The role of temporal fine structure information in the perception of complex sounds for normal-hearing and hearing-impaired subjects

Kathryn Hopkins
St Catharine’s College

November 2008
This dissertation is submitted for the degree of Doctor of Philosophy
Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this work has been submitted as part of any other qualification.
Summary

The peripheral auditory system acts as an array of bandpass filters. Temporal information at the output of each filter can be classified as temporal fine structure (TFS), the rapid oscillations close to the filter centre frequency, and temporal envelope, modulations in amplitude superimposed on this TFS. The roles of TFS in the perception of complex tones and speech were investigated for normal-hearing (NH) and hearing-impaired (HI) subjects.

The discrimination of harmonic and frequency-shifted bandpass-filtered complexes was used as a measure of TFS sensitivity. HI subjects were much less sensitive to TFS in complex tones than NH subjects, with many HI subjects scoring no better than chance when only TFS cues were available for discrimination.

The use of TFS information in speech was assessed using a speech signal that was split into a large number of frequency bands, which were either vocoded, to contain only temporal envelope information, or unprocessed, and so contained both envelope and TFS information. HI subjects benefitted less than NH subjects as the number of unprocessed bands was increased, indicating that HI subjects made less effective use of TFS.

The hypothesis that TFS information is important for dip listening was tested using a similar method, with NH subjects. As the number of unprocessed bands was increased, more benefit was measured in amplitude-modulated noise than steady noise, suggesting that TFS is particularly important for listening in a fluctuating background.

The importance of TFS at different spectral regions in speech was investigated. NH subjects benefitted similarly from TFS added to each of the spectral regions tested (spanning 100-8000 Hz). HI subjects showed no significant benefit when TFS was added to only one spectral region and benefitted less than NH subjects when TFS information was added to the whole spectrum. Benefit from TFS information in speech for HI listeners was significantly correlated with a psychophysical measure of TFS sensitivity.

Finally, a processing scheme that removed temporal envelope cues, leaving TFS cues relatively intact, was assessed. The ability of the auditory system to recover temporal envelope information from the processed signal was investigated, as was the effect of adding a low-level, low-fluctuation noise to each bandpass channel of the speech signal before further processing. This was hypothesised to improve the fidelity of TFS coding by reducing large excursions in instantaneous frequency in low-level signal portions. For NH listeners, the added noise improved intelligibility for certain types of sentences.

Together, these experiments suggest that TFS information is important for speech perception when listening in fluctuating background noise. The relative insensitivity of HI subjects to TFS may partly account for their particular difficulties listening to speech in such environments.
Acknowledgements

I would like to thank:

My supervisor, Brian Moore for his encouragement, inspiration, patience, and remarkable efficiency in reading many drafts of my work.

Micheal Stone, for allowing me to modify his code for a noise vocoder, which forms the basis of much of the signal processing described in Chapters 3, 4, 5 and 6.

Michael Stone, Brian Glasberg, and Tom Baer for patiently answering my many questions about signal processing, programming and statistics.

Jo Robinson and Sarah Creeke for their support, advice and friendship throughout the last three years.

Simon Goldman, Helen Waller, Thomas Stainsby, Posy Marriage, Christian Füllgrabe, Anne Schlüeter, Paddy Boyle, and Debi Vickers for being excellent lab mates.

All of my subjects for the many hours of testing. Without them, none of my work would have been possible.

The Medical Research Council for funding my PhD and Deafness Research UK and the Acoustical Society of America for providing support to attend conferences.
# Contents

## 1 Introduction

1.1 Temporal information in complex sounds ........................................... 16  
1.2 The role of temporal fine structure information in pitch perception .......... 17  
1.3 Temporal fine structure information in speech .................................... 21  
1.4 Temporal fine structure information and cochlear hearing loss ................. 25  
1.5 Summary of chapters .............................................................................. 27  

## 2 Moderate cochlear hearing loss leads to a reduced ability to use temporal 

fine structure information ........................................................................ 30  

2.1 Introduction ............................................................................................ 30  
2.2 Rationale .................................................................................................. 35  
2.3 Method ...................................................................................................... 37  
2.3.1 Subjects ............................................................................................... 37  
2.3.2 Stimuli .................................................................................................. 39  
2.3.3 Signal generation .................................................................................. 43  
2.3.4 Procedure .............................................................................................. 44  
2.3.5 Statistics ............................................................................................... 47  
2.4 Results ....................................................................................................... 49  
2.5 Discussion .................................................................................................. 56  
2.5.1 Data for the normal-hearing subjects .................................................... 56  
2.5.2 Data for the hearing-impaired subjects ................................................ 57  
2.6 Discrimination of harmonic and frequency-shifted tones with spectral shap-

ing and components added in random phase ............................................. 60  
2.6.1 Rationale ............................................................................................... 60  
2.6.2 Method .................................................................................................. 60  
2.6.3 Results ................................................................................................... 61  
2.6.4 Discussion .............................................................................................. 63  
2.7 Conclusions ............................................................................................... 64
3 Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech 65

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>65</td>
</tr>
<tr>
<td>3.2 Rationale</td>
<td>71</td>
</tr>
<tr>
<td>3.3 Method</td>
<td>72</td>
</tr>
<tr>
<td>3.3.1 Subjects</td>
<td>72</td>
</tr>
<tr>
<td>3.3.2 Speech material</td>
<td>72</td>
</tr>
<tr>
<td>3.3.3 Processing and equipment</td>
<td>74</td>
</tr>
<tr>
<td>3.3.4 Procedure</td>
<td>75</td>
</tr>
<tr>
<td>3.3.5 Audibility calculations</td>
<td>76</td>
</tr>
<tr>
<td>3.3.6 Training</td>
<td>77</td>
</tr>
<tr>
<td>3.3.7 Testing</td>
<td>77</td>
</tr>
<tr>
<td>3.3.8 Analysis</td>
<td>79</td>
</tr>
<tr>
<td>3.4 Results</td>
<td>80</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>83</td>
</tr>
<tr>
<td>3.6 Experiment two</td>
<td>87</td>
</tr>
<tr>
<td>3.6.1 Rationale</td>
<td>87</td>
</tr>
<tr>
<td>3.6.2 Method</td>
<td>88</td>
</tr>
<tr>
<td>3.6.3 Results</td>
<td>89</td>
</tr>
<tr>
<td>3.6.4 Discussion</td>
<td>93</td>
</tr>
<tr>
<td>3.7 Conclusions</td>
<td>96</td>
</tr>
</tbody>
</table>

4 The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise 97

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>97</td>
</tr>
<tr>
<td>4.2 Methods</td>
<td>99</td>
</tr>
<tr>
<td>4.2.1 Subjects and materials</td>
<td>99</td>
</tr>
<tr>
<td>4.2.2 Processing and equipment</td>
<td>100</td>
</tr>
<tr>
<td>4.2.3 Training</td>
<td>100</td>
</tr>
<tr>
<td>4.2.4 Testing</td>
<td>101</td>
</tr>
<tr>
<td>4.2.5 Analysis</td>
<td>101</td>
</tr>
<tr>
<td>4.3 Results</td>
<td>102</td>
</tr>
<tr>
<td>4.4 Discussion</td>
<td>107</td>
</tr>
<tr>
<td>4.5 Conclusions</td>
<td>110</td>
</tr>
</tbody>
</table>

5 The importance of TFS information in speech at different spectral regions for normal-hearing and hearing-impaired subjects 111

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 The importance of TFS information in speech at different spectral regions for normal-hearing and hearing-impaired subjects</td>
<td>111</td>
</tr>
</tbody>
</table>
5.1 Introduction .................................................. 111
5.2 Method ....................................................... 114
  5.2.1 Subjects and Materials ................................. 114
  5.2.2 Processing and equipment ............................ 115
  5.2.3 Procedure ............................................... 116
5.3 Results ...................................................... 117
5.4 Discussion .................................................. 117
5.5 Experiment two ............................................. 121
  5.5.1 Rationale ................................................ 121
  5.5.2 Methods ................................................ 122
  5.5.3 Results ................................................ 124
  5.5.4 Discussion .............................................. 126
5.6 Experiment three: Measuring sensitivity to TFS .............. 131
  5.6.1 Methods ................................................ 131
  5.6.2 Results ................................................ 134
  5.6.3 Discussion .............................................. 134
5.7 Conclusions ................................................ 137

6  Processing speech to remove temporal envelope information 138
  6.1 Introduction .............................................. 138
  6.2 Processing ................................................ 142
  6.3 Analysis of stimuli ....................................... 144
  6.4 Experiment one .......................................... 152
    6.4.1 Rationale ............................................ 152
    6.4.2 Method .............................................. 153
    6.4.3 Results and discussion ............................ 156
  6.5 Experiment two .......................................... 158
    6.5.1 Rationale ............................................ 158
    6.5.2 Methods .............................................. 159
    6.5.3 Results .............................................. 160
    6.5.4 Discussion .......................................... 162
  6.6 Experiment three ......................................... 165
    6.6.1 Rationale ............................................ 165
    6.6.2 Method .............................................. 166
    6.6.3 Results and discussion ............................ 166
  6.7 Discussion ................................................. 168
  6.8 Conclusions ............................................... 170
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>TFS and envelope for a sample of narrow-band speech</td>
<td>18</td>
</tr>
<tr>
<td>1.2</td>
<td>Flow diagram of noise- or tone-vocoder processing to remove TFS information. Here, two channels are shown explicitly, but a vocoder can have any number of channels between one and N.</td>
<td>22</td>
</tr>
<tr>
<td>2.1</td>
<td>Schematic spectra of stimuli</td>
<td>36</td>
</tr>
<tr>
<td>2.2</td>
<td>Air conduction audiometric thresholds for the test ears of the hearing-impaired subjects</td>
<td>38</td>
</tr>
<tr>
<td>2.3</td>
<td>Excitation patterns for SHAPED stimuli</td>
<td>42</td>
</tr>
<tr>
<td>2.4</td>
<td>Mean $d'$ values for normal-hearing subjects for discrimination of harmonic and frequency-shifted complex tones</td>
<td>50</td>
</tr>
<tr>
<td>2.5</td>
<td>$d'$ values for individual hearing-impaired subjects for discrimination of harmonic and frequency-shifted complex tones</td>
<td>51</td>
</tr>
<tr>
<td>2.6</td>
<td>$d'$ values for F0 discrimination by normal-hearing and hearing-impaired subjects</td>
<td>55</td>
</tr>
<tr>
<td>2.7</td>
<td>Mean $d'$ values for discrimination of harmonic and frequency-shifted tones by normal-hearing subjects with components added in sine or random phase</td>
<td>62</td>
</tr>
<tr>
<td>3.1</td>
<td>Air conduction audiometric thresholds for the test ears of the hearing-impaired subjects</td>
<td>73</td>
</tr>
<tr>
<td>3.2</td>
<td>Excitation levels of the target speech and excitation levels at threshold for individual hearing-impaired subjects</td>
<td>78</td>
</tr>
<tr>
<td>3.3</td>
<td>Mean SRTs for normal-hearing and hearing-impaired subjects, plotted as a function of CO/N</td>
<td>81</td>
</tr>
<tr>
<td>3.4</td>
<td>Individual SRTs for the hearing-impaired subjects, plotted as a function of CO/N</td>
<td>84</td>
</tr>
<tr>
<td>3.5</td>
<td>Mean SRTs for normal-hearing subjects and hearing-impaired subjects for tone vocoded stimuli, plotted as a function of CO/N</td>
<td>90</td>
</tr>
<tr>
<td>4.1</td>
<td>Benefit of adding TFS information as measured by the SRT relative to that for CO=0, for steady and modulated noise</td>
<td>102</td>
</tr>
</tbody>
</table>
4.2 Masking release as a function of CO

4.3 Psychometric functions for each value of CO, for steady and modulated noise

5.1 SRTs plotted as a function of CO for TFS-high and TFS-low conditions

5.2 Audiometric thresholds of the test ears of the hearing-impaired subjects

5.3 Mean results for the normal-hearing subjects

5.4 Mean results for the hearing-impaired subjects

5.5 Individual results for the hearing-impaired subjects

5.6 Individual results for the hearing-impaired subjects and mean results for the normal-hearing subjects for experiments two and three

6.1 The amplitude and instantaneous frequency of the channel signal from the 6th channel of a 12-channel filter bank with and without the addition of low-noise noise

6.2 Mean correlations between the envelope and TFS of the TFS-processed and original signals at the output of 25 ERB_N-wide gammatone filters for TFS speech processed using the TFS-LNN and TFS-no-LNN schemes

6.3 Mean correlations between the envelopes of unprocessed and TFS-processed speech at the outputs of 25 1-ERB_N-wide gammatone filters. TFS speech was processed using the TFS-no-LNN scheme, with and without re-filtering after TFS extraction

6.4 Mean correlations between the logarithm of the envelopes of the TFS-processed and original signals at the output of 25 ERB_N-wide gammatone filters for TFS speech processed using the TFS-LNN and TFS-no-LNN schemes

6.5 Flow diagram showing the processing scheme used for experiment one

6.6 Percent correct scores for speech processed using the TFS-LNN scheme and then vocoded with a tone vocoder to assess the usability of recovered-envelope cues

6.7 Percent correct scores for TFS speech processed using the TFS-LNN and TFS-no-LNN schemes for N=3, 6 and 12

6.8 Excitation patterns evoked by white noise with the same frequency spectra as the channel signals when N=3, 6 and 12

6.9 Distribution of levels for channel signals derived from ASL and IEEE sentence lists

6.10 Percent correct scores for speech processed using the TFS-LNN and TFS-no-LNN schemes as a function of run number
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Largest differences in excitation level between harmonic and frequency-shifted SHAPED tones</td>
<td>41</td>
</tr>
<tr>
<td>2.2</td>
<td>Conditions for which the non-adaptive procedure was used for the normal-hearing subjects</td>
<td>45</td>
</tr>
<tr>
<td>2.3</td>
<td>Conditions for which the non-adaptive procedure was used for the hearing-impaired subjects</td>
<td>45</td>
</tr>
<tr>
<td>2.4</td>
<td>Summary of results for the normal-hearing subjects for conditions where all subjects completed the adaptive procedure</td>
<td>48</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary of the ANOVA outcomes for hearing-impaired subjects tested with SHAPED and NON-SHAPED stimuli</td>
<td>53</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary of conditions for which performance was not significantly better than chance</td>
<td>54</td>
</tr>
<tr>
<td>3.1</td>
<td>Differences between the mean SRTs measured with different values of CO for normal-hearing subjects in experiment one</td>
<td>82</td>
</tr>
<tr>
<td>3.2</td>
<td>As Table 3.1 but for the hearing-impaired subjects</td>
<td>82</td>
</tr>
<tr>
<td>3.3</td>
<td>As Table 3.1 but for experiment two, which used two values of N</td>
<td>92</td>
</tr>
<tr>
<td>3.4</td>
<td>As Table 3.3 but for hearing-impaired subjects</td>
<td>92</td>
</tr>
<tr>
<td>3.5</td>
<td>Differences between mean SRTs measured for N=8 and N=16 for each value of CO/N, for normal-hearing and hearing-impaired subjects in experiment two</td>
<td>92</td>
</tr>
<tr>
<td>4.1</td>
<td>Differences between mean SRTs measured with different pairs of values of CO for steady and modulated noise</td>
<td>104</td>
</tr>
<tr>
<td>4.2</td>
<td>Mean slopes of the psychometric functions estimated for each condition</td>
<td>106</td>
</tr>
<tr>
<td>4.3</td>
<td>Masking release and benefit from additional TFS information expressed in percentage points</td>
<td>108</td>
</tr>
<tr>
<td>5.1</td>
<td>Differences between mean SRTs measured with different pairs of values of CO for the TFS-high and TFS-low conditions</td>
<td>118</td>
</tr>
</tbody>
</table>
6.1 Mean SDs of the instantaneous frequency in N channels in low-level signal portions of concatenated ASL sentences. .................................................. 152

6.2 Results of linear regression analysis for the intelligibility of TFS-processed sentences with TFS-LNN and TFS-no-LNN processing schemes ............ 168
List of abbreviations

**ASL**  Adaptive sentence list.

**ANOVA**  Analysis of variance.

**BKB**  Bench-Kowel-Bamford.

**CRM**  Co-ordinate response matrix.

**ERBN**  The equivalent rectangular bandwidth of the auditory filter for young, normally hearing subjects.

**FIR**  Finite impulse response.

**HI**  Hearing impaired.

**F0**  Fundamental frequency.

**F0DL**  Fundamental frequency difference limen.

**LNN**  Low-noise noise.

**LSD**  Least significant difference.

**NH**  Normal hearing.

**rms**  Root mean squared

**SBR**  Signal to background ratio.

**SD**  Standard deviation.

**SNR**  Signal to noise ratio.

**SRT**  Speech reception threshold.

**TFS**  Temporal fine structure.

**VCV**  Vowel-consonant-vowel.
Chapter 1

Introduction

1.1 Temporal information in complex sounds

The cochlea acts as a frequency analyser, splitting complex sounds into their component frequencies. Consequently, information about the frequency content of a sound could be derived from the pattern of excitation along the basilar membrane (Helmholtz, 1863; von Békésy, 1960; Zwicker, 1970). Additionally, a temporal frequency-coding scheme is believed to operate in mammals when the cochlea is undamaged. Physiological recordings have shown that auditory nerve fibres fire repeatedly at the same phase of a sinusoidal waveform, provided that the repetition rate of that waveform is not too high, allowing the repetition rate of the waveform to be determined from the times between neural spikes (Rose et al., 1967). The frequency at which this phase locking breaks down varies between species (Palmer and Russell, 1986). In humans with normal hearing, it has often been assumed that this limit is between 4000 and 5000 Hz. This has not been measured physiologically but is consistent with measurements in other species and with psychophysical data, which show a change in this region, suggesting a difference in coding mechanism (Moore, 2003). For example, detection of low-rate (2 Hz) frequency modulation, which is thought to depend on detection of changes in the pattern of phase-locking, worsens for frequencies above about 4000 Hz (Moore and Sek, 1996). Pure tone frequency discrimination also worsens above 4000 Hz, and listeners with absolute pitch are poor at naming the notes of tones with frequencies above 4000-5000 Hz (Ohgushi and Hatoh, 1991).

It is not certain, however, that temporal information does not play some role at frequencies above 5000 Hz. Heinz et al. (2001b) showed by computational modelling that psychophysical frequency discrimination performance was better predicted if temporal information in the auditory nerve was included than if only information about the place of excitation was included, for frequencies up to at least 10,000 Hz.

Temporal information in narrowband sounds can be divided into two categories. ‘Tem-
poral fine structure’ (TFS) refers to the rapidly fluctuating variations in amplitude of the waveform at a rate close to the centre frequency of the band. ‘Temporal envelope’ refers to slower amplitude modulations superimposed on this TFS. This categorisation is illustrated in Figure 1.1, which shows the TFS and envelope in a short section of speech filtered into a narrow frequency band.

Figure 1.1: TFS (thin solid line) and envelope (dashed line) for a sample of narrow-band speech. The figure was produced by passing a recording of the word ‘down’ through an array of eight contiguous bandpass filters with centre frequencies arranged on an equal-ERB scale from 100-8000 Hz [Glasberg and Moore 1990]. The output of the 3rd filter in the array is plotted. The envelope of this signal was calculated by finding the absolute value of the analytic signal, using the Hilbert transform.

