contributed to the effect.

6.2.2. Effect of manipulating duration contours

Replacing duration contours in natural speech with synthetic duration contours significantly degraded the intelligibility of the speech. The reduction in intelligibility of the natural target speech implies that the synthetic duration contours carry different information than natural duration contours. Natural duration reliably predicts phrase (Streeter, 1978) and lexical boundaries (Nakatani & Schaffer, 1978). It also clearly marks lexical stress (Fry, 1958). Disrupting duration cues may therefore cause difficulty in correctly parsing speech into phrases, perceiving word boundaries and lexical stress. Duration has been shown to be a stronger predictor of lexical stress (Kochanski et al., 2005) and phrase boundary location (Streeter, 1978) than F0. This may explain why synthetic duration contours produced a greater detrimental effect on the natural target speech than the synthetic F0 contours. However, this difference may reflect the fact that synthetic duration contours are inferior to the synthetic F0 contours.

6.3. Further Research

6.3.1. Synthesised speech

Synthesised speech is much less intelligible than natural speech. The majority of research regarding synthesised speech investigates either the perceived naturalness of the speech subjectively or how similar to natural speech it is objectively. Few directly investigate its intelligibility. Experiments in this thesis found that neither the synthesised F0 contour nor synthesised duration contour fully emulate the intelligibility of the natural contours. Synthesised duration contours degraded natural speech intelligibility significantly and no consistent effect of F0 inversion was found for the synthesised F0 contours. Given the large detrimental effect of the synthetic duration contour, it would be interesting to investigate the impact of manipulating these contours further.

There was no improvement when natural F0 or duration contours were imposed on synthesised speech. This finding may suggest that the phonetic quality of the synthesised speech is too poor to exploit the additional cues within the natural speech contours. Further research testing the intelligibility of the phonetics of synthesised speech is required to improve speech synthesis. Considering the reduced quality of the synthetic prosody, it may be easier to detect improvements to the phonetics of the system if the prosody is accurate. Therefore, it may be useful to place natural prosodic contours onto the synthesised speech whilst attempting to improve the phonetics. Similarly, since no improvement to synthetic speech intelligibility was noted when natural prosodic contours were placed onto the speech, it may be easier to detect improvements in synthetic prosodic cues if they are tested on natural speech where the phonetics are correct. As the synthetic prosodic contours begin to more accurately mimic the intelligibility of the natural contours, the degradation in natural speech intelligibility using the synthetic contours should lessen.

6.3.2. Prosodic cues

Listeners actively search for the sentence focus and F0 cues are important in indicating the position of this focus. The experiments in this thesis, however, did not investigate the role of other prosodic cues such as duration and intensity in highlighting content words. In both the monotonised and inverted F0 conditions, prosodic cues other than F0 remain intact. Despite the reduced intelligibility of both the monotonised and inverted F0 utterances compared to normally-intonated sentences, they are not completely unintelligible: listeners are able to transcribe the sentence, albeit at higher signal-to-noise ratios. It is possible that the listener can glean some information about the sentence focus position from cues other than the F0. All stimuli were low-predictability sentences making it unlikely that context would be a particularly strong cue. Duration is an important cue to word boundary perception (Nakatani & Shaffer, 1978), prosodic boundary perception (Streeter, 1978) and lexical stress (Fry, 1958). Although intensity is not as influential as duration or F0 cues (Holmes and Holmes, 2001), it has been found to aid perception of lexical stress (Fry, 1958; Morton and Jassem, 1965). Results from experiment 7 found that the duration cues from the synthesised speech were highly detrimental to the intelligibility of the natural speech. Although no strong conclusions can be drawn from this result due to the potential confound of speech rate, it does indicate a need to investigate the contribution of additional aspects of prosody to speech intelligibility, if only to determine the relative strength of each cue.