TFS information could be carried in the pattern of phase locking to the stimulus waveform, and temporal envelope by the changes in firing rate over time. Alternatively, neurones could phase lock to peaks in the temporal envelope rather than peaks in the TFS.

1.2 The role of temporal fine structure information in pitch perception

The waveform of a pure tone (a single sine wave) has a flat envelope. A pitch is heard which varies according to the frequency of the sine wave, with higher-frequency sine waves evoking the perception of a higher pitch. The frequency of the sine wave could
be coded by the place of maximum excitation on the basilar membrane, higher-frequency
sine waves evoking excitation patterns which peak in more basal regions (von Békésy
1960). However, a temporal frequency-coding scheme is believed to operate in mammals
at low frequencies, where phase locking is robust.

It seems likely that a similar strategy is used to code the frequencies of resolved, lower
harmonics in complex tones, either to extract the frequencies of constituent tones to be
fed into a central ‘pattern recogniser’ (Goldstein 1973; Terhardt, 1974) or to establish
a common time interval across harmonics that can be used to deduce the fundamental
frequency (F0) of the complex (Moore 1982). Unresolved upper harmonics interact on
the basilar membrane to form a complex periodic waveform that repeats with a period
of 1/F0. Complexes with only unresolved harmonics have a pitch that corresponds to
their F0, so this period must be extracted by temporal mechanisms (Schouten 1940). It
remains unclear whether two different pitch mechanisms are responsible for the extraction
of pitch from the resolved and unresolved harmonics of complex tones, or whether there
is one mechanism that extracts information from all components (Gockel et al. 2004;
Shackleton and Carlyon 1994; Carlyon and Shackleton 1994). In either case, for a pitch
corresponding to F0 (‘low pitch’) to be perceived from unresolved components alone, some
form of periodicity-extraction mechanism is required.

Periodicity extraction from a complex waveform could occur by two mechanisms: mea-
surement of the time period between corresponding peaks in the envelope of the waveform,
or measurement of the time interval between corresponding peaks in the TFS. The use of
TFS cues would allow greater accuracy in coding waveform periodicity, but would also re-
quire greater temporal resolution by the hair cells and auditory nerve fibres of the cochlea,
as well as by more central mechanisms.

de Boer (1956a) and Schouten et al. (1962) measured the shift in pitch that occurred
when all components in a harmonic complex tone were shifted upwards by the same
amount in Hz. The shifted and harmonic complexes had the same envelope repetition rate,
but different TFS. They found that the pitch of the shifted complex was higher than that
of the harmonic complex, leading to the suggestion that TFS information was important
in determining pitch. However, shifting all components in a complex upwards in frequency
results in an upward shift in the excitation pattern, as well as a change in the TFS. It
is possible that the pitch shifts observed by de Boer (1956a) and Schouten et al. (1962)
were a result of such shifts in the excitation patterns of the complexes. To test whether
this was the case, Moore and Moore (2003b) measured the pitch shift between harmonic
and frequency-shifted complexes that had components that were filtered into the same
spectral region to minimise the differences in excitation patterns evoked by the shifted
and harmonic complexes. The filter was centred on the 11th harmonic (intermediate
condition) or the 16th (unresolved condition). Pitch shifts were found for complexes with components that were intermediate, but not very high relative to F0, suggesting that TFS information was important in the perception of pitch for complexes with components that have intermediate harmonic numbers (8-11), but that for complexes with only high harmonics, envelope repetition rate, but not TFS is important for pitch perception.

An alternative explanation for the pitch shift observed in the intermediate condition by Moore and Moore (2003b) is that subjects matched partially-resolved harmonics in the test and matching stimuli rather than using TFS cues. Other authors have presented evidence that harmonics up to the 11th or 12th could be resolved (Bernstein and Oxenham 2003), although previous work suggested that only harmonics below the eighth were resolvable (Plomp, 1964; Moore and Ohgushi, 1993).

To test whether the use of TFS information could explain the pitch shifts observed by Moore and Moore (2003b) for intermediate harmonic numbers (8-11), Moore et al. (2006a) measured difference limens for the F0 (F0DLs) of three-component complex tones for normal-hearing subjects. Complexes had varying lowest harmonic numbers (N), but the same centre component frequency (2000 Hz), which meant that the F0 of the complex also varied. For example, for a complex for which N=8, the F0 was equal to 222.2 Hz and the components were 1777.8 Hz, 2000 Hz and 2222.2 Hz. Discrimination was tested when all components were presented with the same starting phase (cosine phase) and when the phase of the central component was shifted by 90 degrees (alternating phase). The relative phase of the components was only expected to affect discrimination when components were unresolved. For conditions where harmonics interacted on the basilar membrane, the cosine-phase stimuli would have had a more peaky temporal envelope, and so performance based on envelope or TFS information was expected to be better. For complexes containing harmonics up to the 7th, F0DLs were low, and were unaffected by component phase, suggesting that components were resolved. For complexes containing only harmonics above the 14th, F0DLs were larger, and performance was significantly better in the cosine-phase condition than the alternating-phase condition. For intermediate harmonic numbers, there was a transition in performance; performance worsened as harmonic number increased. In this intermediate region, there was a significant effect of component phase, suggesting that these components were unresolved, and that performance was not based on the use of partially-resolved harmonics but on the use of TFS information.
1.3 Temporal fine structure information in speech

Information in speech is redundant. For normal-hearing people, this means that the signal is robust to corruption, and that speech remains intelligible under adverse listening conditions, such as in high levels of background noise. In the normal auditory system, a complex sound like speech is filtered into frequency channels on the basilar membrane. The signal at a given place can be considered as a time-varying envelope superimposed on the more rapid fluctuations of a carrier (TFS) whose rate depends partly on the centre frequency and bandwidth of the channel. The relative envelope magnitude across channels conveys information about the spectral shape of the signal and changes in the relative envelope magnitude indicate how the short-term spectrum changes over time. The TFS could carry information both about the F0 of the sound (when it is periodic) and about its short-term spectrum. For example, if at a particular time there is a formant centred at frequency fx (and hence one or more relatively intense stimulus components near fx), then, provided that these stimuli components are unresolved, channels centred close to fx will show TFS synchronised to fx, and this will be reflected in the patterns of phase locking in those channels (Young and Sachs, 1979).

Figure 1.2: Flow diagram of noise- or tone-vocoder processing to remove TFS information. Here, two channels are shown explicitly, but a vocoder can have any number of channels between one and N.

The relative importance of envelope and TFS information in speech has been studied using noise- and tone-vocoder processing techniques, where a speech signal is filtered into a number of frequency channels, and the original TFS in each channel signal is replaced with either noise or a sine wave (Dudley, 1939; Shannon et al., 1995). A flow diagram
illustrating vocoder processing is shown in Figure 1.2. The speech signal is first filtered using an array of bandpass filters, and the temporal envelope of each of the channel signals is extracted, either by finding the absolute values of the Hilbert transform of each channel signal, or by rectifying and then low-pass filtering each channel signal. The channel envelopes are then multiplied by carrier signals. For a noise vocoder this is usually broadband noise, and for a tone vocoder it is usually a sine wave with a frequency equal to the centre frequency of the channel. Each modulated carrier is then typically filtered to restrict its spectrum to the original channel bandwidth and the modulated carriers are combined to give the vocoded signal. The vocoded signal has little of the original TFS information, and also has reduced spectral information compared with the original signal. The amount of spectral information that is preserved depends on the number and spacing of the analysis channels, with narrower spacing preserving more spectral information. If the number of channels and their spacing is similar to the number and spacing of auditory filters in the peripheral auditory system, the vocoder processing has little effect on spectral information available to the central auditory system, allowing the effects of removing TFS information alone to be assessed.

Shannon et al. (1995) measured the intelligibility of noise-vocoded sentences and showed that removal of TFS information had little effect on speech intelligibility in quiet, even when spectral information was much reduced. However, when similar processing was applied to a speech signal with a competing background, intelligibility was much lower than for the unprocessed signal (Dorman et al. 1998, Fu et al. 1998, Qin and Oxenham 2003, Stone and Moore 2003). This reduction in intelligibility is likely to be partly due to the reduced spectral information available in the vocoded signal. However, Başkent (2006) showed that intelligibility was reduced for vocoded speech with a competing background even when the analysis channels were similarly spaced to auditory filters in the normal auditory system, suggesting that TFS information is important for listening to speech in a competing background.

TFS information could improve performance when listening in a competing background by allowing more accurate identification of the F0 of the target speech. As discussed in Section 1.2, TFS information is thought to be important for accurate estimation of the frequencies of resolved harmonics, and may also be important for coding the periodicity of tones containing unresolved harmonics. Both mechanisms would allow the F0 of target speech to be identified more accurately. In tonal languages, F0 carries information about word meaning, and even in non-tonal languages F0 can convey linguistic information - for example, distinguishing a question from a statement. F0 may be important for separating target and background into different auditory streams, especially when the background is a competing talker (Bregman 1990). For normal-hearing listeners, intelli-
gibility improves as the F0s of the target and competing talkers become more different, suggesting that the accurate identification of F0 may improve intelligibility (Brokx and Nooteboom 1982; Stickney et al. 2007). Subjects listening with a cochlear implant, and normal-hearing subjects listening to a vocoder do not benefit as much as normal-hearing listeners from a difference in F0 between target and competing talkers (Stickney et al. 2004). Such listeners have little or no access to TFS information, and this is likely to partly account for this reduction in benefit. Stickney et al. (2007) showed that by reintroducing some simplified TFS information to a vocoder, using a ‘frequency-amplitude modulation encoding’ strategy (known as FAME), the benefit from the difference in F0 of target and competing speakers was partially restored.

It has been suggested that TFS information is particularly important when listening in a background that fluctuates. Normal-hearing listeners perform better when listening to speech in a fluctuating background than when listening in a steady background, when the signal-to-background ratio is the same (Festen and Plomp 1990). This improvement in performance when listening in a fluctuating background is known as ‘masking release’. A similar psychophysical phenomenon is observed; the detection threshold for a sine wave is lower in the presence of a fluctuating masker than in a steady masker with the same overall level (Buus 1985; Moore and Glasberg 1987). Moore and Glasberg (1987) measured the threshold for detecting a sine wave in the presence of a single sine wave masker and a double sine wave masker. Both the single and double sine wave maskers were centred at frequencies 1.8 times lower than that of the target sine wave. The double sine-wave masker had an envelope that fluctuated in amplitude because of ‘beating’ between the two components. The rate of beating was determined by the spacing between the two components. When the masker and target were presented at low frequencies, the detection threshold was much lower for the signal in the double sine wave masker. At high frequencies, the difference between the two conditions was much smaller. Moore and Glasberg (1987) suggested that this difference could arise because, at low frequencies, a signal in the dips of a fluctuating masker could be detected as a difference in the TFS when the signal was present. They argued that this cue would not be available at high frequencies where phase locking is believed to be weak or absent, which could account for the reduced masking release. A similar mechanism has been proposed to account for speech masking release (Lorenzi et al. 2006a). When listening to speech in a fluctuating background, such as a competing talker, TFS may be important in identifying signal portions in which there is a target speech signal in the dips of the fluctuating background.
1.4 Temporal fine structure information and cochlear hearing loss

People with raised audiometric thresholds caused by cochlear damage often suffer deficits in the processing of complex sounds, even when signals are presented well above the detection threshold. These deficits can be partly attributed to damage to outer hair cells resulting in a broadening of the auditory filters (Glasberg and Moore, 1986). Auditory filter broadening reduces the resolving power of the cochlea, leading to poorer discrimination and increased masking by noise that is remote in frequency from the target signal. However, auditory filter broadening cannot explain all deficits associated with sensorineural hearing impairment.

Moore and Peters (1992) found that there was only a weak correlation between pure-tone frequency discrimination ability and auditory filter sharpness at a particular frequency. The ability to understand speech in noise is poorly predicted by audiometric thresholds, suggesting that a number of cochlear pathologies may contribute to deficits other than a loss in sensitivity. One animal study suggested that cochlear damage may lead to a deficit in phase locking, which would impair the ability to use TFS information (Woolf et al., 1981). However, another study showed no phase-locking deficits in guinea pigs with kanamycin-induced outer hair cell damage (Harrison and Evans, 1979b). These apparently contradictory results may reflect a species difference or a difference in the methods used to induce cochlear damage. It is unclear whether humans with cochlear hearing loss suffer phase locking deficits, but if this were the case, then a reduced ability to use TFS information would be expected.

A phase locking deficit is not the only pathology that may result in a reduced ability to make use of TFS information, however. A change in cochlear tuning caused by outer hair cell damage results in a change in the phase response properties of the auditory filters. It has been suggested that TFS information may be extracted by cross-correlation of auditory filter outputs - a particular delay in response between closely spaced places along the basilar membrane could indicate a particular signal frequency (Shamma, 1985; de Cheveigne and Pressnitzer, 2006). Changes in filter phase response due to a loss of outer hair cells would disrupt this mechanism and reduce the ability to extract TFS information.

A third reason that subjects with cochlear hearing loss may have a reduced ability to use TFS information is that they usually have broader auditory filters than normal-hearing subjects (Glasberg and Moore, 1986). As well as resulting in deficits associated with poorer frequency resolution, broader filters could impair the ability to use TFS information. In a broadband sound with many components, such as speech, the waveform.
at the output of a broader filter is much more complex than that at the output of a narrower filter centred at the same frequency. The TFS information at the output of such a broad filter may be un-interpretable by the central auditory system because of its complexity.

A number of psychoacoustic studies suggest that hearing-impaired subjects have a reduced ability to use TFS information (Buss et al., 2004; Moore and Moore, 2003a; Moore et al., 2006b). Moore and Moore (2003a) showed that hearing-impaired subjects relied more than normal-hearing subjects on spectral cues to discriminate the F0 of complex tones. They also showed that the hearing-impaired subjects that they tested had similar difference limens for tones with intermediate harmonics and high harmonics. This is in contrast to results for normal-hearing subjects, who showed smaller difference limens for the intermediate-harmonic conditions. The authors suggested that this difference could occur because the hearing-impaired subjects could not use TFS information and relied instead on envelope cues, which remained similar as harmonic number increased.

Moore et al. (2006b) provided further evidence to suggest that the ability to use TFS information is reduced in hearing-impaired subjects. F0DLs for complex tones were measured for hearing-impaired subjects using the same procedures and stimuli as Moore et al. (2006a). For a centre frequency of 2000 Hz, hearing-impaired subjects showed the same pattern of responses as normal-hearing subjects showed for a 5000-Hz centre frequency, where phase locking is believed to be reduced. The authors interpreted this result as evidence that hearing-impaired subjects were not using TFS information even at frequencies where phase locking is believed to be robust for normal-hearing subjects. Lacher-Fougère and Demany (2005) showed that some hearing-impaired subjects were sensitive to TFS at low frequencies, as they could detect inter-aural phase differences, which would require some TFS sensitivity to the monaural signals. However, detection of differences in inter-aural phase was poorer for hearing-impaired subjects than for normal-hearing subjects, suggesting some deficit in TFS processing even at low frequencies.

A deficit in TFS processing could partially account for the reduced ability of hearing-impaired subjects to understand speech in background noise, even when the speech is at a level that is audible. Hearing-impaired subjects show less masking release than normal-hearing subjects (Festen and Plomp, 1990; Peters et al., 1998; Nelson et al., 2003; Lorenzi et al., 2006a) (see Section 1.3 for a discussion of masking release in normal-hearing subjects). If masking release in speech is partially dependent on TFS information, a deficit in TFS processing could explain the reduced masking release for speech for hearing-impaired subjects.
1.5 Summary of chapters

This thesis describes a series of experiments investigating TFS processing for normal-hearing and hearing-impaired subjects. Each chapter is self-contained, so that it can be read without reference to the rest of the thesis. Consequently, information from this introductory chapter is sometimes repeated in the introductory sections of each chapter, when that information is necessary to introduce the questions that each chapter addresses.

In Chapter 2, the sensitivity of normal-hearing and hearing-impaired subjects to TFS information was measured using a psychophysical task. Subjects were required to discriminate harmonic complexes from corresponding frequency-shifted complexes, in which each component was shifted upwards by the same amount in Hz. Such complexes had the same envelope repetition rate, but different TFS. The harmonic and frequency-shifted complexes were filtered into the same spectral region using a bandpass filter to minimise changes in excitation patterns. The F0 of the harmonic complexes, and the centre frequency of the bandpass filter were varied. When the bandpass filter was centred at a frequency of 11F0 (for which it was assumed that all audible components in the passband were unresolved), normal-hearing subjects could discriminate the harmonic and frequency-shifted complexes on the basis of their TFS for all of the F0s tested (100, 200 and 400 Hz). Most of the hearing-impaired subjects could not discriminate the harmonic and frequency-shifted complexes for a centre frequency of 11F0, and scored no better than chance. It was concluded that moderate cochlear hearing loss leads to an insensitivity to TFS information in complex tones with unresolved harmonics.

To test whether this insensitivity to TFS in complex tones reflected a more general insensitivity to TFS in complex sounds, the ability of normal-hearing and hearing-impaired subjects to use TFS information to identify speech in a competing talker background was tested in Chapter 3. TFS information was progressively introduced to a vocoded signal by including unprocessed information in the low-frequency channels up to a ‘cut-off channel’. The cut-off channel was varied and SRTs were measured. Normal-hearing listeners benefitted much more than hearing-impaired listeners from the TFS information, although the results for the hearing-impaired subjects were variable, with one subject benefitting as much as the normal-hearing listeners, and some subjects showing very little or no benefit.

We suggested that the reason that the normal-hearing listeners benefitted from TFS information in speech when listening in a competing talker background in Chapter 3 may be because TFS information could be important for identifying when and what signal was present in the dips of a fluctuating background. This hypothesis was tested in Chapter 4. A similar method was used as in Chapter 3. SRTs were measured for signals processed to contain variable amounts of TFS, by including unprocessed information in channels up to and including a cut-off channel. Channels above the cut-off channel were
tone vocoded. The cut-off channel was varied, and SRTs were measured in both steady and modulated noise. SRTs decreased for both noise types as the cut-off channel was increased and more TFS was included, but subjects benefitted more from the additional TFS when listening in a modulated-noise background than when listening in a steady-noise background, supporting the idea that TFS information is important for dip listening.

In Chapter 5, the work of Chapters 3 and 4 was extended to more accurately identify the spectral regions for which TFS information is important for speech perception in a competing talker background for normal-hearing and hearing-impaired subjects. The benefit of adding TFS information to five 6-ERB wide spectral regions was measured by comparing SRTs when there was no information in the band (there was a spectral gap), tone-vocoded information in the band (envelope information only), and when there was unprocessed information in the band (envelope and TFS information). Information outside of the spectral region under test was tone-vocoded using 1-ERB wide analysis bands to remove TFS information. Normal-hearing subjects benefitted similarly from TFS information added in each of the five spectral regions that were tested (spanning 100-8000 Hz), suggesting that TFS information is important for cues other than representing the F0 of the target speaker. As a group, the hearing-impaired subjects did not show any significant benefit when TFS information was added to a single spectral region. However, the hearing-impaired subjects did show significantly better SRTs when the signal was entirely unprocessed than when it was entirely vocoded. The amount of benefit was significantly correlated with a psychophysical measure of TFS sensitivity.