6.3.3. Interference

Our experiments determined there was a more pronounced effect of F0 manipulation that occurred against a single-talker interferer than when speech was presented in speech-shaped noise. Manipulating the F0 contour of the interfering speech did not affect its masking effect, indicating that as long as the interferer is speech, it does not matter what the F0 contour of the interferer is doing. It is not clear why the use of a speech interferer accentuates these effects. A single-talker interferer produces less masking than modulated noise. This result, therefore, cannot be explained in terms of an increased masking effect, causing the listener to rely more heavily on the F0 contour. Considering informational masking, it may be that with a single-talker interferer, the listener is forced to follow the target F0 contour more closely in order to differentiate between the two talkers. With a speech-shaped noise interferer, there is potentially a larger perceived difference between the target speech and interferer; hence the F0 contour is not as salient a cue. However, the 9 semitone difference between the F0 contours of the target and interferer across these experiments should minimise similarity and hence reduce any effect of informational masking. Analysing the listeners' transcripts did not indicate any intrusions from the interfering speech making it difficult to ascribe this effect to informational masking. In order to pursue this issue further, it would be interesting to use different interferers such as continuous nonsense speech, reversed speech and multiple-talker interferers. If the effect is associated with the content of the interfering speech, it should be reduced when using nonsense words in the interferer. Increasing the number of interfering voices should also reduce the interference from sentence context, if indeed informational masking is the problem.

6.4. Summary

The F0 contour is important to speech intelligibility, particularly when speech is heard against a speech interferer. Listeners rely on the F0 contour to guide them to the sentence focus. Reduced or misleading F0 cues reduce the intelligibility of the utterance. This thesis would suggest that the inclusion and correct interpretation of F0 in ASR systems, especially in difficult listening conditions, is therefore computationally worthwhile. Synthesised speech is much less intelligible than natural speech. Although the role of F0 seems to be small in this decreased intelligibility, the role of other prosodic cues such as duration could be large. The intelligibility of the phonetics of the synthesised speech requires further investigation.

BIBLIOGRAPHY

Akker, E., & Cutler, A. (2003). Prosodic cues to semantic structure in native and non-native listening. *Bilingualism: Language and cognition, 6* (2), 81-96.

Arai, T., Pavel, M., Hermansky, H., & Avendano, C. (1996). Intelligibility of speech with filtered time trajectories of spectral envelopes. *Proceedings of the International Conference of Spoken Language Processing, Philadelphia, (4)*, 2490-2493.

Assmann, P. F. (1999). Fundamental frequency and the intelligibility of competing voices. *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, Aug. 1-7, 1999*, 179-182.

Atoye, R. O. (2005). Non-native perception and interpretation of English intonation. *Nordic Journal of African Studies*, 14 (1), 26-42.

Beckman, M., & Pierrehumbert, J. L. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255-310.

Bird, J., and Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sentences. *Paper for the 11th International Conference on Hearing.*

Black, A. W., Taylor, P. A., & Caley, R. J. (2003). *The Festival Speech Synthesis System.* University of Edinburgh.

Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, 14, 109-149.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America, 84* (1), 101-114.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10*, 23-36.

Brungart, D. S. (2001). Evaluation of speech intelligibility with the coordinate response measure. *Journal of the Acoustical Society of America, 109* (5), 2276-2279.

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America, 110* (5), 2527-2538.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *Journal of the Acoustical Society of America, 45*, 694-703.

Chen, A., Gussenhoven, C., & Rietveld, T. (2004). Language-specificity in the perception of paralinguistic intonational meaning. *Language and Speech, 47* (4), 311-349.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America, 93*, 3271-3290.

de Cheveigné, A. (1997). Concurrent vowel identification III: A neural model of harmonic interference cancellation. *Journal of the Acoustical Society of America, 101*, 2857-2865.

Clark, R. A. J., & Dusterhoff, K. E. (1998). Objective methods for evaluating synthetic intonation. *Eurospeech'99: Budapest*, 1623-1626

Connine, C. M., Clifton, C. Jr., & Cutler, A. (1987). Effects of lexical stress on lexical categorisation. *Phonetica, 44* (3), 133-146.