Finally, in Chapter 6, a method for processing speech signals to remove (as far as possible) temporal envelope information was investigated. If such a signal could be produced it could be used to investigate speech intelligibility based mainly on TFS cues in different listening conditions. We proposed that by adding 'low-noise' noise (Hartmann and Pumplin, 1988) to the speech signal at a low level before application of the processing to remove temporal envelope information, fewer spurious large excursions in instantaneous frequency in the output signal would occur. Simple sentences were more intelligible when low-noise noise was added before processing, and subjects reached a stable level of performance more quickly when low-noise noise was added before processing.

The final chapter summarises the main findings of each chapter, and discusses possibilities for future research.
Chapter 2

Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information

A version of this chapter was published as: Hopkins, K., and Moore, B. C. J. (2007). ”Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information,” J. Acoust. Soc. Am. 122, 1055-1068.

2.1 Introduction

The frequency of a sine wave could be coded by the place of excitation on the basilar membrane (von Békésy, 1960; Zwislocki and Nguyen, 1999) or by the temporal information in the auditory nerve (phase locking), which is widely believed to be available for frequencies up to about 5000 Hz (Rose et al., 1967), although the exact upper limit varies across species (Palmer and Russell, 1986). Psychoacoustic evidence from humans suggests that the temporal mechanism plays an important role for frequencies up to about 5000 Hz (Moore, 1973, 2003; Plack and Oxenham, 2005), although Heinz et al. (2001b) have shown computationally that human psychophysical frequency discrimination performance is better predicted by modelling if temporal information in auditory nerve firing patterns is included for frequencies up to at least 10,000 Hz.

It seems likely that temporal information is used to code the frequencies of resolved, lower harmonics in complex tones (Hartmann and Doty, 1996; Moore et al., 1984, 2006c), either to allow extraction of the frequencies of constituent tones to be fed into a central pattern recogniser (Goldstein, 1973; Terhardt, 1974) or to establish a common time interval across harmonics that can be used to deduce the fundamental frequency (F0) of the
complex (Moore, 1982; van Noorden, 1982). Unresolved harmonics interact on the basilar membrane to form a complex periodic waveform with a period of 1/F0. Complexes with only unresolved harmonics have a pitch corresponding to their F0, so this period must be extracted by temporal mechanisms. The temporal information has two forms. Temporal fine structure (TFS) refers to the rapidly fluctuating variations in the amplitude of the waveform. Envelope refers to the slower modulation superimposed on this TFS, which occurs at a rate equal to F0. Periodicity extraction from a complex waveform could occur by measurement of the time period between corresponding peaks in the envelope of the waveform, or by measurement of the time interval between corresponding peaks in the TFS. The use of TFS cues would allow greater accuracy in coding waveform periodicity (Moore et al., 2006a), but would require greater precision in temporal coding. For waveforms with TFS frequencies greater than the limit of phase locking (commonly believed to be 4000-5000 Hz in humans), periodicity extraction must be on the basis of envelope repetition rate alone. For lower TFS frequencies, either coding strategy is possible.

To assess the role of TFS, de Boer (1956b) and Schouten et al. (1962) obtained pitch matches to frequency-shifted complex tones, derived from harmonic complexes by shifting each component upwards (or downwards) by the same amount in Hz. The envelope repetition rate was the same as for the original harmonic complex, but the time intervals between peaks in the TFS were smaller (or larger) (for an illustration of the effect of frequency shifting on the envelope and TFS of harmonic tones, see: Moore (2003, Fig. 6.7, page 214)). The matched pitch was found to shift with the shift of the component frequencies, suggesting that TFS plays an important role in the pitch perception of complex tones. However, Moore and Moore (2003b) argued that these results might have been influenced by shifts in the spectrum or excitation pattern produced by the frequency shift of the components. To reduce this effect, Moore and Moore (2003b) conducted a similar experiment to those of de Boer (1956b) and Schouten et al. (1962), but both the test and matching tones were spectrally shaped so that they evoked very similar excitation patterns on the basilar membrane when no resolved components were present. They used complexes with components that were resolved (low relative to F0), unresolved (high relative to F0), or intermediate, but probably containing mainly unresolved harmonics (Moore et al., 2006a). For the resolved condition, matching tones were made up of harmonics in a different spectral region to the test tone to prevent comparisons between the frequencies of individual components. For the intermediate and resolved conditions, subjects matched the inharmonic tones (with upward-shifted components) to harmonic tones with higher envelope repetition rates. This suggests that TFS information does play a role in the pitch perception of complex tones with intermediate harmonic numbers, as the complexes in the intermediate condition contained harmonics that were probably unresolved and
no pitch shift would be expected if only envelope cues were used. In contrast, subjects matched harmonic and inharmonic tones with the same envelope repetition rate for the unresolved condition. This suggests that only envelope cues were used to compare the pitch of these tones, and that TFS cues did not play a role, even when TFS fluctuations were below the frequency at which phase locking is thought to break down.

An alternative explanation for the pitch shifts observed in the intermediate condition by Moore and Moore (2003b) is that subjects matched the partially resolved harmonics in the test and matching stimuli rather than using TFS cues. One paper has presented evidence that harmonics up to the 11th or 12th may be resolved (Bernstein and Oxenham, 2003), although other work suggests that only harmonics below the eighth are resolvable (Plomp, 1964; Moore and Ohgushi, 1993; Moore et al., 2006c). To address this issue, Moore et al. (2006a) measured difference limens for the F0 (F0DLs) of three-component complex tones for normal-hearing subjects. The nominal frequency of the centre component was fixed, but the harmonic number of the lowest component, N, was varied. For example, for a centre frequency of 2000 Hz, a three-component complex with N=8 would have components with frequencies 1777.8, 2000 and 2222.2 Hz (F0=222.2 Hz). A complex with N=9 would have components of 1800, 2000 and 2200 Hz (F0=200 Hz). Discrimination was tested when all components had the same starting phase (cosine phase) and when the phase of the central component was shifted by 90 degrees (alternating phase). It was argued that the presence of a phase effect would indicate that the components were not resolved (Moore, 1977; Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Bernstein and Oxenham, 2005).

For complexes with N equal to six or seven, there was no significant phase effect. However, F0DLs were smaller for complexes presented in cosine phase than for those presented in alternating phase when N was greater than eight. The phase effect suggests that components were not resolved for complexes with N of eight or more. This suggests in turn that the pitch shifts observed by Moore and Moore (2003b) for complexes with intermediate harmonic numbers was due to the use of TFS information rather than comparison of excitation patterns of partially resolved components.

Listeners with cochlear hearing loss usually show a poor ability to discriminate the pitch of complex sounds, even when the sounds are presented well above the detection threshold (Moore and Carlyon, 2005). These deficits can be partly attributed to damage to outer hair cells resulting in a broadening of the auditory filters (Pick et al., 1977; Liberman and Kiang, 1978; Glasberg and Moore, 1986). This would lead to a reduced ability to resolve the partials in complex tones and to more complex waveforms at the outputs of the auditory filters (Rosen, 1987). Auditory filter broadening cannot, however, explain all deficits associated with cochlear hearing loss. Moore and Peters (1992) found
only a weak correlation between pure-tone frequency discrimination and auditory filter sharpness at a particular frequency, though recent work by Bernstein and Oxenham (2006) showed a correlation between frequency selectivity in hearing-impaired subjects, and the position of the transition in F0 discrimination ability from good to poor as the number of the lowest harmonic was increased from low values towards higher values.

One animal study suggested that cochlear damage may lead to a deficit in phase locking, which would impair the ability to use TFS information (Woolf et al., 1981). However, another study showed no phase-locking deficits in guinea pigs with kanamycin-induced outer hair cell damage (Harrison and Evans, 1979b). These contradictory results may reflect a species difference or a difference in the methods used to induce cochlear damage. It is unclear whether humans with hearing impairments suffer phase locking deficits, but if they do, then a reduced ability to use TFS information would be expected. A phase locking deficit is not the only pathology that may result in a reduced ability to make use of TFS information, however. A change in cochlear tuning caused by outer hair cell damage can result in a change in the phase response properties of auditory filters. It has been suggested that TFS information may be extracted by cross correlation of auditory filter outputs. A particular phase shift in the response between adjacent places along the basilar membrane would indicate a particular signal frequency (Loeb et al., 1983; Shamma, 1985; de Cheveigné and Pressnitzer, 2006). Changes in filter phase response through a loss of the active mechanism might disrupt this cue and reduce the ability to extract TFS information.

A number of psychoacoustic and speech perceptual studies suggest that hearing-impaired listeners have a reduced ability to use TFS information (Lacher-Fougère and Demany, 1998, 2005; Moore and Skrodzka, 2002; Moore and Moore, 2003a; Buss et al., 2004; Lorenzi et al., 2006a; Moore et al., 2006b). Moore and Moore (2003a) showed that hearing-impaired subjects relied more than normal-hearing subjects on spectral cues to discriminate the F0 of complex tones. They also showed that hearing-impaired subjects had similar F0DLs for tones with intermediate harmonics and with high harmonics. This contrasts with results for normal-hearing subjects, where smaller F0DLs were found for tones with intermediate harmonic numbers. Moore and Moore (2003a) suggested that this difference might have occurred because the hearing-impaired subjects could not use TFS information and relied instead on envelope or spectral cues, which remained similar as harmonic number increased.

Further evidence suggesting that the ability to use TFS information is reduced for hearing-impaired listeners comes from recent research by Moore et al. (2006b). F0DLs were measured for hearing-impaired listeners using the same procedures and stimuli as in Moore et al. (2006a). For a centre frequency of 2000 Hz, F0DLs tended to improve with
increasing N, while for normally hearing subjects F0DLs worsened with increasing N in the range N=8 to 13. The pattern of results for the hearing-impaired subjects was similar to that found for normal-hearing subjects at a centre frequency of 5000 Hz, a frequency for which phase locking is often believed to be greatly reduced. The authors interpreted this result as evidence that hearing-impaired subjects were not using TFS information even at frequencies where phase locking is believed to be robust in normally hearing subjects.

Here, we attempted to determine more directly the extent to which hearing-impaired subjects can use TFS information, by measuring the discrimination of harmonic and frequency-shifted complex tones under conditions where TFS cues were available, but envelope and spectral cues were limited or absent.

2.2 Rationale

Moore and Moore (2003b) described an experiment in which a harmonic complex tone was matched in pitch to a frequency-shifted tone; the two tones were bandpass filtered so that they evoked very similar excitation patterns when components were unresolved. For frequency-shifted complexes containing only components with high frequencies relative to F0, no pitch shift was measured for normal-hearing subjects, suggesting that TFS information in these stimuli was inaccessible. If hearing-impaired subjects cannot use TFS information, even for complexes with lower harmonic numbers, then no pitch shift for these frequency-shifted complexes would be expected. However, when the experiment was attempted with hearing-impaired subjects, results were very erratic. Hearing-impaired subjects could not make consistent pitch matches, so no conclusions about their ability to use TFS information could be drawn.

Here, similar stimuli were used, but a three-interval forced-choice task rather than a pitch-matching task was chosen, as this was expected to be easier for non-musically-trained hearing-impaired subjects to perform. Subjects were required to discriminate frequency-shifted complexes from harmonic complexes; as in Moore and Moore (2003b), stimuli were bandpass filtered so that they would evoke almost the same excitation pattern on the basilar membrane when only unresolved components were present. The harmonic number of the centre component in the bandpass filter, N, was varied. If the hearing-impaired subjects were unable to use TFS information, they should be unable to discriminate the frequency-shifted and non-frequency-shifted complexes when N was such that components in the passband were unresolved.

A concern with this method is that, if hearing-impaired subjects do indeed perform this task very poorly, this could be because they have failed to understand the task or because performance was limited by cognitive factors, such as the ability to remember
the three stimuli within a trial. Cognitive factors were a potential concern since some of the hearing-impaired subjects included in the present study were elderly. To exclude this interpretation, and to further investigate the cues used for discrimination by normal-hearing and hearing-impaired subjects, two additional types of stimuli were used. The three stimulus types are illustrated in Figure 2.1.

![Figure 2.1: Schematic spectra of stimuli for SHAPED, NON-SHAPED and F0-DISCRIM stimulus types. Reference (unshifted) stimuli are shown in the bottom row, and shifted stimuli are shown in the top row. For the SHAPED condition, the bandpass filter was centred on the 11th harmonic.](image)

For stimulus type one (SHAPED), the stimuli to be discriminated were the bandpass filtered harmonic and frequency-shifted tones described above. The components of the frequency-shifted tone had the same spacing as for the harmonic comparison tone, but each component was shifted upwards by the same amount in Hz. For stimuli with only unresolved components (high frequencies relative to F0), the main cue for discrimination was changes in TFS, although changes in envelope might have been usable to a small extent (this is discussed in more detail later). For conditions with lower-frequency components relative to F0, the task could be performed by comparing the frequencies of individual resolved components.

For stimulus type two (NON-SHAPED), subjects discriminated harmonic and inharmonic tones similar to those for the SHAPED condition, except that complexes were made up of five equal-amplitude components only, meaning that upwards shifts in frequency were accompanied by an upward shift in the excitation pattern, even when harmonics were unresolved.

For stimulus type three (F0-DISCRIM), subjects were required to detect a change in
the F0 of a five-component harmonic complex with a centre component with harmonic number N. F0DLs have been measured many times previously with similar stimuli (Moore and Glasberg, 1988; Moore and Peters, 1992; Arehart, 1994), and it has been found that even severely hearing-impaired subjects can perform the task if the difference in F0 is made large enough. Envelope, TFS and spectral cues were all available for these stimuli.

An additional reason for using the NON-SHAPED and F0-DISCRIM stimuli was to prevent the hearing-impaired subjects from becoming unduly discouraged; we found in pilot studies that the performance of these subjects with SHAPED stimuli was often very poor.

2.3 Method

2.3.1 Subjects

Ten subjects with normal hearing and seven hearing-impaired subjects were recruited for this experiment. Five of the ten normal-hearing subjects and four of the seven hearing-impaired subjects had previous experience of psychoacoustical experiments, and one of the normal-hearing subjects and two of the hearing-impaired subjects were musically trained. Normal-hearing subjects were between 20 and 22 years of age and had absolute thresholds less than 15 dB HL in their test ears at all audiometric frequencies. Audiograms of the test ears of the hearing-impaired subjects are shown in Figure 2.2. These were measured using a Grason-Stadler GSI 61 audiometer with Telephonics TDH 50-P earphones. The age of each subject is indicated in the figure.

All hearing-impaired subjects had air-bone gaps less than 15 dB at octave frequencies between 500 Hz and 4000 Hz, indicating that there was no large conductive component of the hearing loss. Hearing-impaired subjects were tested for cochlear dead regions for octave frequencies between 500 and 4000 Hz, using the method described by Moore et al. (2004). No dead regions were found in any subject. Two of the hearing-impaired subjects (HI 6 and HI 7) had little or no hearing loss at high frequencies.

The ear with the smaller variation in pure tone thresholds across frequency was chosen as the test ear for each subject. For HI 6, the worse hearing ear was used as the test ear, so it was necessary to prevent cross hearing. Masking noise generated with an IVIE IE-20B pink noise generator was presented to the non-test ear at a level of 30.6 dB/ERB\(_N\) at 1000 Hz (where ERB\(_N\) refers to the equivalent rectangular bandwidth of the auditory filter for young, normally hearing subjects tested at moderate sound levels; Glasberg and Moore, 1990). This noise level was calculated from the maximum stimulus level per ERB\(_N\) in the test ear. The level reaching the contralateral ear was determined by subtracting 40 dB
Figure 2.2: Air conduction audiometric thresholds for the test ears of the hearing-impaired subjects. The age of each subject is also shown.

from this value, which is the inter-aural attenuation measured for the headphones that were used. The final level of the contralateral masker was calculated by adding 6 dB to this level to ensure masking.

Five normal-hearing subjects were tested with SHAPED stimuli (NH 1-NH 5), five with NON-SHAPED stimuli (NH 6-NH 10) and four with the F0-DISCRIM stimuli (NH 6, NH 7, NH 9 and NH 10). All seven hearing-impaired subjects were tested using SHAPED stimuli, four were tested using NON-SHAPED stimuli (HI 3, HI 4, HI 6, and HI 7), and three were tested using F0-DISCRIM stimuli (HI 1, HI 2 and HI 5). The order in which conditions were tested was randomised for each subject.

Subjects were paid for their time, apart from one of the authors, HI 7. Subjects underwent training until performance appeared to be stable; this usually took one hour, but took longer for some subjects.

2.3.2 Stimuli

Nominal F0s of 100, 200 and 400 Hz were used. For normal-hearing subjects, discrimination was measured with N=7, 11 and 18 for each F0, making nine conditions for each stimulus type. The condition with F0=400 Hz and N=18 was not tested for five of the seven hearing-impaired subjects (HI 1-HI 5), as residual hearing in this high-frequency region (around 7200 Hz) was very poor.
Stimulus type one (SHAPED): Discrimination of harmonic and frequency-shifted complexes spectrally shaped to reduce excitation pattern cues

A trial consisted of three intervals, two containing a harmonic complex and one containing a frequency-shifted complex. For the harmonic complex, harmonics 1 to 35 of the F0 were added, each starting in sine phase. The inharmonic complex was formed in the same way as the harmonic complex except that each component was shifted upwards in frequency by the same amount in Hz. For example, for a harmonic complex that contained (among others) components with frequencies of 600, 700 and 800 Hz, the corresponding inharmonic complex, with a shift of 20 Hz, would contain components with frequencies of 620, 720 and 820 Hz. The amplitudes of the components were defined using a bandpass filter function with a central flat region with a width of 5F0 and skirts that decreased in level at a rate of 30 dB/octave; see Figure 2.1.

The maximum shift was 0.5F0 Hz. Excitation patterns for harmonic and frequency-shifted complexes with the largest possible frequency shift (0.5F0) were calculated using the model proposed by Moore et al. (1997), to check that differences were small when the components in the passband were unresolved. The same levels of stimulus components and threshold equalizing noise were used as in the experiment for normal-hearing subjects (see later). Table 2.1 shows the largest difference in excitation level at any single frequency for the harmonic and frequency-shifted complexes for each condition. Examples of excitation patterns for N=7, 11 and 18 are shown in Figure 2.3. The largest excitation-pattern differences were about 2 dB for N=11 and about 0.5 dB for N=18. However, the model used to calculate these differences applies to normal-hearing subjects; excitation-pattern differences for hearing-impaired subjects would be smaller than calculated using this model, at least in frequency regions of hearing loss, probably by a factor of two or more, since the auditory filters typically broaden with increasing hearing loss (Glasberg and Moore, 1986). Thresholds for detecting a change in excitation level in a restricted frequency region are typically 2-4 dB (Moore et al., 1989). Hence, the excitation-pattern changes would have been barely, if at all, detectable for both the normal-hearing and hearing-impaired subjects.

Note that when N=7, components would be resolved for normally hearing subjects, so the excitation pattern is clearly different for the harmonic complex (solid line) and the frequency-shifted complex (dotted line). This could allow discrimination using excitation pattern cues, although such cues would be less effective for hearing-impaired subjects.
Table 2.1: Largest differences in excitation level between harmonic and frequency-shifted SHAPED tones for the maximum frequency shift of 0.5F0. Excitation patterns were calculated using a model for normal hearing (Moore et al., 1997).

<table>
<thead>
<tr>
<th>F0</th>
<th>N</th>
<th>Largest difference in excitation level (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>7</td>
<td>4.1</td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>1.3</td>
</tr>
<tr>
<td>100</td>
<td>18</td>
<td>0.4</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>5.5</td>
</tr>
<tr>
<td>200</td>
<td>11</td>
<td>1.8</td>
</tr>
<tr>
<td>200</td>
<td>18</td>
<td>0.4</td>
</tr>
<tr>
<td>400</td>
<td>7</td>
<td>5.8</td>
</tr>
<tr>
<td>400</td>
<td>11</td>
<td>2.2</td>
</tr>
<tr>
<td>400</td>
<td>18</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Stimulus type two (NON-SHAPED): Discrimination of harmonic and frequency-shifted complexes without spectral shaping

Stimuli were synthesised in a similar way as for the SHAPED stimuli. However, harmonic and frequency-shifted complexes were made up of only five equal-amplitude components. No flanking components were presented. Consequently, in addition to TFS information, a change in excitation pattern could be used as a cue to identify the frequency-shifted complex, even when components were unresolved.

Stimulus type three (F0-DISCRIM): Discrimination of the F0 of a harmonic complex

Harmonic complexes were formed from five consecutive harmonics added in sine phase. The centre component had a harmonic number of N (7, 11 or 18). Two intervals contained a complex tone with F0 equal to the nominal value (100, 200 or 400 Hz), and the other interval contained a complex tone with F0 equal to the nominal value plus δ. Envelope, spectral and TFS cues could be all used for this stimulus type.

Threshold equalizing noise (TEN)

Threshold equalizing noise (TEN) (Moore et al., 2000), extending from 200 to 16,000 Hz, was used to mask combination tones for all task types, and, for the SHAPED stimuli, to help ensure that the audible parts of the excitation patterns evoked by the harmonic and
Figure 2.3: Excitation patterns (Moore et al., 1997) for SHAPED stimuli for F0=400 Hz and N=7, 11 and 18, presented in TEN with a spectrum level of 20 dB/ERB_N below the level of the most intense stimulus component. The patterns are plotted only over the frequency range where the SHAPED stimuli produced excitation comparable to or above that produced by the noise. Patterns for harmonic and frequency-shifted stimuli are plotted as solid and dotted lines respectively. The frequency shift was 0.5F0 Hz (the maximum shift).

frequency-shifted tones were the same. TEN was chosen because it is designed to give equal masked thresholds (in dB SPL) across frequency for subjects with normal hearing, and approximately equal masked thresholds for subjects with cochlear hearing loss, but without dead regions (Moore et al., 2000). This meant that we could be confident that combination tones of a particular level would be masked irrespective of their frequency. For normal-hearing subjects, the TEN level at 1000 Hz was set to 20 dB/ERB_N below the level of the most intense component in the complex tones. For hearing-impaired subjects, the TEN level was set to 30 dB/ERB_N below the level of the most intense component, to prevent the noise from being uncomfortably loud while still providing sufficient masking.

2.3.3 Signal generation

Stimuli were produced by a Tucker Davies Technologies (TDT) system II, using a 16-bit digital-to-analog converter (TDT DA4) with a sampling rate of 50,000 Hz. Levels of the tones and the TEN were controlled independently using two TDT PA4 attenuators and stimuli were low-pass filtered at 20,000 Hz using Kemo (VBF8) dual variable filters. Stimuli were presented via Sennheiser HD580 headphones in a double-walled sound-attenuating chamber.

For the normal-hearing subjects, each complex was presented at an overall level of 65 dB SPL. Complexes were presented to hearing-impaired subjects at a sensation level of
20 dB, except for subjects HI 1 and HI 7 who found this level to be too quiet. For them, a level of 30 dB SL was used. The audiograms of the hearing-impaired subjects were used to calculate sensation levels at the frequency corresponding to N for each condition. Linear interpolation between audiometric frequencies and conversion to dB SPL allowed the appropriate level to be calculated. Two hearing-impaired subjects (HI 6 and HI 7), had normal audiometric thresholds at some frequencies, so for some conditions, a level of 20 or 30 dB SL would give levels much lower than those presented to normal-hearing subjects. For these conditions, an overall level of 65 dB SPL was used, which was the same as the level used for normal-hearing subjects.

2.3.4 Procedure

The same experimental procedure was used for all stimulus types. When two stimulus types were tested for the same subject, runs for the two stimuli types were interleaved and the subject was not informed of which stimulus type was being presented.

A trial consisted of three successive stimuli, indicated by lights on a response box. Each stimulus was 540 ms long including 20-ms raised-cosine onset and offset ramps. The inter-stimulus interval was 200 ms. Two intervals contained the same stimulus and the third, chosen at random, contained a different one. Subjects were instructed to press the button corresponding to the interval that sounded different. Feedback was given after every trial via lights on the response box.

Initially, we tried to use an adaptive procedure to estimate thresholds for discrimination for each stimulus type. We assumed that discriminability would increase monotonically with increasing frequency shift, even for the SHAPED stimuli, since, for normally hearing subjects, the pitch shift increases monotonically with increasing frequency shift up to at least 0.25F0 \cite{MooreMoore2003b}. This assumption was confirmed by the orderly nature of the adaptive tracks obtained for conditions where performance was relatively good. However, especially for the SHAPED stimuli, performance was sometimes too poor for the adaptive procedure to be used, since the procedure called for a change larger than the maximum possible value (a shift of 0.5F0 for the SHAPED stimuli). In such cases, performance was measured with the shift fixed at 0.5F0. Details of the subjects and conditions where the non-adaptive procedure was used are shown in Tables 2.2 and 2.3 for the normal-hearing and hearing-impaired subjects, respectively. Column headings SH and N-SH refer to SHAPED and NON-SHAPED stimulus types and the abbreviation na indicates conditions for which the non-adaptive procedure was used. The two types of procedure are described below.
Table 2.2: Conditions for which the non-adaptive procedure (na) was used for the normal-hearing subjects. Stimulus types SHAPED (SH) and NON-SHAPED (N-SH) are shown only, as all subjects completed the adaptive procedure for all conditions in the F0-DISCRIM task.

<table>
<thead>
<tr>
<th>F0</th>
<th>N</th>
<th>NH 1</th>
<th>NH 2</th>
<th>NH 3</th>
<th>NH 4</th>
<th>NH 5</th>
<th>NH 6</th>
<th>NH 7</th>
<th>NH 8</th>
<th>NH 9</th>
<th>NH 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SH</td>
<td>SH</td>
<td>SH</td>
<td>SH</td>
<td>SH</td>
<td>N-SH</td>
<td>N-SH</td>
<td>N-SH</td>
<td>N-SH</td>
<td>N-SH</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>100</td>
<td>18</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>200</td>
<td>11</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>200</td>
<td>18</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>400</td>
<td>7</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>400</td>
<td>11</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>400</td>
<td>18</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
</tbody>
</table>

Table 2.3: As Table 2.2 but for the hearing-impaired subjects. Dashes indicate conditions that were not tested.

<table>
<thead>
<tr>
<th>F0</th>
<th>N</th>
<th>HI 1</th>
<th>HI 2</th>
<th>HI 3</th>
<th>HI 4</th>
<th>HI 5</th>
<th>HI 6</th>
<th>HI 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SH</td>
<td>SH</td>
<td>N-SH</td>
<td>SH</td>
<td>N-SH</td>
<td>SH</td>
<td>N-SH</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>100</td>
<td>18</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>200</td>
<td>11</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>200</td>
<td>18</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>400</td>
<td>7</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>400</td>
<td>11</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>400</td>
<td>18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>na</td>
</tr>
<tr>
<td></td>
<td></td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
</tr>
</tbody>
</table>

36
Adaptive procedure

The tracking variable (the frequency shift of the components for the SHAPED and NON-SHAPED stimuli or the value of $\delta$ for the F0-DISCRIM stimuli) was varied adaptively using a three-down one-up tracking procedure (Levitt, 1971). It was increased by a factor $k$ after one incorrect response and was decreased by the same factor after three consecutive correct responses. For the first four turnpoints, $k$ equaled 1.414 and for the last eight turnpoints, it was reduced to 1.189. Twelve turnpoints were obtained and the geometric mean of the frequency differences at the last eight was taken to be the threshold corresponding to 79.4% correct. The standard deviation (SD) of the logarithms of the turnpoint values was also calculated. If this SD was greater than 0.2 then results of the run were discarded and the condition was repeated.

Each condition was tested three times and the geometric mean of the threshold estimates was calculated. The SD of the log values of the threshold estimates was determined, and if it was greater than 0.15, then an extra run was obtained, and the final threshold was taken as the geometric mean of all four runs.

If a value of 0.5F0 was reached after the fourth turnpoint for the SHAPED and NON-SHAPED stimuli, then the run was aborted. Before the fourth turnpoint, the frequency difference was allowed to reach 0.5F0 four times before the run was aborted, as the first four turnpoints were not used in calculation of the run geometric mean. When a run was aborted, the subject was later re-tested on the same condition, using the adaptive procedure. If a subject was consistently unable to complete runs using the adaptive procedure, the non-adaptive procedure described below was used. It should be noted that the abortion of some runs, but acceptance of others for the same condition, might have led to a bias to accept runs for which performance was better; thus the thresholds estimated using the adaptive procedure may be underestimates of the true thresholds in cases where the threshold approached 0.5F0.

Non-adaptive procedure

This procedure was only used for SHAPED and NON-SHAPED stimulus types, as all subjects could complete the adaptive procedure for the F0-DISCRIM stimuli for all conditions.

Subjects were given the same instructions as for the adaptive procedure. They selected the different interval from three alternatives and feedback was given as before. A run consisted of 55 trials with the last 50 trials being used to calculate a percent correct value. The frequency shift was fixed at the maximum difference of 0.5F0. Subjects completed five non-adaptive runs. The mean of the percent correct values for each run was used to estimate the final percent correct value.
2.3.5 Statistics

To allow results from the two types of procedure (adaptive and non-adaptive) to be compared, thresholds from the adaptive procedure and percent correct values from the non-adaptive procedure were converted to $d'$ values \cite{Green1974}. Conversion was by use of a table of $d'$ values for m-alternative forced-choice procedures \cite{Hacker1979}. The adaptive procedure tracked the 79.4% correct point on the psychometric function, which corresponds to a $d'$ of 1.63 for a three-alternative forced-choice task. When the adaptive procedure was used, the $d'$ value that would have been measured for a difference of 0.5F0 Hz was calculated by dividing 1.63 by the threshold measured in the adaptive procedure, and multiplying this value by 0.5F0. This method for calculating $d'$ assumes that $d'$ is proportional to the frequency shift in Hz \cite{Nelson1986}. However, this assumption is not critical for interpreting the results. This method sometimes yielded values of $d'$ that were much larger than would normally be encountered as, for some conditions, a frequency difference of 0.5F0 Hz was much larger than the threshold value. Such large values of $d'$ would be difficult or impossible to measure in practice, and should not be taken too literally. The main point here is that the calculated $d'$ values are inversely proportional to the estimated threshold values, so that large $d'$ values indicate good performance. Table 2.4 summarises results for conditions for which the adaptive procedure was completed by all of the normally hearing subjects. The thresholds in Hz measured using the adaptive procedure are shown, as well as the calculated $d'$ values, so that the two can be related.

Statistical tests were performed on the square root of the absolute $d'$ values, as this transformation gave a roughly uniform variance (across subjects in the case of data for the normal-hearing subjects and within subjects in the case of the data for the hearing-impaired subjects) across conditions. Values that were negative before the transformation were multiplied by -1 after the transformation to restore their sign.

2.4 Results

Mean $d'$ values for the normal-hearing subjects for the SHAPED and NON-SHAPED stimuli are shown in Figure 2.4. Open and filled symbols denote results for the SHAPED and NON-SHAPED stimuli, respectively. Values of $d'$ are plotted on a square-root scale, as the SD of the $d'$ values was roughly proportional to the square root of their magnitude. Some conditions yielded $d'$ values that were not significantly different from zero (chance performance; see later for details of how this was determined). These points were assigned a different symbol (a square) and were plotted at zero. No error bars are shown for these points.
Table 2.4: Summary of results for normal-hearing subjects for conditions where all subjects completed the adaptive procedure. Geometric means of the thresholds measured using the adaptive procedure are shown in Hz. The $d'$ values that would have been measured for discrimination of tones with a frequency shift of 0.5 F0 Hz were estimated by extrapolation from the thresholds measured using the adaptive procedure for each subject. The mean of these $d'$ values across subjects for each condition are shown (see Section 2.3.5 for details).

<table>
<thead>
<tr>
<th>F0</th>
<th>N</th>
<th>SHAPED</th>
<th>NON-SHAPED</th>
<th>F0-DISCRIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hz</td>
<td>$d'$</td>
<td>Hz</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
<td>5.0</td>
<td>18.1</td>
<td>4.2</td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>14.6</td>
<td>7.0</td>
<td>9.7</td>
</tr>
<tr>
<td>100</td>
<td>18</td>
<td></td>
<td>22.5</td>
<td>3.7</td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>6.8</td>
<td>26.8</td>
<td>6.9</td>
</tr>
<tr>
<td>200</td>
<td>11</td>
<td>16.2</td>
<td>12.1</td>
<td>22.3</td>
</tr>
<tr>
<td>200</td>
<td>18</td>
<td></td>
<td>39.3</td>
<td>4.2</td>
</tr>
<tr>
<td>400</td>
<td>7</td>
<td>11.70</td>
<td>32.7</td>
<td>18.9</td>
</tr>
<tr>
<td>400</td>
<td>11</td>
<td>31.9</td>
<td>13.7</td>
<td>33.2</td>
</tr>
<tr>
<td>400</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.4 shows that performance worsened as N increased. This was true for all F0s tested and for both stimulus types. There was little difference between $d'$ values for the two stimulus types when N=7 or 11, but when N=18, performance was better for the NON-SHAPED stimuli than for the SHAPED stimuli. A mixed model analysis of variance (ANOVA) was performed with within-subject factors F0 (100, 200 and 400 Hz), N (7, 11 and 18) and between-subject factor stimulus type (SHAPED and NON-SHAPED). Throughout this chapter, the degrees of freedom used to calculate p values for each factor were corrected using the Greenhouse-Geisser correction. The effect of N was significant ($F(2,16)=136.9; p<0.001$), but the main effects of F0 and stimulus type were not. The interaction between N and stimulus type was significant ($F(1.26,16)=7.9; p=0.014$), and post hoc tests of the effect of stimulus type using the least-significant-differences (LSD) test (using the pooled error term from the ANOVA) (Keppel, 1991) showed that the effect of stimulus type was significant for N=18 ($p=0.024$), but not for N=7 or 11 ($p=0.462$ and 0.757, respectively). The implications of these results will be discussed later.

Figure 2.4: Mean $d'$ values for normal-hearing subjects for discrimination of harmonic and frequency-shifted complex tones, plotted as a function of N. Each panel shows results for one F0. Open and filled circles show $d'$ values for SHAPED and NON-SHAPED stimuli, respectively. Error bars show ± one SD of the mean. The square symbol indicates that the $d'$ value is not significantly different from zero. This point is plotted at zero.

The results for the hearing-impaired subjects showed marked individual differences, so further analysis was performed on individual subject data only. Data for the four subjects who were tested using SHAPED and NON-SHAPED stimuli are shown in Figure 2.5.
Figure 2.5: As Figure 2.4, but showing individual $d'$ values for the hearing-impaired subjects. The unfilled and filled square symbols indicate $d'$ values that are not significantly different from zero for SHAPED and NON-SHAPED stimuli, respectively. These points are plotted at zero.
Symbols have the same meaning as for Figure 2.4. The data for the three other hearing-impaired subjects who were tested with SHAPED stimuli (HI 1, HI 2 and HI 5) are not plotted, as they performed very poorly with those stimuli; performance was close to the chance level for most conditions, as described in more detail later. Subjects HI 3 and HI 4 showed better performance in all conditions for NON-SHAPED than for SHAPED stimuli. HI 6 showed a similar pattern for F0=100 and 200 Hz, but for F0=400 Hz, the pattern of results was more like that seen for normal-hearing subjects, with similar performance for the two stimulus types when N=7 or 11 but better performance for the NON-SHAPED stimuli for N=18. The fact that HI 6 showed similar performance to the normal-hearing subjects for F0=400 Hz probably reflects the near-normal absolute thresholds of HI 6 for frequencies of 2000 Hz and above. HI 7 showed better performance for the NON-SHAPED than for the SHAPED stimuli for all N for F0=100 Hz, but for F0=200 and 400 Hz, the pattern of performance for HI 7 was similar to that for the normal-hearing subjects. Again, this probably reflects the fact that HI 7 had near-normal absolute thresholds for frequencies of 3000 Hz and above.

A full factorial ANOVA was performed on the individual data with the same factors as for the ANOVA performed on the data for the normal-hearing subjects. The last three estimates of d′ for each condition were treated as replications. A summary of the results is shown in Table 2.5 (subjects HI 3, HI 4, HI 6, and HI 7). The degrees of freedom shown in the table are those before the Greenhouse-Geisser correction was applied; after correction, the degrees of freedom differed slightly across subjects. All p values incorporate the correction. For subjects who did not complete the conditions where F0=400 Hz and N=18, those conditions were treated as missing values for the purpose of the analysis. The missing values were estimated using the statistical package GENSTAT, although the ANOVAs were conducted using another package, SPSS. The main effects of F0 and stimulus type were significant for all subjects and the effect of N was significant for all subjects expect HI 4 (p>0.05). Subjects HI 6 and HI 7 showed significant two-way and three-way interactions between all of the factors. HI 4 showed no significant interaction effects, and HI 3 showed a significant interaction effect between F0 and stimulus type. The nature of these interactions is discussed in Section 2.5.

For some conditions, subjects performed very poorly. To assess whether poor performance (d′ < 0.5) was significantly better than chance level (d′=0), the standard error of the mean of the transformed d′ values for that condition was calculated (see Section 2.3.5 for details of the transformation). Performance was considered not significantly better than chance if the value zero fell within two standard errors of the mean of the transformed d′ values for a particular condition. Conditions where performance was not significantly better than chance are shown in Table 2.6. Stimulus types SHAPED and
Table 2.5: Summary of the ANOVA outcomes for hearing-impaired subjects tested with SHAPED and NON-SHAPED stimuli.

<table>
<thead>
<tr>
<th>Subject</th>
<th>F0</th>
<th>N</th>
<th>Stimulus type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F(2,36)</td>
<td>p</td>
<td>F(2,36)</td>
</tr>
<tr>
<td>HI 3</td>
<td>22.1</td>
<td>0.02</td>
<td>208.3</td>
</tr>
<tr>
<td>HI 4</td>
<td>28.0</td>
<td>0.03</td>
<td>5</td>
</tr>
<tr>
<td>HI 6</td>
<td>21.2</td>
<td>0.04</td>
<td>35.4</td>
</tr>
<tr>
<td>HI 7</td>
<td>116.8</td>
<td>0.00</td>
<td>314.3</td>
</tr>
</tbody>
</table>

Table 2.5: Summary of the ANOVA outcomes for hearing-impaired subjects tested with SHAPED and NON-SHAPED stimuli.

NON-SHAPED are denoted by column headings SH and N-SH, respectively. For HI 1 to HI 4, performance for SHAPED stimuli was extremely poor, with discrimination significantly better than chance in only a few conditions, typically those where N=7. Performance was, however, significantly better than chance for all conditions and subjects for the NON-SHAPED condition, except for HI 7 and HI 3 when F0=200 Hz and N=18.