Cruttenden, A. (1986). *Intonation.* Cambridge University Press.

Crystal, D. (1969). *Prosodic Systems and intonation in English.* Cambridge: Cambridge University Press.

Culling, J. F., & Colburn, H. S. (2000). Binaural sluggishness in the perception of tone sequences and speech in noise. *Journal of the Acoustical Society of America, 107* (1), 517-527.

Culling, J. F., & Darwin, C. J. (1994). Perceptual and computational separation of simultaneous vowels: cues arising from low-frequency beating. *Journal of the Acoustical Society of America*, 95 (3), 1559-1569.

Culling, J. F., Hodder, K. I., & Toh, C. Y. (2003). Effects of reverberation on perceptual segregation of competing voices. *Journal of the Acoustical Society of America, 114*, 2871-2876.

Culling, J. F., Linsmith, G. M., & Caller, T. L. (2005). Evidence for a cancellation mechanism in perceptual segregation by differences in fundamental frequency. *Journal of the Acoustical Society of America, 117* (4), p.2600.

Culling, J. F., & Summerfield, Q. S. (1995). Perceptual segregation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America, 98* (2), 785-797.

Cutler, A. (1976). Phoneme-monitoring reaction times as a function of preceding intonation contour. *Perception and Psychophysics,. 20*, 55-60.

Cutler, A. (1997). The comparative perspective on spoken-language processing. *Speech Communication, 21*, 3-15.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31*, 218-236.

Cutler, A., & Clifton, C. (1984). The use of prosodic information in word recognition. *Attention and Performance, X*, 183-198.

Cutler, A., Dahan, D., & Donselaar, W. van (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech, 40,* 141-201.
Cutler, A., & Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition, 7*, 49-59.

Cutler, A., Murty, L., & Otake, T. (2003). Rhythmic similarity effects in non-native listening. *Proceedings of the 15th International Conference of Phonetics: Barcelona.*

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America, 116*, 3668-3678.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *Journal of the Acoustical Society of America, 114* (5), 2913-2922.

Diehl, R. L. (1991). The role of phonetics within the study of language. *Phonetica, 48*, 120-134.

Drullman, R., & Bronkhorst, A. W. (2004). Speech perception and talker segregation: Effects of level, pitch and tactile support with multiple simultaneous talkers. *Journal of the Acoustical Society of America, 116* (5), 3090-3098.

Drullman, R., Festen, J.M., & Plomp, R. (1994a). Effect of temporal smearing on speech reception. *Journal of the Acoustical Society of America, 95* (2), 1053-1064.

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America, 95* (5), 2670-2680.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, S., & Kidd, G. (2003). Informational masking: counteracting the effects of stimulus incertainty by decreasing target-masker similarity. *Journal of the Acoustical Society of America, 114* (1), 368-379.

Dusterhoff, K., & Black, A.W. (1997). Generating F0 contours for speech synthesis using the Tilt intonation theory. *Proceedings of the 1997 ESCA workshop on intonation: Athens, Greece.*

Festen, J.M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America, 88* (4), 1725-1736.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41* (1), 78-104.

Fletcher, H., & Galt, E. H. (1950). The perception of speech and its relation to telephony. *Journal of the Acoustical Society of America, 22*, 89-151.

French, N. R., & Steinberg, J. C. (1946). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19* (1), 90-119.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America, 115*, 2246-2256.

Fry, D.B. (1958). Experiments in the perception of stress. *Language and Speech, 1*, 126-152.

Grabe, E., Post, B., & Nolan, F. (2001). Modelling intonational variation in English: The IViE system. In *Proceedings of Prosody 2000,* ed. S. Puppel and G. Demenko, 51-57.

Graddol, D. (1986). Discourse specific pitch behaviour. *Intonation in Discourse.* London and Sidney: Croom Helm, 221–237.