Results for the F0-DISCRIM task are shown in Figure 2.6. Mean results are plotted for the four normally hearing subjects who were tested, and individual results are shown for the three hearing-impaired subjects. The hearing-impaired subjects were all able to perform the task, and could complete the adaptive procedure consistently. However, the normal-hearing subjects performed better than all of the hearing-impaired subjects. For the normal-hearing subjects, discrimination worsened as N increased, for all F0s. The effect of N on performance was smaller for the hearing-impaired subjects, although it was marked for F0=100 Hz.
Figure 2.6: \(d'\) values for F0 discrimination by normal-hearing and hearing-impaired subjects. Individual data for three hearing-impaired subjects are plotted (HI 1, HI 2 and HI 5, filled symbols), and average data for four normal-hearing subjects are shown in each frame for comparison (open symbols). Error bars indicate ± one SD of the mean.
Table 2.6: Summary of conditions for which performance was not significantly better than chance, as indicated by x. Only data for SHAPED (SH) and NON-SHAPED (N-SH) stimuli are shown. A dash indicates that a condition was not tested for that subject.

### 2.5 Discussion

#### 2.5.1 Data for the normal-hearing subjects

The data for the normal-hearing subjects for the SHAPED stimuli showed that performance worsened with increasing N. This is broadly consistent with the finding of Moore and Moore (2003b) that the pitch shift produced by a given frequency shift of the components decreased with increasing N, since one would expect that performance in our task would depend on the magnitude of the pitch shift produced by the frequency shift. However, Moore and Moore (2003b) found no pitch shift when SHAPED stimuli with N=16 were used, a result which they attributed to an inability to use TFS information when only very high harmonics were present. In the present study, performance for N=18 was poor, but was significantly better than chance when F0=100 Hz and 200 Hz. There are a number of possible explanations for this apparent discrepancy.

1. Subjects may have used remaining spectral (excitation-pattern) cues for discrimination for conditions when F0=100 and 200 Hz. However, this seems unlikely, given the small differences in the excitation patterns of the harmonic and frequency-shifted SHAPED complexes. Also, an explanation based on excitation-pattern cues does not account for why performance was not significantly better than chance for F0=400 Hz, where the maximum difference in excitation level was similar to when F0=100 and 200 Hz (see Table 2.1).
2. Although the envelope repetition rate was the same for the harmonic and frequency-shifted complexes, the shape of the envelope was somewhat different. It is possible that the normally hearing subjects used the change in envelope shape to discriminate the harmonic and frequency-shifted complexes. Such a change in shape would not have caused a change in low pitch, which would account for why Moore and Moore (2003b) found no pitch shift. This explanation can also account for the finding that performance was not significantly above chance for F0=400 Hz; it seems likely that discrimination of changes in envelope shape would be very poor for such a high F0, as sensitivity to modulation decreases for rates above about 120 Hz (Kohlrausch et al. 2000) and discrimination of envelope shape may require detection of higher harmonics of the envelope repetition rate (Dau et al. 1997). The possibility that the subjects here were able to use envelope cues is addressed in a supplementary experiment, which is described later.

3. For complexes containing only high-frequency harmonics, TFS information may contribute to sound quality but not to low pitch. Subjects may have discriminated harmonic and frequency-shifted complexes on the basis of differences in timbre. If discrimination was based on a limited ability to use TFS information, chance performance would be expected when F0=400 Hz and N=18, as the components of this complex fall entirely into the frequency region where it is believed that phase locking is much reduced. The results were consistent with this interpretation.

2.5.2 Data for the hearing-impaired subjects

For the hearing-impaired subjects, discrimination of SHAPED stimuli with N=11 or 18 was very poor whenever the stimuli fell in a frequency region where the hearing loss was 30 dB or more, suggesting that they could make very little use of TFS information to discriminate harmonic and frequency-shifted complexes. This is an important result, as other studies have presented data that indirectly imply an inability of subjects with cochlear hearing loss to make use of TFS information, but this has not previously been shown directly (Moore and Moore 2003a, Buss et al. 2004, Lacher-Fougère and Demany 2005, Lorenzi et al. 2006a, Moore et al. 2006b). Note that the poor performance of the hearing-impaired subjects for SHAPED stimuli with N=11 or 18 confirms that these subjects were not able to use excitation-pattern cues to perform the task for these conditions.

Some subjects showed above-chance performance for SHAPED stimuli when N=7. Although this may reflect a limited ability to use TFS information for the frequencies in question, discrimination could also have been based on comparison of the frequencies of individual components, as, despite the poorer frequency selectivity seen in hearing-
impaired subjects, components with low harmonic numbers may have been resolved for some subjects (Moore et al., 2006b).

Subjects HI 6 and HI 7 performed better than the other hearing-impaired subjects with SHAPED stimuli for some conditions. This can be attributed to the reduced severity of their hearing impairments. Their audiometric thresholds were within the normal range at high frequencies, and the varying deficits seen across frequency reflect this. For both subjects, performance was worst when F0=100 Hz, in which case the components fell into the frequency region where their hearing loss was greatest. Performance in conditions when F0=400 Hz was normal for both subjects, as would be predicted by their near-normal audiometric thresholds for high frequencies (see Figure 2.2).

The significant interactions between factors for the individual subjects highlighted by the ANOVAs may result from the differences across hearing-impaired subjects in the patterns of the audiometric thresholds. When components fell into frequency regions over which audiometric thresholds were nearly normal, performance was relatively good, especially for the conditions where spectral cues could be used (for the NON-SHAPED stimuli and for the SHAPED stimuli when N=7). When components fell into a region of significant hearing loss, performance based on spectral cues was poorer, as the auditory filters typically broaden with increasing hearing loss (Glasberg and Moore, 1986). Subjects HI 6 and HI 7 appeared to be able to use TFS information for some F0s, which also led to significant interactions between factors, as they showed a pattern of results similar to those of normal-hearing subjects for some conditions, and a pattern of results similar to those of hearing-impaired subjects for others.

It is unlikely that the poor performance of the hearing-impaired subjects for SHAPED stimuli can be explained by poor understanding of the task. Performance was markedly better for the NON-SHAPED and F0-DISCRIM stimuli than for the SHAPED stimuli, even though subjects were not informed of the stimulus type being presented. The good performance of HI 6 and HI 7 for SHAPED stimuli in some conditions also indicates that, for these subjects at least, poor performance cannot be attributed to cognitive factors.

The poor performance of the three hearing-impaired subjects tested in the F0-DISCRIM task is consistent with previous data (Hoekstra and Ritsma, 1977; Moore and Peters, 1992; Arehart, 1994; Moore et al., 2006b). The complex tones could have been discriminated using envelope, TFS, or spectral information, or any combination of these. Poor performance could, in part, be due to poor frequency selectivity, which would reduce the number of resolved harmonics that could be directly compared. However, the results are also consistent with a loss of ability to use TFS information. F0DLs for complexes with N=11 and 18 were similar (see Figure 2.6), suggesting that envelope cues alone may have been used for both values of N, as proposed by Moore and Moore (2003a).
For conditions for which the components were unlikely to be resolved (N=11 and 18),
the components always fell in frequency regions above about 900 Hz. Thus, our results do
not rule out the possibility that subjects with cochlear hearing loss are able to use TFS
information at lower frequencies. Indeed, there are data indicating that hearing-impaired
subjects have at least some ability to use TFS cues for lateralization (Lacher-Fougère and
Demany, 2005); for a review, see Moore (2007). It may be that hearing impairment has
a greater effect on the ability to use TFS information at high than at low frequencies,
perhaps because the precision of phase locking is reduced for frequencies above about 1000
Hz, even in the normal peripheral auditory system (Johnson, 1980; Palmer and Russell,
1986).

Up to this point, it has been assumed that normal-hearing subjects used TFS cues
to discriminate harmonic and frequency-shifted tones when components were unresolved,
and that a deficit in TFS processing was responsible for the poor performance by hearing-
impaired subjects in these conditions. However, as previously mentioned, it is possible
that envelope shape may have been used as a cue, especially when N=18.1 To investigate
the extent to which this cue was used to discriminate SHAPED stimuli, an additional
condition was tested.

2.6 Discrimination of harmonic and frequency-shifted
tones with spectral shaping and components added
in random phase

2.6.1 Rationale

In experiment one all components were added with the same starting phase (sine phase),
so harmonic and frequency-shifted complexes within a trial would have had somewhat
different envelope shapes. This could have contributed to discrimination ability for the
SHAPED stimuli, particularly when N=18. To assess the possible role of this cue, the
discrimination of spectrally shaped stimuli was measured when the components were
added with random starting phase (referred to as SHAPED-RP stimuli). The starting
phase varied across stimuli within a trial, so each stimulus had a different envelope shape,
but only one (the frequency-shifted tone) had a different TFS.

1The shape of the Hilbert envelope was the same for the harmonic and frequency-shifted tones in the
acoustic signal, but the representation of the envelope shapes in the auditory system may be somewhat
different, because the representation is expected to be more faithful when there are more TFS peaks per
envelope period as was the case for the frequency-shifted tones.
2.6.2 Method

Four normal-hearing subjects were tested. Two had previously been tested with SHAPED stimuli (NH 1 and NH 5), and two had no previous experience of the task (NH 11 and NH 12). Three of these four subjects had musical training (NH 1, NH 11 and NH 12) and two had previous experience of psychoacoustic experiments (NH 5 and NH 12). Subjects were trained for one hour prior to data collection, after which performance appeared to be stable.

The SHAPED-RP stimuli were created in the same way as the SHAPED stimuli, except that components were added with randomly chosen starting phases (uniform distribution between 0 and 360°). Harmonic tones were synthesised prior to the start of each run, to reduce the delay between trials. Twenty harmonic complexes were synthesised, each with a different random selection of starting phases. A trial was made up of two harmonic complexes randomly selected from these pre-generated complexes, plus a frequency-shifted complex that was synthesised freshly for each trial.

The same procedure was used as for the SHAPED stimuli. Subjects were instructed to choose the interval containing the tone that sounded the most different. Subjects were advised that, for most conditions, the correct interval would have a different pitch than the other two, and if no pitch shift could be heard, they should use the feedback lights to try and identify the quality of the sound that they should use to identify the correct interval.

All subjects were unable to complete the adaptive procedure for all F0s when N=18, so a non-adaptive procedure was used for these conditions (see Section 2.3.4 for details of the adaptive and non-adaptive procedures).

2.6.3 Results

Mean d’ values for the SHAPED-RP stimuli are shown in Figure 2.7, with mean d’ values for the SHAPED stimuli re-plotted for comparison. For all F0s, performance was very similar for the two stimulus types when N=7 and 11. For N=18 and F0=100 and 200 Hz, performance was worse in the SHAPED-RP condition than in the SHAPED condition and, using the same criteria as earlier, was not significantly better than chance. However, as described earlier, performance in the shaped condition with N=18 and F0=100 and 200 Hz was significantly better than chance. Therefore, the randomization of phase had a deleterious effect when N=18 and F0=100 and 200 Hz. Performance was not significantly better than chance in the condition when F0=400 Hz and N=18 for either SHAPED or SHAPED-RP stimuli.
Figure 2.7: The filled symbols show $d'$ values for discrimination by normal-hearing subjects of harmonic and frequency-shifted tones with components added with random starting phases (SHAPED-RP stimuli). The open symbols show $d'$ values for discrimination of SHAPED stimuli by normal-hearing subjects, for comparison (data from the main experiment). The square symbols indicate $d'$ values that are not significantly different from zero. These points are plotted at zero.
2.6.4 Discussion

The absence of a significant difference between performance for the SHAPED and SHAPED-RP stimuli when N=7 and 11 suggests that envelope shape was not an important cue for the SHAPED stimuli for these values of N. It seems likely that no components were resolved for N=11 (see Figure 2.3). Hence, these results suggest that, for the normal-hearing subjects, TFS cues allowed better performance than envelope shape cues, so the latter were redundant. Similarly, the absence of a significant difference between the performance of normal-hearing subjects for SHAPED and NON-SHAPED stimuli when N=7 and 11 indicates that subjects did not use the extra excitation-pattern cue that was available for NON-SHAPED stimuli. In contrast, better performance for NON-SHAPED than for SHAPED stimuli when N=18 suggests that the additional spectral cue allowed better performance for these conditions. Additionally, the better performance for SHAPED than for SHAPED-RP stimuli when N=18 and F0=100 or 200 Hz suggests that subjects used a change in envelope shape to achieve above-chance performance in these conditions for SHAPED stimuli. This is consistent with the idea that use of TFS information for N=18 was either limited or absent, as suggested by Moore and Moore (2003b).

It is of interest that the normally hearing subjects were able to use envelope-shape cues to discriminate the SHAPED stimuli when N=18 and F0=100 or 200 Hz, while the hearing-impaired subjects apparently were not able to use these cues, since they mostly performed close to chance for these conditions. This suggests that the hearing-impaired subjects had some deficit in the ability to process envelope cues, perhaps because the cues were subtle. Previous work has suggested that hearing-impaired subjects have a near-normal ability to detect amplitude modulation (Bacon and Gleitman 1992; Moore et al. 1992) and to discriminate changes in amplitude modulation rate (Grant 1998), but they may show deficits in more complex tasks requiring the use of envelope cues (Lorenzi et al. 1997; Sek and Moore 2006). In any case, the results suggest that, for the SHAPED stimuli with N=11, the normally hearing subjects mainly used TFS cues to perform the task, whereas the hearing-impaired subjects were unable to use these cues.

2.7 Conclusions

1. Normal-hearing subjects appear to be able to use TFS information to discriminate inharmonic and frequency-shifted complexes with components that are unresolved, but not too high relative to the nominal F0.

2. For components with frequencies that are high with respect to the spacing between components, normal-hearing subjects appear to be unable to access TFS informa-
tion, consistent with the conclusions of Moore and Moore (2003b).

3. The results suggest that hearing-impaired subjects with moderate cochlear hearing loss have very little or no ability to use TFS cues to discriminate harmonic and frequency-shifted complex tones.
Chapter 3

Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech

A version of this chapter was published as: Hopkins, K., Moore, B. C. J. and Stone, M. A. (2008). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech." J. Acoust. Soc. Am. 123(2):1140-53

3.1 Introduction

Information in speech is redundant. For normal-hearing people, this means that the signal is robust to corruption, and that speech remains intelligible under adverse listening conditions, such as in high levels of background noise. In the normal auditory system, a complex sound like speech is filtered into frequency channels on the basilar membrane. The signal at a given place can be considered as a time-varying envelope superimposed on the more rapid fluctuations of a carrier, (temporal fine structure, TFS) whose rate depends partly on the centre frequency and bandwidth of the channel. The relative envelope magnitude across channels conveys information about the spectral shape of the signal and changes in the relative envelope magnitude indicate how the short-term spectrum changes over time. The TFS could carry information both about the F0 of the sound (when it is periodic) and about its short-term spectrum. For example, if at a particular time there is a formant centred at frequency fx (and hence one or more relatively intense stimulus components near fx), then channels centred close to fx will show TFS synchronised to fx,
and this will be reflected in the patterns of phase locking in those channels (Young and Sachs, 1979).

Many studies have assessed the relative importance of TFS and envelope information for speech intelligibility, for normal-hearing subjects. Vocoder processing has been used to remove TFS information from speech, so allowing speech intelligibility based on envelope and spectral cues to be measured (Dudley, 1939; Van Tasell et al., 1987; Shannon et al., 1995). A speech signal is filtered into a number of channels (N), and the envelope of each channel signal is used to modulate a carrier signal, typically a noise (for a noise vocoder) or a sine wave with a frequency equal to the channel centre frequency (for a tone vocoder). The modulated signal for each channel is filtered to restrict the bandwidth to the original channel bandwidth and the modulated signals from each channel are then combined. For a single talker, provided that N is sufficiently large, the resulting signal is highly intelligible to both normal-hearing and hearing-impaired subjects (Shannon et al., 1995; Turner et al., 1995; Başkent, 2006; Lorenzi et al., 2006a). However, if the original signal includes both a target talker and a background sound, intelligibility is greatly reduced, even for normal-hearing subjects (Dorman et al., 1998; Fu et al., 1998; Qin and Oxenham, 2003; Stone and Moore, 2003), leading to the suggestion that TFS information may be important for separation of a talker and background into separate auditory streams (Friesen et al., 2001).

As well as removing TFS information, vocoder processing also smears spectral information, an effect that is greatest when N is small. If the analysis filters are of a similar width to the auditory filters, however, the spectral information that is available to the auditory system is only slightly reduced compared to normal, though TFS information is still disrupted.

Another method for assessing the roles of different types of temporal information in speech is to attempt to remove envelope information but to leave TFS information (partially) intact. This was firstly attempted by infinite peak clipping of a wide-band speech signal (Licklider and Pollack, 1948), and later by using the Hilbert transform (Bracewell, 1986) to separate envelope and TFS information in each of a number of frequency channels (Smith et al., 2002). The TFS in each channel is preserved, but the envelope cues are removed, and the channel signals are then combined. Effectively, the processing in each channel behaves like a very fast compressor with an infinite compression ratio. For brevity, we will refer to this signal-processing method as TFS processing. At first sight, this method may appear to remove temporal envelope information and leave TFS intact. However, because envelope and TFS information are correlated, the envelope can be partially re-introduced by filtering in the peripheral auditory system, especially when the channels used in the processing have large bandwidths (Ghitza, 2001). This
problem is reduced if the signal is split into many narrow channels before removal of the envelope information, although some envelope cues may remain.

Gilbert and Lorenzi (2006) investigated the extent to which these recovered cues could be used to identify VCV nonsense syllables. They subjected nonsense syllables to TFS processing, and then passed the resulting signals through an array of filters that simulated filtering in the normal peripheral auditory system. The envelopes of the outputs of these filters were extracted and used to modulate tones at the centre frequencies of the filters. This is similar to tone-vocoder processing. The modulated tones were summed and presented to normal-hearing subjects, who were asked to identify the consonant that was presented. When the TFS processing used a small number of broad channels, subjects could identify the consonant accurately from the recovered envelope information. However, when the number of channels used in the TFS processing was eight or more, subjects scored close to chance. The authors concluded that if a signal is filtered into a sufficiently large number of channels before removing envelope cues, any recovered envelope cues are insufficient for intelligibility of VCVs. VCV syllables that are only TFS processed with a large number of analysis channels are reasonably intelligible to normally hearing listeners, after some training (Lorenzi et al., 2006a), which suggests that TFS cues alone can convey useful speech information.

Results from several studies have led to the suggestion that the ability to use TFS information is adversely affected by cochlear hearing loss. Much of this work has investigated the discrimination of synthetic complex sounds by hearing-impaired subjects (Lacher-Fougère and Demany, 1998, 2005; Moore and Skrodzka, 2002; Moore and Moore, 2003a; Moore et al., 2006b; Hopkins and Moore, 2007). For example, Hopkins and Moore (2007, Chapter 2) tested the ability of normal-hearing and hearing-impaired subjects to discriminate a harmonic complex tone from a frequency-shifted tone, in which all components were shifted up by the same amount in Hz (de Boer, 1956b). The frequency-shifted tone had very similar temporal envelope and spectral envelope characteristics to the harmonic tone, but a different TFS. All tones were passed through a fixed bandpass filter, to reduce excitation-pattern cues. When the filter was centred on the 11th component, so that the components within the passband were unresolved, subjects with moderate cochlear hearing loss performed poorly, while normal-hearing subjects could do the task well. Hopkins and Moore (2007) concluded that moderate cochlear hearing loss usually led to a reduced ability to use TFS information.