Greenberg, S. (1996). Understanding speech understanding: Towards a unified theory of speech perception. *Proceedings of the ESCA Workshop on the "Auditory Basis of Speech Perception,"*: Keele University, 1-8.

Gussenhoven, C. (1983). Focus, mode and the nucleus. *Journal of Linguistics*, 19, 377-417.

Gussenhoven, C. (2004). *The phonology of tone and intonation.* Cambridge University Press.

Gustafsson, H. A., & Arlinger, S. D. (1993). Masking of speech by amplitude-modulated noise. *Journal of the Acoustical Society of America, 95* (1), 518-529.

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2003). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America, 115* (2), 833-843.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies.* Chicago: University of Chicago Press.

Hazen, T. J. (2006). Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Transactions on Audio, Speech and Language Processing. 14* (3), 1082-1089.

Hillenbrand, J. M. (2003). Some effects of intonation contour on sentence intelligibility. *Journal of the Acoustical Society of America, 114* (4), 2338.

Hirst, D., & di Cristo, A. (1998). A survey of intonation systems. *Intonation systems: a survey of twenty languages*. Cambridge University Press.

Holmes, J., & Holmes, W. (2001). *Speech Synthesis and Recognition.* Second Edition. Taylor and Francis.

Jilka, M., Syrdal, A.K., Conkie, A.D., & Kapilow, D.A. (2003). Effects on TTS quality of methods of realizing natural prosodic variations. *Proceedings of ICPhS, Barcelona.*

Kanedera, N., Arai, T., Hermansky, H. & Pavel, M. (1999). On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication, 28,* 43-55.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59* (5), 1208-1221.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America, 118* (2), 1038-1054.

Kucera, H., & Francis, W. N. (1982). *Frequency analysis of English usage.* Boston: Houghton and Mifflin.

Ladd, D. R. (1979). *Basic bibliography of English intonation.* Bloomington: Indiana University Linguistics Club.

Ladd, D. R., (1996). *Intonational Phonology.* Cambridge University Press.

Ladd, D. R., & Silverman, K. E. A. (1984). Vowel intrinsic pitch in connected speech. *Phonetica, 41,* 31-40.

Ladefoged, P. (1993). *A course in phonetics.* Third edition. Hardcourt Brace.

Laures, J. S., & Bunton, K. (2003). Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of Communication Disorders, 36,* 449-464.

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language and Hearing Research, 42,* 1148-1156.

Lecumberri, M. L. G., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *Journal of the Acoustical Society of America, 119* (4), 2445-2454.

Lehiste, I. (1970) *Suprasegmentals.* Cambridge, MA: MIT Press.

Lieberman, P. (1967). *Intonation, perception and language.* Cambridge, MA: MIT Press.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America,* 69 (6), 1763-1769.

Mattys, S. L. (2004). Stress versus coarticulation: Toward an integrated approach to explicit speech segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 30* (2), 297-408.

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of second language acquisition and perception of speech in noise. *Journal of Speech, Language and Hearing Research, 40*, 686-693.

McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *Journal of the Acoustical Society of America, 86* (6), 2148-2159.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of verbal learning and verbal behaviour, 20*, 298-305.

Mennen, I. (1999). The realisation of nucleus placement in second language intonation. *Proceedings from the 14th International Congress of Phonetic Sciences (ICPhS 99), San Francisco, August 1-7, 1999.*

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22* (2), 167-173.

Morton, J., & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech, 8*, 159-181.

Müsch, H. (2000). Review and computer implementation of Fletcher and Galt's method of calculating the Articulation Index. *Acoustics Research Letters Online, 2* (1), 25-30.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America, 63* (1), 234-245.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, 771-774.

Norris, D., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception and Psychophysics, 43*, 541-550.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language, 32*, 258-278.

Peissig, J. and Kollmeir, B. (1997). Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *Journal of the Acoustical Society of America, 101* (3), 1660-1670.

Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. PhD thesis, MIT.

Pike, K. L. (1945). *The intonation of American English.* Ann Arbor: University of Michigan Press.

Piske, T., Flege, J. E., Mackay, I. R. E., & Meador, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica, 59* (1), 49-71.

Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception thresholds for sentences. *Audiology, 18,* 43-52.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90* (6), 2956-2970.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H.R., Urbanek, G. E., & Weinstock, M. (1969). I.E.E.E. recommended practice for speech quality measurements. *I.E.E.E. Trans. Audio Electroacoust., 17,* 227-246.

Sanders, L. D., Neville, H. J., & Woldorff, M. G. (2002). Speech segmentation by native and non-native speakers: the use of lexical, syntactic and stress-pattern cues. *Journal of Speech, Language and Hearing Research, 45,* 519-530.

Segui, J. (1984). The syllable: a basic perceptual unit in speech processing? *Attention and Performance, X,* 165-181.

Shadle, C. H. (1985). Intrinsic fundamental frequency of vowels in sentence context. *Journal of the Acoustical Society of America, 78* (5), 1562-1567.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. *Proceedings of ICSLP92,* 2, 867-870.

Silverman, K., & Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. In *Papers in laboratory phonology, vol. 1: Between the grammar and physics of speech.* Ed. Kingston & Beckman. Cambridge University Press. 71-106.

Steele, S. A. (1986). Interaction of Vowel F0 and prosody. *Phonetica, 43,* 92-105.

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America, 67,* 318-326.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America, 64* (6), 1582-1592.

Stubbs, R. J., & Summerfield, Q. (1990). Effects of signal-to-noise ration, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms. *Journal of the Acoustical Society of America, 89 (3),* 1383-1393.

Taylor, P. (2000). Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America, 107*, 1697-1714

Traunmüller, H., & Branderud, P. (1989). Paralinguistic speech signal transformations. *STL-QSPR*, 63-68.

Qin, M. K. & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *Journal of the Acoustical Society of America, 114* (1), 446-454.

Umeda, N. (1981). Influence of segmental factors on fundamental frequency in fluent speech. *Journal of the Acoustical Society of America, 70* (2), 350-355.

Vroomen, J, van Zon, M., & de Gelder, B. (1996). Cues to speech segmentation: Evidence from juncture misperception and word spotting. *Memory and Cognition, 24 (6),* 744-755.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics, 23*, 349-366.

Williams, B. (1985). Pitch and duration in Welsh stress perception: the implications for intonation. *Journal of Phonetics, 13* (4), 381-406.

Williams, B. (1999). Text-to-speech synthesis for Welsh and Welsh English. *Proceedings of Eurospeech 99, Budapest, Hungary.*

Wingfield, A., Lombardi, L., & Sokol, S. (1984). Prosodic features and the intelligibility of accelerated speech: Syntactic versus periodic segmentation. *Journal of Speech and Hearing Research, 27*, 128-134.