The reason for this is not clear. One possibility is that the precision of phase locking is reduced by cochlear hearing loss. One study found that phase locking was reduced in animals with induced hearing loss (Woolf et al., 1981), but another study found normal phase locking in such animals (Harrison and Evans, 1979a). It is unclear whether the
types of pathologies that cause cochlear hearing loss in humans lead to reduced phase locking. Another possible reason for a reduced ability to use TFS information is that TFS information could be decoded by cross-correlation of the outputs of two points on the basilar membrane (Loeb et al., 1983; Shamma, 1985; de Cheveigne and Pressnitzer, 2006). A deficit in this process, produced by a change in the traveling wave on the basilar membrane, would impair the ability to use TFS information even if phase locking were normal. The broader auditory filters typically associated with cochlear hearing loss (Liberman and Kiang, 1978; Glasberg and Moore, 1986) could also lead to a reduced ability to use TFS information. The TFS at the output of these broader filters in response to a complex sound would have more rapid fluctuations and be more complex than normal. Such outputs may be uninterpretable by the central auditory system (Sek and Moore, 1995; Moore and Sek, 1996).

A reduced ability to use TFS information could explain some of the perceptual problems of hearing-impaired subjects (Lorenzi et al., 2006b). TFS information may be important when listening in background noise, especially when the background is temporally modulated, as is often the case when listening in real life, for example, when more than one person is speaking. Normal-hearing subjects show better speech intelligibility (or lower speech reception thresholds, SRTs) when listening in a fluctuating background than when listening in a steady background (Festen and Plomp, 1990; Baer and Moore, 1994; Peters et al., 1998; Füllgrabe et al., 2006), an effect which is sometimes called masking release. Hearing-impaired subjects show a much smaller masking release, and it has been suggested that this may be because they are poorer at listening in the dips of a fluctuating masker than normal-hearing subjects (Duquesnoy and Plomp, 1983; Peters et al., 1998; Lorenzi et al., 2006a). Reduced audibility may account for some of the reduction in masking release measured for hearing-impaired subjects (Bacon et al., 1998), although the effect persists even when audibility is restored (Peters et al., 1998; Lorenzi et al., 2006b). TFS information may be important in dip listening tasks, as it could be used to identify points in the stimulus when the level of the target is high relative to the level of the masker; if the target and masker do not differ in their TFS, or no TFS information is available, dip listening may be ineffective.

Some studies have investigated the ability of hearing-impaired subjects to use TFS information in speech. Buss et al. (2004) showed that there was a correlation between temporal processing as assessed with psychoacoustic tasks and the ability of hearing-impaired subjects to recognise words in quiet. Lorenzi et al. (2006a) attempted to measure the ability of young and elderly hearing-impaired subjects to use TFS information in speech more directly. They applied 16-channel TFS processing to VCV nonsense syllables and asked subjects to identify the consonant in each syllable. According to Gilbert and
Lorenzi (2006), this number of channels should be sufficient to prevent the use of recovered envelope cues. Hearing-impaired subjects performed poorly at this task, while normal-hearing subjects scored around 90% correct after some training. Lorenzi et al. interpreted this result as indicating that the hearing-impaired subjects had a very limited ability to use the TFS information in the speech, whereas it was usable by normal-hearing subjects. Lorenzi et al. (2006a) also measured masking release for the young hearing-impaired subjects when listening to unprocessed speech in steady and modulated noise. The amount of masking release was found to be correlated with the score obtained for speech in quiet that had been subjected to TFS processing. This result is consistent with the argument made earlier, that the ability to use TFS is important for listening in the dips of a background sound.

A potential problem with the use of TFS-processed signals is that during gaps in the speech in a particular processing channel, low-level recording noise is amplified to the same level as the speech information. This is because the process is equivalent to multi-channel compression with an infinite compression ratio; whatever the original envelope amplitude in a given channel, the output envelope amplitude is constant. Channels with no speech information at a particular time are filled with distracting background sound. As a result, TFS-processed speech sounds harsh and very noisy. This may pose a particular problem to hearing-impaired subjects who, because of their broadened auditory filters, would suffer more from masking between channels. The problem becomes worse as the signal is split into more channels, as this results in more across-channel masking. Also, hearing-impaired listeners would be poorer at recovering any envelope cues that may still be available, again as a result of their broadened auditory filters. This could account for some of the difference in performance between normal-hearing and hearing-impaired subjects when listening to TFS speech. Here, a different approach was used to assess the use of TFS information by normal-hearing and hearing-impaired subjects. Rather than creating a signal that contains speech information only in its TFS, performance was measured as a function of the number of channels containing TFS information; the other channels were noise or tone vocoded, so that they conveyed only temporal envelope information.

3.2 Rationale

Hopkins and Moore (2007) found that subjects with moderate cochlear hearing loss could make little use of TFS information to discriminate complex tones. If similar subjects were completely unable to use TFS information in speech, they would be expected to perform as well when listening to speech that had been vocoded to remove TFS information as when
listening to unprocessed speech, provided that $N$ was sufficiently large that the frequency selectivity of the processing was similar to or better than that of the peripheral auditory system of the subject, thus avoiding significant loss of spectral information. However, Başkent (2006) found that hearing-impaired subjects performed better in a phoneme identification task when the syllables were unprocessed than when they were processed with a 32-channel noise-band vocoder. The disparity might arise because hearing-impaired subjects may be able to use TFS information at low carrier frequencies, but may be unable to use it at high frequencies. Hopkins and Moore (2007) showed that hearing-impaired subjects had a greatly reduced ability to discriminate the TFS of complex tones with unresolved components when all components were above 900 Hz, but they did not investigate sensitivity to TFS for lower frequencies. It is possible that subjects with moderate cochlear hearing loss are able to use TFS information below 900 Hz, which could explain why they performed better in the unprocessed condition than in the 32-channel vocoded condition in the study of Başkent (2006). If subjects with moderate cochlear hearing loss can use TFS information only at low carrier frequencies, progressively replacing vocoded information with unprocessed information, starting at low frequencies, should improve performance only up to a cut-off frequency above which TFS information cannot be used. This hypothesis was tested here.

SRTs corresponding to 50% correct keyword identification were measured for signals that were unprocessed for channels up to and including channel number CO and were vocoded for higher-frequency channels. The value of CO, which determined the amount of TFS information available in the signal was varied from 0 to 32. A competing-talker background was used, because, as described earlier, TFS information may be particularly important for listening in backgrounds that have temporal dips.

## 3.3 Method

### 3.3.1 Subjects

Nine normal-hearing subjects and nine hearing-impaired subjects took part in the experiment. The normal-hearing subjects were aged between 18 and 27 years and had audiometric thresholds of 15 dB HL or less at octave frequencies between 250 and 8000 Hz. The audiograms of the test ears of the hearing-impaired subjects are shown in Figure 3.1 (subjects HI 1 to HI 9) and the age of each subject is shown in brackets. All hearing-impaired subjects had air-bone gaps of 15 dB or less, and normal tympanograms, suggesting that their hearing loss was cochlear in origin. Hearing-impaired subjects were tested with the TEN HL test, which indicated no cochlear dead region for any subject.
Figure 3.1: Air conduction audiometric thresholds of the test ears of the hearing-impaired subjects for experiments one and two. The ages of the subjects (in years) are shown in brackets.

3.3.2 Speech material

Subjects were asked to repeat sentences presented in a competing talker background. The background began 500 ms before the target sentence, and continued after the target sentence had finished for about 700 ms (the exact value depended on the length of the target sentence). Each sentence list was added to a randomly chosen portion of a passage of continuous prose spoken by a competing talker. Long gaps between sentences and pauses for breath were removed from the background passage by hand editing. The same passage was used in both training and testing sessions. Both the target sentences and the competing talker passage had the same long-term spectral shape; for frequencies up to 500 Hz, the spectrum level was roughly constant, and for frequencies above 500 Hz.
the spectrum level fell by 9 dB per octave. For the training session, IEEE sentences were used (Rothauser et al., 1969). For the testing session, sentences were taken from the adaptive sentence list (ASL) corpus (MacLeod and Summerfield, 1990). Both target and competing talkers were male speakers of British English. The target talker had a fundamental frequency (F0) range of about 130-200 Hz, and the competing talker had a larger F0 range of about 130 to 280 Hz. The target and background speech were added together at the appropriate signal-to-background ratio (SBR) before processing.

3.3.3 Processing and equipment

Speech signals were split into 32 channels with centre frequencies spanning the range 100 to 10000 Hz, with an array of linear-phase, finite-impulse-response (FIR) filters. The filters had a variable order so that the transition bands of each filter had similar slopes when plotted on a logarithmic frequency scale. Each filter was designed to have a response of -6 dB at the frequencies at which its response intersected with the responses of the two adjacent filters. Channel edges were regularly spaced on an equivalent-rectangular-bandwidth (ERB) number scale and each channel was 1-ERB wide (Glasberg and Moore, 1990). This filtering was designed to simulate the frequency selectivity of the normal auditory system, so that the processing preserved nearly all of the spectral information available in the original signal. The signals from each channel were time aligned to compensate for the time delays introduced by the bandpass filtering. Stimuli were processed with nine values of CO (0, 4, 8, 12, 16, 20, 24, 28, 32). Channels with channel numbers up to and including CO were not processed further. Channels with channel numbers above CO were vocoded. The signals from these channels were half-wave rectified and these rectified signals were used to modulate white noise. Each modulated noise was subsequently filtered with the initial analysis filters and shaped to have the same spectral shape as the long-term spectrum of the original target speech from that channel. Consequently, envelope fluctuations with frequencies greater than half of the channel bandwidth were attenuated. After processing, the signals from the vocoded and unprocessed channels were added together.

All signals were generated with a high-quality 16-bit PC soundcard (Lynx One) at a sampling rate of 22050 Hz, passed through a Mackie 1202-VLZ mixing desk and presented to the subject monaurally via Sennheiser HD580 headphones. Subjects were seated in a

1Normally, the rectification would be followed by lowpass filtering, or the Hilbert transform would be used to extract the envelope. The omission of this stage in our processing meant that the modulator contained high-frequency components related to the TFS of the signal. However, these high-frequency components resulted in sidebands that were removed by the subsequent bandpass filtering. Listening tests and physical measurements confirmed that the processing used here gave results that were almost identical to those obtained when the Hilbert transform was used to extract the envelope.
double-walled sound-attenuating chamber.

### 3.3.4 Procedure

Microphones were placed in both the chamber and control room to allow communication between the experimenter and subject, although the control room microphone was only routed to the chamber headphones in the gaps between stimulus presentations. Target speech was presented to the normal-hearing subjects at a constant rms level of 65 dB SPL, which was equivalent to a spectrum level of 36.6 dB (re 20 $\mu$Pa) between 100 Hz and 500 Hz (see Section 3.3.2 for a description of the spectral shape). The level of the competing talker was varied to give the appropriate SBR, except when the SBR was less than -16 dB. Below this SBR, the level of the competing talker was not increased further, but instead the level of the target speech was reduced, to prevent the combined signal becoming uncomfortably loud. In practice, this was not necessary for any of the hearing-impaired subjects.

Previous studies have shown that audibility can account for some of the difference in performance between normal-hearing and hearing-impaired subjects listening in a temporally modulated background if stimuli are presented at the same level to both groups of subjects (Bacon et al., 1998; George et al., 2006). To reduce such effects, gains were applied to the combined target and background signal as prescribed by the CAMEQ hearing aid fitting method, according to the audiometric thresholds of each subject (Moore et al., 1998). Gains were specified at audiometric frequencies between 250 and 6000 Hz. The CAMEQ gains are designed to ensure speech audibility between these frequencies. Relatively more gain is prescribed for higher frequencies and this compensates for the increased upward spread of masking that is expected at higher overall levels, so helping to avoid the rollover effect on speech intelligibility as overall level increases (Fletcher, 1953; Studebaker et al., 1999). The CAMEQ gains were applied to the processed signals using a linear-phase FIR filter with 443 taps.

### 3.3.5 Audibility calculations

To check that the target speech would be audible for the hearing-impaired subjects after the CAMEQ gains were applied, excitation patterns were calculated for a signal that had the same long-term average spectrum as the target speech signal used for each subject. The spectrum for each subject was obtained by determining the long-term average spectrum of the speech with an overall level of 65 dB SPL and adding the CAMEQ gains at each frequency (with interpolation of gains for frequencies between the values specified by CAMEQ). Mean excitation levels between 100 Hz and 8000 Hz were calculated for
each subject using a model similar to that proposed by Moore and Glasberg (2004), but updated to incorporate the middle ear transfer function proposed by Glasberg and Moore (2006). Excitation levels are calculated relative to the excitation evoked by a 1000-Hz tone presented in free field with frontal incidence at a level of 0 dB SPL. The model allows the audiometric thresholds (in dB HL) of the individual subject to be entered. Default values were assumed for the proportion of the hearing loss attributed to outer hair cell and inner hair cell dysfunction. The model also gave estimates of the excitation level at threshold as a function of frequency for each subject. Figure 3.2 shows the excitation level at threshold and the mean excitation evoked by the (amplified) speech signal for each subject. The excitation level for the target speech was well above threshold excitation level except at very high or very low frequencies for some subjects. For most subjects, and for frequencies between 500 and 5000 Hz (the frequency range that is most important for speech intelligibility), the excitation level of the target speech was more than 15 dB above the excitation level at threshold, meaning that the entire dynamic range of the speech would have been audible, as speech is widely assumed to have a dynamic range of 30 dB, extending 15 dB above and below the root-mean-squared (rms) level (ANSI, 1997).

3.3.6 Training

Previous studies using vocoded speech material have shown large learning effects (Stone and Moore, 2003, 2004; Davis et al., 2005), so a training period was included before testing. Training lasted approximately one hour, and took place separately to the testing session. Firstly, subjects were played two passages of connected discourse to familiarise them with the task of listening in a competing talker background and to introduce the vocoder-processed speech. The first passage was unprocessed, and the second was vocoded across all 32 channels. The level of the competing talker was initially low, but was increased gradually throughout the passages. Subjects were instructed to listen to the target talker for as long as possible. The hearing-impaired subjects found this difficult, and so were given transcripts of the target passages to follow, which made the task easier.

For the next phase of training, IEEE sentences were presented at a fixed SBR. Six lists were presented, each made up of ten sentences. The sentences were processed with different values of CO and an SBR was selected by the experimenter to yield scores of approximately 70% correct. Subjects were required to repeat each sentence and the number of correctly identified key words was recorded. When subjects did not repeat the sentence perfectly, they were told the correct answer, and the sentence was repeated.

Finally, subjects were given an opportunity to practice the task used in the testing session. Four word lists similar to those in the ASL corpus were used. The same procedure was used as for the testing session, as described below.
Figure 3.2: Excitation levels of the target speech (solid lines) and excitation levels at threshold (dashed lines) for individual hearing-impaired subjects. Excitation levels for normal-hearing subjects are shown for comparison in the bottom-right panel.
3.3.7 Testing

Two consecutively presented ASL sentence lists were used for each condition and the order of presentation of conditions was counterbalanced across subjects. The SBR of the target and competing talker was varied adaptively. If a subject identified two or more keywords correctly in a sentence, the next sentence was presented with a SBR that was k dB lower, and if the subject identified fewer than two keywords correctly, the next sentence was presented with a SBR that was k dB higher. Before the third turnpoint was reached, k was equal to 4 dB; subsequently it was equal to 2 dB. The first sentence in each list was initially presented at an adverse SBR, at which the subject was expected to identify no keywords correctly. If the subject scored fewer than two keywords correctly, this sentence was repeated at an SBR that was 4 dB higher until at least two keywords were correctly identified. Subsequent sentences in each list were presented once only. For each sentence list, the total number of keywords presented at each SBR was recorded, as well as the number of keywords that were identified correctly for each SBR. The first sentence in each list was not included in these totals, as subjects could have heard this sentence more than once.

3.3.8 Analysis

For each SBR, the total keywords presented and keywords correct were summed for the two sentences lists that were presented for each condition. These values were used to perform a probit analysis \cite{Finney1971}, from which the SRT corresponding to the SBR required for 50% correct identification was estimated for each subject and each condition. In some cases, because of the scatter in the data, the probit analysis failed to fit the data and gave a slope of the psychometric function that was not significantly different from zero. This happened for at least one of the conditions for five of the normal-hearing subjects, but for only one of the hearing-impaired subjects (HI 8). For these cases, the SRT was estimated by plotting the proportion of correctly identified words against the SBR at which the words were presented. A line was drawn by eye to best fit the data points, and this line was used to exclude points from the probit analysis that did not fit the general trend. The probit analysis was then redone. In one case (NH 5, CO=20), after this procedure the probit analysis still did not give a psychometric function with a slope significantly different from zero, so the SRT for this case was treated as a missing data point for the remaining analysis.

An analysis of variance (ANOVA) was performed on all of the data from the normal-hearing and hearing-impaired subjects, with a within-subjects factor of CO and a between-subjects factor of subject type (normal-hearing or hearing-impaired).
3.4 Results

Figure 3.3 shows the mean data for both normal-hearing and hearing-impaired subjects. Mean SRTs are plotted for each value of CO/N. The hearing-impaired subjects performed more poorly than the normal-hearing subjects in all conditions, but the difference in performance varied with CO; for larger values of CO, the difference in performance between normal-hearing and hearing-impaired subjects was greater. The main effects of subject type and CO were significant \([F(1,8)=95.2, p<0.001\) and \(F(8,128)=53.9, p<0.001\) respectively\] and there was also a significant interaction between subject type and CO \([F(8,128)=12.2, p<0.001]\).

![Figure 3.3: Mean SRTs for normal-hearing and hearing-impaired subjects, plotted as a function of CO/N. The frequency corresponding to CO/N is shown along the top axis. Error bars show ± one standard deviation (SD) of the mean across subjects.](image)

CO had a greater effect on performance for the normal-hearing subjects than for the hearing-impaired subjects. For example, both subject groups performed better when speech was completely unprocessed (CO=32) than when it was completely vocoded (CO=0), but the difference in performance between these conditions was much greater for the
normal-hearing than for the hearing-impaired subjects (mean differences were 15.8 dB and 4.9 dB, respectively). Post hoc Fishers least-significant-difference (LSD) tests were used to determine whether the SRTs measured with different values of CO were significantly different from each other within each subject group. Tables 3.1 and 3.2 show the differences between mean scores for each value of CO for the normal-hearing subjects and hearing-impaired subjects, respectively. Values greater than the LSD are shown in bold.

<table>
<thead>
<tr>
<th>CO</th>
<th>Freq. (Hz)</th>
<th>0</th>
<th>4</th>
<th>8</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>28</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>277</td>
<td>3.8</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>548</td>
<td>7.4</td>
<td>3.7</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>965</td>
<td>10.2</td>
<td>6.4</td>
<td>2.8</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1605</td>
<td>12.4</td>
<td>8.7</td>
<td>5.0</td>
<td>2.2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2590</td>
<td>13.4</td>
<td>9.7</td>
<td>6.0</td>
<td>3.2</td>
<td>1.0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>4102</td>
<td>14.3</td>
<td>10.6</td>
<td>6.9</td>
<td>4.2</td>
<td>1.9</td>
<td>0.9</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>6427</td>
<td>14.6</td>
<td>10.9</td>
<td>7.2</td>
<td>4.5</td>
<td>2.2</td>
<td>1.2</td>
<td>0.3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>10000</td>
<td>15.5</td>
<td>11.8</td>
<td>8.1</td>
<td>5.4</td>
<td>3.1</td>
<td>2.1</td>
<td>1.2</td>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.1: Differences between mean SRTs measured with different values of CO for the normal-hearing subjects in experiment one. Equivalent cut-off frequencies (in Hz) are also shown. The LSD calculated using Fisher’s LSD procedure was 2.0. Differences equal to or above this value are shown in bold.