APPENDIX 1

List of continuously-voiced sentences used for Experiment 4

the YELLOW LION WORE an IRON MUZZLE
EVERY MAN WON a LEMON RAZOR
the OILY RIVER RAN in the RURAL VALLEY
the WEARY WOMAN LAY on the MAIN LAWN
the MEAN ARMY WON EVERY WAR
a LOVELY WELL is NEAR the WOOL MILL
YOUR LAWYER is AWARE of the VALUE of the VAN
MOVE the LIME WIRE over the NEW RAIL
NINE MEN OWE ME MONEY
AIM the ARROW OVER the LOW WALL
ALLOW the LOYAL ANIMAL a WARM LAIR
the ENEMY is AWARE of YOUR INNER ZONE
WARM RAIN RAN over the IVORY MIRROR
MANY WOMEN LIVE a MILE AWAY
WARN OUR MEN of the EARLY ALARM
NONE of THEM RAN VIA the MARINA
EVEN the LIVELY EARL was AWAY ILL
NEARLY ALL the NAVY was WARY of MAYOR
the EVEN WAVES will OVERWHELM the LONELY MARINER
ROYAL LOONIES were ALL OVER the MOOR
ELEVEN REMOVAL MEN were in the LOWER ROOM
the WARNING of NAVAL INVASION was a MARVELLOUS RUSE
an ARRAY of RAW MELON is ALWAYS on the MENU
WEAVE your LINEN on the LARGE NARROW LOOM
the EVIL VILLAIN ALWAYS LIES and is EVASIVE
WOMEN RARELY MARRY in a YELLOW VEIL
the RAVEN ROSE OVER the RIM of the RAVINE
the RARE RHYTHM was the ENEMY of ALL REASON
LAZY LIMBS LAY OVER the WEIR
the WILLOWS WAVE OVER the LEISURELY LAKE
RAISE the ALARM WHENEVER the LEVEL RISES
a NORMAL MALE loses NINE ENAMEL MOLARS
in the ARENA, ROMANS EARN MANY ENEMIES
we MOVE AWAY the REVELERS in the MAUVE ROOM
SMALL MARINE MAMMALS were EVERYWHERE in the MUSEUM
we ALL WOVE WARM WOOLIES in the MILL
we KNEW of the ANIMAL'S LOW EYES and EARS
ALL LOVE UNRULY REVELRY in the EVENING
their ZEAL is ONLY the ILLUSION of EARLY RISING
MOTHER was NEVER in the VILLA with OTHER WOMEN
MINERAL EROSION REMAINS the MAIN WORRY
you will NEVER LEAN EASILY on a LAME LIMB
my MEMOIRS are VENOMOUS in EVERY LINE or WORD
MEASLES are MAINLY a MALAISE of the VERY YOUNG

the REASON WHY we were ALUMNAE was ALWAYS UNKNOWN
REAL MAYONAISE is LOVELY or MELLOW on the NOSE
ANYONE WHO is EARLY will VIEW the MARE
the NEWS of the RALLY INVOLVES NEARLY EVERYONE
the NEW MOVIE REVEALS our ONLY ERROR
REMOVE the RIVAL MILLER ALIVE in the MORNING
LONELY MEN ROAM EVERY REALM
we OIL the REAR RAILWAY ALL YEAR
EVEN a MINIMAL WIN ALARMS MANY
EVERY VOLUME VARIES in a NOVEL WAY
our LONE MEMORY of ONE WARREN was VIVID
the MILLIONAIRE'S MANOR is OVER the NORMAN RIVER
ALL WHO OWN NAVY are NORMAL
we LEARN NEARLY ALL our MANNERS in YOUTH
NORMALLY the LEAN ROWERS WEAR LEATHER
RAW LIVER as a MEAL WORRIES ME
my MORAL LOVER is UNAWARE of the MINER'S NAME
the NERVE in your ARM NEVER ANNOYS YOU
RELY on the WARY EYE of the VAIN ALIEN
you VALUE YOUR REVOLVER MORE than ME
MOREOVER the IRONY of the NOVEL was VERY MINIMAL
RAISE the OLIVE LEVER NEAR YOU
the MINI WALL MURAL is OVER THERE
THESE ROSES LINE the AVENUE EVENLY
NOONE KNOWS the NUN is REALLY a LIAR
we will LOSE YOU in the ZOO'S NARROW MAZE
the NEW ROYALS REIGN EVILY over the REALM
ZOOM in ANYWHERE on the NEW MOON NOW
MIRROR MY EVERY MOVE in the ROOM
the MEN'S MORALE was NEVER a WORRY of MINE
HAVE the ANNUAL REUNION in LOVELY VIENNA
they will MOAN if you REMOVE THEIR LEISURE HOUR
EARLIER the LEAVES were on the REALLY WORN LANE
USE the NEON NAIL in THEIR URN
they will MOURN the LOSS of ONE as MERRY as YOU
NEVER KNEEL on ANY WALL in the VENUE