Figure 3.4 shows results for individual hearing-impaired subjects. The mean results for the normal-hearing subjects are shown in the bottom-right panel for comparison. Between-subject variability in overall performance was larger for the hearing-impaired than for the normal-hearing subjects. The pattern of results across conditions also varied more between hearing-impaired subjects. The benefit gained from the additional TFS information that was present when CO was large varied, with some hearing-impaired subjects benefiting little, if at all (for example, HI 4 and HI 5) and others benefiting almost as much as the normal-hearing subjects (HI 8).

### 3.5 Discussion

Normal-hearing subjects appear to benefit more than hearing-impaired subjects from the replacement of vocoded speech information with unprocessed speech information. This is
Figure 3.4: Individual SRTs for the hearing-impaired subjects, plotted as a function of CO/N. Mean SRTs measured for the normal-hearing subjects are shown in the bottom-right panel for comparison.
consistent with the idea that the hearing-impaired subjects had a reduced ability to use TFS information, which is consistent with previously published results (Lorenzi et al., 2006a; Moore et al., 2006b; Hopkins and Moore, 2007). Both groups did, however, improve as CO increased, though the amount of benefit from the additional TFS information varied across hearing-impaired subjects. This may reflect different abilities to use TFS information among hearing-impaired subjects with broadly similar audiometric thresholds, which could account for the weak correlation between audiometric thresholds and the ability to understand speech in noise previously reported for hearing-impaired subjects (Festen and Plomp, 1983; Glasberg and Moore, 1989). Other studies have also reported large individual differences in performance between hearing-impaired subjects when tasks require the use of TFS information (Buss et al., 2004; Moore et al., 2006b).

One possible concern is that the mean age of the normal-hearing subjects was much less than the mean age of the hearing-impaired subjects (21.9 years and 56.8 years, respectively), so the reduced benefit from the additional TFS might have been due to age rather than to hearing loss per se. Some previous studies have been interpreted as indicating that older subjects with near-normal audiometric thresholds have temporal processing deficits (Pichora-Fuller, 2003; Pichora-Fuller et al., 2006). However, other studies have tested young and elderly normal-hearing subjects listening to target speech in a temporally modulated background similar to that used here, and found relatively small differences in performance between the two groups (Takahashi and Bacon, 1992; Peters et al., 1998; Dubno et al., 2002; Lorenzi et al., 2006b). The differences were much smaller than the difference in performance seen here between the hearing-impaired and normal-hearing subjects in the unprocessed condition (CO=32). It is possible that some

<table>
<thead>
<tr>
<th>CO</th>
<th>Freq. (Hz)</th>
<th>0</th>
<th>4</th>
<th>8</th>
<th>12</th>
<th>16</th>
<th>20</th>
<th>24</th>
<th>28</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>277</td>
<td>0.9</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>548</td>
<td>2.1</td>
<td>1.1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>965</td>
<td>3.1</td>
<td>2.2</td>
<td>1.1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1605</td>
<td>3.5</td>
<td>2.6</td>
<td>1.4</td>
<td>0.4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>2590</td>
<td>4.7</td>
<td>3.7</td>
<td>2.6</td>
<td>1.5</td>
<td>1.2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>4102</td>
<td>5.2</td>
<td>4.3</td>
<td>3.2</td>
<td>2.1</td>
<td>1.8</td>
<td>0.6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>6427</td>
<td>5.6</td>
<td>4.7</td>
<td>3.6</td>
<td>2.5</td>
<td>2.2</td>
<td>1.0</td>
<td>0.4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>10000</td>
<td>4.9</td>
<td>3.9</td>
<td>2.8</td>
<td>1.8</td>
<td>1.4</td>
<td>0.2</td>
<td>0.4</td>
<td>0.8</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.2: As Table 3.1, but for the hearing-impaired subjects.
of the reduced ability to use TFS information seen for the hearing-impaired subjects in the current study could be attributed to their age, rather than their hearing loss. However, the pattern of results for the two young hearing-impaired subjects tested here (HI 2, 23 years and HI 6, 26 years) did not differ markedly from the pattern of the mean data for the hearing-impaired subjects, suggesting that hearing loss, rather than age was the important factor contributing to the reduced ability to use TFS information. The benefit from the addition of TFS was quantified as the difference between the SRT for CO=0 and CO=32. The correlation between benefit and age for the hearing-impaired subjects was not significant (r=-0.26, p=0.50). Again, this suggests that age was not the important factor determining the benefit from added TFS information.

Phase locking in the normal auditory system is widely believed to break down for frequencies above 4000-5000 Hz (Palmer and Russell, 1986; Moore, 2003). If this is true, TFS information above 4000-5000 Hz should be unusable, and so no improvement in performance would be expected when TFS information was added in the higher-frequency channels. Consistent with this, the Fisher LSD tests showed no significant difference in performance for the normal-hearing subjects for CO values from 24 to 32; CO=24 corresponds to a cut-off frequency of 4102 Hz. For the hearing-impaired subjects, performance appears to plateau at a lower value of CO. With one exception, LSD tests revealed no significant difference in performance for values of CO from 16 to 32; CO=16 corresponds to a frequency of 1605 Hz. This is consistent with the idea that hearing-impaired subjects may be able to use TFS information at low frequencies, but are unable to use higher-frequency TFS information, even for frequencies where phase locking is believed to be robust in the normal auditory system. However, the plateau in performance for both normal-hearing and hearing-impaired subjects for high values of CO could have been partly due to redundancy in the TFS information (see Chapter 5).

A possible concern when comparing hearing-impaired and normal-hearing subjects is that differences in performance may be explained by differences in audibility. For our experiment, this explanation seems unlikely. Figure 3.2 shows that the entire dynamic range of speech would have been audible for most of the hearing-impaired subjects for frequencies between 500 and 5000 Hz. Audibility was compromised at very low and high frequencies for some subjects, but this reduced audibility is unlikely to have affected speech intelligibility and cannot account for the large differences between hearing-impaired and normal-hearing subjects.

The improvement in SRT as CO increased has been interpreted so far as reflecting an ability use TFS information, but this is not the only interpretation of these results. Another possibility is connected with the idea that a noise-band vocoder may introduce distracting or masking low-frequency modulations into the signal. Whitmal et al. (2007)
found that normal-hearing subjects scored better when tested with a tone vocoder than with a noise vocoder (see also Dorman et al. 1997 and Stone et al. 2008). They suggested that modulations introduced by the noise carrier may have caused a reduction in speech intelligibility. The modulation spectrum of a bandpass filtered noise is triangular (Schwartz, 1970), with more modulation energy at low frequencies. This means that the modulations introduced by the noise carrier are dominated by modulation frequencies that are similar to those thought to be important in understanding speech (Drullman et al. 1994b,a; Shannon et al. 1995). When the number of analysis channels is large (so the channel widths are small), this is an even greater problem, as higher-frequency modulations are removed when the channel signals are filtered after vocoder processing, leaving the signal even more dominated by low-frequency noise modulations. It is possible that the normal-hearing subjects and the hearing-impaired subjects who showed greater improvement as CO increased did not benefit from the additional TFS information, but performed better because the spurious modulations introduced by the noise carrier were reduced, as the proportion of the signal that was vocoded was reduced.

Another factor that might have influenced the change in performance with increasing CO is connected with the effect of the processing on the representation of high-rate envelope fluctuations. Rosen (1992), suggested that modulations between 50 and 500 Hz are important in providing information about voice periodicity, and this voice periodicity information is important for listening in a competing talker background (Brokx and Nooteboom 1982; Assmann and Summerfield 1990). In experiment one, the speakers were male, with F0s varying between 130 and 280 Hz. The processing used 32 channels, which were equally spaced on an ERB<sub>N</sub>-number scale; each channel was 1-ERB<sub>N</sub> wide. This was intended to simulate the frequency selectivity of the normal auditory system. The highest modulation rate that can be carried by a channel is determined by the bandwidth of the channel. The filtering that was used subsequent to modulation of the channel carriers would have attenuated the side bands produced by the modulation, hence reducing the modulation depth for high rates. As a result, voice periodicity information would have been partially removed from channels tuned to lower centre frequencies.

This restriction of periodicity information would not have any important effects for normally hearing subjects, because the filters used in the processing had comparable widths to the normal auditory filter. However, hearing-impaired subjects generally have broader auditory filters than normal-hearing subjects, so, for unprocessed speech, higher-frequency modulation side bands would be attenuated less by the peripheral auditory system. Consequently, post-processing filtering of the vocoded signal into 1-ERB<sub>N</sub>-wide channels could reduce the periodicity information available to the hearing-impaired subjects, and this could be a reason for their worse performance with CO=0 than with...
CO=32.

These possible explanations for the improvement in performance with increasing CO found for the normal-hearing subjects and some of the hearing-impaired subjects were investigated in experiment two.

### 3.6 Experiment two

#### 3.6.1 Rationale

Experiment two was broadly similar to experiment one, but a tone vocoder was used rather than a noise vocoder. The carrier signals were sine waves with frequencies equal to the channel centre frequencies. No random modulations were introduced by the carrier signals, unlike for the noise vocoder used in experiment one.

Previous work has suggested that subjects with moderate cochlear hearing loss have auditory filters that are between two and four times as broad as those for normal-hearing subjects (Glasberg and Moore, 1986; Moore, 2007). In experiment two, the signal was divided into either 8 or 16 channels before processing, rather than 32, so that each channel was wider, and more comparable to the auditory filters of the hearing-impaired subjects (channels were 4- or 2-ERB_N wide rather than 1, as previously). This avoided the possible loss of modulation at F0 rates. A consequence of splitting the signal into fewer channels before vocoder processing is that more spectral detail from the original signal is lost. If the filters used in the processing are broader than those in the peripheral auditory system, as they would be for normal-hearing subjects, this in itself may lead to poorer performance. To check the effect of decreasing N, and to allow comparison with the results of experiment one, a condition was run with N=32 and CO=0, so that stimuli were fully tone vocoded, but with the same value of N as for experiment one.

#### 3.6.2 Method

**Subjects**

Five of the normal-hearing and seven of the hearing-impaired subjects who took part in experiment one also took part in experiment two. Four normal-hearing and two hearing-impaired subjects were newly recruited. Recruitment criteria were the same as for experiment one. The audiograms and ages of all of the hearing-impaired subjects used in both experiments are shown in Figure 3.1. HI 10 and HI 11 took part in experiment two only, HI 4 and HI 6 took part in experiment one only, and the remaining subjects took part in both experiments.
Speech material

ASL lists were used for training, as most of the subjects had already heard these in the testing session for experiment one. For the testing session, Bench-Kowal-Bamford (BKB) sentence material was used, which is similar in style to the ASL sentence material (Bench and Bamford 1979).

Processing

Sentences were processed in a similar way as for experiment one, but a tone vocoder was used rather than a noise vocoder, and the signal was split into 8 or 16 channels before processing rather than 32 (so channels were 4- or 2-ERB_N wide rather than 1-ERB_N wide). Signals were split into channels, and the envelope of each channel was extracted as before. Sine waves with frequencies equal to the centre frequency of each channel were used as carrier signals rather than noise bands, and these sine waves were modulated with the envelope of the original channel signals. As before, processed channel signals were filtered to remove side bands that were introduced as a result of the processing, so limiting the frequency of modulation that could be carried in each channel.

Conditions and procedure

For N=8, values of CO were 0, 2, 4, 6 and 8. For N=16, values of CO were 0, 4, 8 and 12 (note that the condition where N=16 and CO=16 is the same as N=8 and CO=8, so this condition was not retested). For five of the normal-hearing subjects, and five of the hearing-impaired subjects, an additional condition was tested, with N=32 and CO=0, but still using a tone vocoder. The procedure for training and testing sessions was the same as for experiment one, except for the differences in sentence material, as noted previously. Data were analyzed in the same way as for experiment one.

3.6.3 Results

The results are summarised in Figure 3.5. Mean SRTs are plotted against CO/N. A given value of CO/N corresponds to a fixed frequency, as indicated at the top of the panels in Figure 3.5. As in experiment one, the hearing-impaired subjects performed more poorly than the normal-hearing subjects for all conditions, and the difference in performance between the two groups was greatest when CO/N=1. An ANOVA was performed with N and CO/N as within-subject factors and subject type as a between-subject factor. The main effects and two-way interactions were all highly significant (p<0.001) and the three-way interaction was also significant [F(4,64)=2.83, p=0.03].
Figure 3.5: Mean SRTs for normal-hearing subjects (left) and hearing-impaired subjects (right), plotted as a function of CO/N. Note that the points for CO/N=1 for N=8 and N=16 plot the same data; these conditions were identical (the whole spectrum contained unprocessed information), and so were not tested twice. Error bars show ± one SD of the mean across subjects.
Differences between mean results for different values of CO/N are shown in Tables 3.3 and 3.4 for the normal-hearing and hearing-impaired subjects, respectively. For the normal-hearing subjects, SRTs did not differ significantly for CO/N=0.75 and 1 (for N=8) or for CO/N=0.5, 0.75 and 1 (for N=16). Thus, performance reached a plateau for higher values of CO/N, as found in experiment one. For the hearing-impaired subjects, SRTs reached a plateau at a lower value of CO/N. SRTs did not differ significantly for CO/N=0.5, 0.75 and 1 (for N=8) or for CO/N=0.5, 0.75 and 1 (for N=16). For N=16, the SRT for the normal-hearing subjects decreased by 16.0 dB as CO/N was increased from 0 to 1. This is similar to, but slightly larger than the decrease found in experiment one for 32-channel processing. For N=8, the decrease was larger, at 22.0 dB, because the SRT was higher for N=8 than for N=16 when the signal was fully vocoded (CO/N=0). For the hearing-impaired subjects, the decrease in SRT with increasing CO/N was much smaller, 4.1 dB for N=16 and 6.4 dB for N=8. Thus, as found in experiment one, the benefit of progressively adding TFS information (by increasing CO/N) was much smaller for the hearing-impaired than for the normal-hearing subjects, despite the use of a tone vocoder and a smaller number of channels in experiment two.

Fisher LSD tests revealed that normal-hearing subjects performed better for N=16 than for N=8, except when CO/N≥0.75 (see Table 3.5). Hearing-impaired subjects performed better for N=16 when CO=0, but not for higher values of CO. For CO/N=0.25, performance was significantly better for N=8 than N=16 for the hearing-impaired subjects, and when CO/N ≥ 0.5, there was no significant difference in performance for N=8 and 16.

Three students t-tests were performed to assess whether there was a significant effect of number of channels (N=8, 16 or 32) when CO=0 (i.e., when the signal was completely vocoded), for those hearing-impaired and normal-hearing subjects who were tested in all conditions. A Bonferroni correction for multiple comparisons was applied. The normal-hearing subjects performed significantly better when N=32 than when N=16 (p=0.01), whereas the hearing-impaired subjects did not perform significantly differently for the two conditions (p=0.12). However, the SRT of the hearing-impaired subjects did increase when the value of N was decreased further to 8, and the difference in SRT between 32 and 8 channels was significant (p=0.05).

Performance in the condition when N=32 and CO=0 was much better than for the same condition in experiment one, for both normal-hearing and hearing-impaired subjects (the mean SRTs were 7.8 dB and 4.7 dB lower, respectively, in experiment two).
<table>
<thead>
<tr>
<th>CO/N</th>
<th>N=8</th>
<th>N=16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 0.25 0.5 0.75 1</td>
<td>0 0.25 0.5 0.75 1</td>
</tr>
<tr>
<td>Freq. (Hz)</td>
<td>100 548 1605 4102 10000</td>
<td>100 548 1605 4102 10000</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>0.25</td>
<td>548</td>
<td>13.7 0</td>
</tr>
<tr>
<td>0.5</td>
<td>1605</td>
<td>18.2 4.6 0</td>
</tr>
<tr>
<td>0.75</td>
<td>4102</td>
<td>21.0 7.4 2.8 0</td>
</tr>
<tr>
<td>1</td>
<td>10000</td>
<td>22.0 8.3 3.8 1.0 0</td>
</tr>
</tbody>
</table>

Table 3.3: As Table 3.1 but for experiment two, which used two values of N. The LSD calculated using Fisher’s LSD procedure was 2.2.

<table>
<thead>
<tr>
<th>CO/N</th>
<th>N=8</th>
<th>N=16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 0.25 0.5 0.75 1</td>
<td>0 0.25 0.5 0.75 1</td>
</tr>
<tr>
<td>Freq. (Hz)</td>
<td>100 548 1605 4102 10000</td>
<td>100 548 1605 4102 10000</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>0.25</td>
<td>548</td>
<td>5.3 0</td>
</tr>
<tr>
<td>0.5</td>
<td>1605</td>
<td>5.9 0.6 0</td>
</tr>
<tr>
<td>0.75</td>
<td>4102</td>
<td>7.1 1.8 1.2 0</td>
</tr>
<tr>
<td>1</td>
<td>10000</td>
<td>6.4 1.1 0.5 0.7 0</td>
</tr>
</tbody>
</table>

Table 3.4: As Table 3.3 but for hearing-impaired subjects

<table>
<thead>
<tr>
<th>CO/N</th>
<th>Normal hearing</th>
<th>Hearing impaired</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 0.25 0.5 0.75 1</td>
<td>0.4 1.5 0</td>
</tr>
<tr>
<td>Freq. (Hz)</td>
<td>100 548 1605 4102 10000</td>
<td>10000</td>
</tr>
<tr>
<td>Normal hearing</td>
<td>6.0 3.5 2.0 0.5 0</td>
<td>2.3 1.7 0.4 1.5 0</td>
</tr>
<tr>
<td>Hearing impaired</td>
<td>2.3 1.7 0.4 1.5 0</td>
<td>2.3 1.7 0.4 1.5 0</td>
</tr>
</tbody>
</table>

Table 3.5: Differences between mean SRTs measured for N=8 and N=16 for each value of CO/N, for normal-hearing and hearing-impaired subjects in experiment two. The LSD calculated using Fishers LSD procedure was 1.6. Differences equal to or above this value are shown in bold.
3.6.4 Discussion

The pattern of results was similar for experiments one and two, for both the normal-hearing and hearing-impaired subjects. This suggests that neither the random amplitude fluctuations introduced by the noise vocoder nor the partial removal of high-rate envelope modulations by the relatively narrow filters used in experiment one entirely explain the (small) benefit of adding TFS information found for the hearing-impaired subjects in experiment one. Rather, the results are consistent with the idea that the improvement in SRT as CO increased resulted mainly from the use of TFS information, and the improvement was smaller for the hearing-impaired than for the normal-hearing subjects because the latter have a greatly reduced ability to use TFS information.

Performance was better for the tone vocoder (experiment two) than for the noise vocoder (experiment one) with matched N, but different sentence material was used for the two experiments, which makes the comparison difficult. Better performance has been reported for the ASL sentence lists used in experiment one than for the BKB lists used in experiment two (MacLeod and Summerfield 1990). If the same sentence material had been used for the two experiments, an even larger difference might have been observed. Overall, the comparison of results for experiments one and two with N=32 and CO=0 is consistent with previous results (Dorman et al. 1997, Whitmal et al. 2007, Stone et al. 2008) and with the hypothesis that the random amplitude fluctuations introduced by the noise vocoder have a deleterious effect on performance.

Normal-hearing subjects benefitted from the greater spectral information in the vocoded signal when N=32 than when N=16, whereas the hearing-impaired subjects did not. This is consistent with what would be expected from the greater auditory-filter bandwidths that are typically found for hearing-impaired subjects (Glasberg and Moore 1986, Moore 2007). The normal-hearing subjects, who were expected to have relatively sharp filters, benefitted significantly from the greater spectral information provided by more channels, while the hearing-impaired subjects, who were expected to have relatively broad filters, benefitted little, if at all. These findings are consistent with those of Başkent (2006), who found a similar plateau in performance as N increased above 16 for hearing-impaired subjects.

Previous work has concentrated on reduced frequency selectivity as an explanation for the supra-threshold deficits associated with moderate cochlear hearing loss. Reduced frequency selectivity means that hearing-impaired listeners are more susceptible to masking across frequencies, and this partially explains why they perform poorly when listening in background sounds. The different patterns of performance (i.e., the difference in benefits from increasing the value of CO) for the normal-hearing and hearing-impaired subjects in the results presented here cannot be accounted for by differences in across-frequency
masking. Similar amounts of masking would be expected in all of the conditions that were tested, so if deficits caused by cochlear hearing loss were only a result of across-frequency masking, a similar pattern of performance would have been expected for the normal-hearing and hearing-impaired subjects. Reduced spectral resolution may account for the differences in performance between the subject groups when CO=0, when no TFS information was available. Indeed, the fact that the hearing-impaired subjects were tested using higher overall sound levels than the normal-hearing subjects might have exacerbated this effect, since auditory filters tend to broaden at high levels (Glasberg and Moore 1990). However, changes in auditory-filter bandwidth with level tend to be smaller for hearing-impaired than for normally hearing subjects (Moore, 2007), so the effect of level is unlikely to be large. For whatever reasons, speech intelligibility worsens at very high sound levels for both normal-hearing and hearing-impaired subjects (Summers and Cord, 2007), so the higher level used here for the hearing-impaired subjects may have contributed to their poorer overall performance. However, the increasing deficit as TFS information was added is unlikely to reflect this rollover effect, since the speech level was the same for all values of CO.

Another possible factor that may have influenced our results is that some of the hearing-impaired subjects may not have been able to make effective use of information conveyed by the higher-frequency components in the speech, even though those components would have been audible. In other words, that lack of benefit from adding TFS information may reflect a general lack of ability to use information from the higher-frequency components in speech. However, a reduced ability to use information from the higher-frequency (> 2000 Hz) components in speech has mainly been found for subjects with hearing losses greater than about 60 dB (Ching et al., 1998; Hogan and Turner, 1998; Vickers et al., 2001). Hearing-impaired subjects with hearing losses less than 60 dB do seem to be able to make effective use of information from such high frequency components (Skinner and Miller, 1983; Vickers et al., 2001; Baer et al., 2002). Several of our subjects had hearing losses of 60 dB or less for frequencies up to about 4000 Hz, but they still failed to show a clear benefit as CO/N was increased above 0.5 (corresponding to a frequency of 1605 Hz). For example, HI 4 had audiometric thresholds of 55 dB or better for all frequencies up to 6000 Hz, but did not show any benefit of increasing CO/N.

Overall, it seems likely that the increasing deficit of the hearing-impaired subjects as CO/N was increased reflects a different ability to use TFS information between the two groups. It is possible that reduced frequency selectivity may itself contribute to a reduced ability to use TFS information, however. The outputs of broader auditory filters would have a more complex TFS than the outputs of narrower filters, as found in normal-hearing subjects. It is possible that such complex outputs may not be interpretable by the central
auditory system. Deficits in phase locking would also be expected to reduce the ability to use TFS, as inaccuracies in phase locking would degrade information about TFS available to the central auditory system.

The individual differences in benefit from the addition of TFS information found here between hearing-impaired subjects may explain the relatively poor correlation between audiometric thresholds and speech intelligibility in noise ([Festen and Plomp, 1983] [Glasberg and Moore, 1989]). For the subjects tested here, the amount of benefit gained from addition of TFS information \([\text{SRT for CO}=0]-(\text{SRT for CO}=32)\] was not significantly correlated with the mean of audiometric thresholds at 250, 500, 1000, 2000 and 4000 Hz \((r=-0.04, p=0.92)\). The ability to use TFS information may be a factor affecting speech intelligibility that is not well predicted by traditional audiometry.

### 3.7 Conclusions

Hearing-impaired subjects benefitted less than normal-hearing subjects from TFS information added to a vocoded speech signal when listening in a competing talker background. The amount of benefit varied between subjects, with some not benefiting at all. The same general pattern of results was found regardless of whether a noise vocoder or a tone vocoder was used. It is argued that subjects with moderate cochlear hearing loss have a limited ability to use TFS information. This may explain some of the speech perception deficits found for such subjects, especially the reduced ability to take advantage of temporal dips in a competing background.
Chapter 4

The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise

A version of this chapter is to be published as: Hopkins, K. and Moore, B. C. J. (2009) ‘The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise.’ J. Acoust. Soc. Am. 125(1) 442-446

4.1 Introduction

For a given signal-to-noise ratio (SNR), normally hearing subjects can identify speech better when it is presented in an amplitude-modulated noise background than in a steady noise background (Festen and Plomp 1990). Similarly, speech reception thresholds (SRTs) for normal-hearing subjects are lower in modulated noise than steady noise (Duquesnoy 1983, Peters et al. 1998). This improvement in performance when listening in modulated noise has been termed ‘masking release’. Masking release is thought to arise because subjects can listen to signal portions where the masker level is low and so the short-term SNR is high. This strategy has been labelled ‘listening in the dips’.

For this strategy to be successful, temporal dips that contain information about the signal must be identified. One cue for identification could be a reduction in the modulation depth of the combined signal, although this cue is unlikely to be very salient, especially at low SNRs. Another possible cue is the change in temporal fine structure (TFS) in the dips of a masker when there is a signal present. This could allow identification of portions
of the sound with the highest SNR. For this strategy to be successful, at least some of the TFS information of the target speech must be preserved, both in the physical signal and by the peripheral auditory system.

For subjects with cochlear hearing loss and cochlear-implant users, masking release is often small or absent (Festen and Plomp 1990; Peters et al. 1998; Nelson et al. 2003; Lorenzi et al. 2006a). Such subjects perform similarly when listening in steady and modulated noise. There are several possible reasons for this reduced ability to benefit from the dips in a masker:

1. Hearing-impaired subjects often have broader auditory filters than subjects with normal hearing (Glasberg and Moore 1986). Cochlear implants often have a small number of effective channels. This could reduce masking release in backgrounds that are modulated asynchronously across frequency, such as competing speech, because of increased spread of masking, which may mean that signals in the dips of a masker at one frequency are masked by masker components at nearby frequencies that are higher in level.

2. Hearing-impaired subjects and cochlear implant users are often more susceptible to forward masking than normal-hearing subjects (Glasberg et al. 1987). Peaks in a modulated masker would elevate the threshold in the following dips, making signal identification more difficult.

3. Current cochlear implant processors discard much TFS information, and subjects with moderate cochlear hearing loss have a reduced ability to use TFS information (Hopkins and Moore 2007, Chapter 2). TFS information may be important for identifying a signal in the dips of a fluctuating masker, as discussed above. This hypothesis is supported by data from Lorenzi et al. (2006a). They measured the ability of hearing-impaired subjects to identify vowel-consonant-vowel (VCV) syllables processed to remove envelope information but to leave TFS information nearly intact. They found that the ability to identify these VCV syllables was correlated with masking release measured for intact VCV syllables in steady and amplitude-modulated noise.

Here we investigated the importance of TFS information for masking release, to test whether a reduced ability to use TFS information could explain the reduced masking release measured for hearing-impaired subjects and cochlear implant users. We used normal-hearing subjects and tone-vocoder processing, so that the effect of manipulating TFS information could be investigated without the possible confounding effects of changes in forward masking and frequency selectivity. The amount of TFS information that
was preserved from the original signal was varied by vocoding information above a cut-off channel (CO), but leaving lower frequency information intact \cite{Hopkins2008}. The value of CO was varied and performance was measured for both steady and amplitude-modulated noise.

### 4.2 Methods

#### 4.2.1 Subjects and materials

Ten normally hearing subjects were tested. Subjects had thresholds of 15 dB HL or less at standard audiometric frequencies, and were paid for their time. The target speech for both training and testing was IEEE sentences \cite{Rothauser1969}, spoken by a male native British English speaker. Sentences were presented in noise that had the same long-term spectrum as the target speech. The noise was either steady or amplitude modulated at a rate of 8 Hz. The modulation was sinusoidal on a dB scale with a peak-to-valley ratio of 30 dB. The equation specifying the modulated noise was:

\[
F(t) = N(t) \cdot 10^{30[\sin(2\pi\cdot 8t)-1]/40}
\]  

where \(N(t)\) was the waveform of the unmodulated speech-shaped noise.

This combination of modulation rate and depth was chosen because pilot studies showed that these parameters led to a large amount of masking release with unprocessed stimuli, for the sentence material chosen for this study.

#### 4.2.2 Processing and equipment

Processing was similar to that described by \cite{Hopkins2008}. Speech and noise were summed at the required SNR and filtered into 32 channels using an array of finite-impulse-response (FIR) filters, equally spaced on an ERB\(_N\)-number scale \cite{Glasberg1990} between 100 and 10,000 Hz. The order of each filter was chosen so that its frequency response was approximately -6 dB at the point that the response intersected with the response of adjacent filters. Each filter was approximately 1-ERB\(_N\) wide, so that the filter bank roughly simulated the frequency selectivity of the normal auditory system. Channels up to and including a cut-off channel (CO) were not processed further, whereas channels above CO were tone vocoded. The Hilbert transform was used to find the analytic signal for each channel signal and the envelope was calculated as the absolute value of each analytic signal. Each channel envelope was used to modulate a sine wave with frequency equal to the channel centre frequency. Each modulated tone was subsequently filtered with the initial analysis filters. Consequently, envelope fluctuations with
frequencies greater than half of the channel bandwidth were attenuated. The unprocessed
channel signals and vocoded signals were time aligned and then combined. All signals
were generated with a high-quality 16-bit PC soundcard (Lynx One) at a sampling rate of
22,050 Hz, passed through a Mackie 1202-VLZ mixing desk and presented to the subject
monaurally via Sennheiser HD580 headphones. Subjects were seated in a double-walled
sound-attenuating chamber.

4.2.3 Training

A period of training was conducted before testing to allow familiarisation with the vocoder
processing and the procedure. Eleven IEEE lists were presented with a fixed SNR for each
list. Subjects were required to repeat each sentence and the number of correctly iden-
tified key words was recorded. When the sentence was not repeated perfectly, subjects
were told the correct answer, and the sentence was repeated. The first list had CO=32
(corresponding to unprocessed signals) and steady noise, and each of the conditions was
subsequently tested in a random order. For each list, the SNR was chosen to give perfor-
mance of around 50-80% correct. At the end of the training session, subjects completed
two runs with the same procedure used in the testing session (described below) to allow
familiarisation with the adaptive testing procedure.

4.2.4 Testing

Five values of CO (0, 8, 16, 24 and 32) were used for each noise type (steady or modulated),
making ten conditions in total. For each condition the SNR needed to achieve 50% correct
(the speech reception threshold, SRT) was measured using an adaptive procedure. Two
IEEE sentence lists were presented consecutively to make a run. The first sentence in
each run was presented at an adverse SNR at which subjects were expected to identify
no words correctly. The SNR was increased by 4 dB, and the sentence repeated until
the subject correctly identified three or more of the five key words. Subsequent sentences
were presented once only. If the subject identified three or more keywords correctly, the
following sentence was presented at a SNR that was k dB lower, and if two or fewer
keywords were correctly identified, the following sentence was presented with an SNR
that was k dB higher. For sentences before the first two turnpoints, k was equal to 4 dB;
subsequently k was equal to 2 dB.

4.2.5 Analysis

Within a run, for each SNR, the number of keywords presented and keywords correctly
identified was recorded. These data were transformed to probit units (Finney, 1971), and
linear regression analysis was used to estimate the slope and intercept parameters of the transformed data. The resulting probit function was then converted back to proportion correct units for each SNR and used as an estimate of the psychometric function for each subject and condition. The 50% correct points on these psychometric functions were identified to give SRTs that were used to compare performance across conditions.

### 4.3 Results

Subjects performed better for both noise types as CO increased and more TFS information was available. The amount of benefit produced by the additional TFS information was found by calculating the improvement in performance (the reduction in SRT) for each value of CO relative to performance when C0=0, for each noise type. These results are plotted in Figure 4.1.

![Figure 4.1: Benefit of adding TFS information as measured by the SRT relative to that for CO=0, for steady and modulated noise. Mean SRTs for CO=0 for steady and modulated noise were -0.7 and -7.4 dB respectively. The frequencies corresponding to each value of CO are shown along the top axis. Error bars show ± one standard error of the mean across subjects.](image)

For both noise types, the largest benefit occurred when CO was increased from 0 to 8.
However, for the modulated noise, performance improved further as more TFS information was added at higher frequencies. A within-subjects analysis of variance (ANOVA) on the SRTs showed that the main effects of noise type and CO were highly significant ($p<0.001$) and that there was a significant interaction between the two factors [$F(4,36)=4.9$, $p=0.003$]. A Fishers least-significant-difference (LSD) test was performed to assess the effect of CO for each noise type. The results of this analysis are shown in Table 4.1. Performance improved significantly for both noise types when CO was increased from 0 to 8. For the steady noise, no significant improvement in performance occurred as CO was increased further, but for the conditions where modulated noise was used there were significant improvements in performance for higher values of CO. For modulated noise, adding TFS information above 1605 Hz led to a significant improvement in the SRT. Thus, the benefit of TFS information is not restricted to the range covered by the voice fundamental frequency or its low harmonics.

<table>
<thead>
<tr>
<th>CO</th>
<th>Frequency (Hz)</th>
<th>0</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>548</td>
<td>1.4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1605</td>
<td>2.1</td>
<td>0.7</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>4102</td>
<td>1.7</td>
<td>0.3</td>
<td>-0.4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>10,000</td>
<td>2.3</td>
<td>0.9</td>
<td>0.2</td>
<td>0.6</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CO</th>
<th>Frequency (Hz)</th>
<th>0</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>548</td>
<td>3.5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1605</td>
<td>5.0</td>
<td>0.6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>4102</td>
<td>5.0</td>
<td>1.6</td>
<td>1.0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>10,000</td>
<td>6.2</td>
<td>6.2</td>
<td>2.2</td>
<td>1.2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.1: Differences between mean SRTs measured with different pairs of values of CO for steady and modulated noise. Equivalent cut-off frequencies (in Hz) are also shown. The LSD calculated using Fisher’s procedure was 1.3. Differences equal to or above this value are shown in bold.

Figure 4.2 shows the masking release for each value of CO (SRT for steady noise minus...
Figure 4.2: Masking release plotted as a function of CO. The frequencies corresponding to each value of CO are shown along the top axis. Error bars show ± one standard error of the mean across subjects.
the SRT for modulated noise). The amount of masking release increased as CO increased, and more of the original TFS was preserved. However, masking release still occurred when CO=0, when none of the original TFS information was preserved.

One concern with using the SRT as a measure of masking release is that psychometric functions measured using modulated noise are typically less steep than psychometric functions measured using steady noise (Qin and Oxenham, 2003; Stickney et al., 2004), so the amount of masking release that is measured may depend on the percent correct that is tracked. Here, if a higher percentage correct was tracked, the masking release might have been smaller. Additionally, if the slopes of the psychometric functions changed for different values of CO, this might partially account for the different rates of change in SRT for steady and modulated noise as CO increased (as illustrated in Figure 4.1).

To investigate these possibilities, a two-way within-subjects ANOVA was performed on the slopes of the fitted probit functions (as described earlier) for each subject and condition, with factors noise type and CO. The effect of noise type was significant [F(1,9)=13.8, p=0.005], but the effect of CO was not [F(4,36)=0.15, p=0.96], and there was no significant interaction between the two factors [F(4,36)=0.24, p=0.91]. The mean slopes, together with the associated standard deviations (SDs), are shown in Table 4.2. The significant effect of noise type on the slopes of the psychometric functions is consistent with previous studies (Qin and Oxenham, 2003; Stickney et al., 2004). The shallower slope of the psychometric functions for modulated noise than for steady noise means that the masking release measured in dB by tracking the SNR needed to achieve a particular percent correct depends on the percent correct that is tracked; the higher the percent correct that is tracked, the less masking release is measured. The pattern of results that we report here (rather than the absolute values for masking release) should, however, be the same regardless of the percent correct that is tracked, as there was no significant effect of CO on the slopes of the psychometric functions, and no significant interaction between noise type and CO.

Psychometric functions are plotted in Figure 4.3, based on the mean slopes and intercepts across subjects of the probit functions for each condition. From these psychometric functions, we derived a second measure of masking release by finding the difference in percent correct performance predicted by these psychometric functions for an SNR at which performance in the steady noise condition was predicted to be 10% correct for each value of CO. This gives an estimate of the masking release in percentage points, rather than dB, so allowing comparison of our data with previous studies that used this measure. The masking release values estimated in this way are shown in the second column of Table 4.3. The values increased progressively from 48 to 67 percentage points as CO was increased from 0 to 32. Similarly, the benefit of additional TFS information was
Table 4.2: Mean slopes of the psychometric functions estimated for each condition. The SDs of the means are also shown estimated in percentage points, by measuring the improvement in performance in steady and modulated noise as CO increased for an SNR for which performance was estimated to be 20% correct when CO=0. These results are shown in the third and fourth columns of Table 4.3. The maximum benefit (obtained with CO=32) was 31 percentage points for the steady noise and 47 percentage points for the modulated noise. The general pattern of results from these analyses is consistent with that inferred from the SRTs; masking release increased as the number of channels containing TFS information increased, and the benefit of adding TFS information was greater for modulated than for steady noise.

![Psychometric functions for each value of CO, for steady and modulated noise. The plotted functions are based on mean values of the slopes and intercepts of the probit functions fitted to the data for each subject and condition.](image)

<table>
<thead>
<tr>
<th>CO</th>
<th>Mean slope (probit units/dB)</th>
<th>SD</th>
<th>Mean slope (probit units/dB)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.39</td>
<td>0.23</td>
<td>0.20</td>
<td>0.11</td>
</tr>
<tr>
<td>8</td>
<td>0.41</td>
<td>0.13</td>
<td>0.23</td>
<td>0.09</td>
</tr>
<tr>
<td>16</td>
<td>0.40</td>
<td>0.19</td>
<td>0.24</td>
<td>0.11</td>
</tr>
<tr>
<td>24</td>
<td>0.40</td>
<td>0.17</td>
<td>0.25</td>
<td>0.11</td>
</tr>
<tr>
<td>32</td>
<td>0.38</td>
<td>0.20</td>
<td>0.26</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Table 4.3: Masking release and benefit from additional TFS information expressed in percentage points. Masking release was calculated from the difference in psychometric functions in steady and modulated noise, at an SNR for which performance in steady noise was predicted to be 10% correct. The benefit from additional TFS as CO increased was measured by finding the difference in the psychometric function at a SNR for which performance when CO=0 was predicted to be 20% correct.

<table>
<thead>
<tr>
<th>CO</th>
<th>Masking release</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Steady noise</td>
</tr>
<tr>
<td>0</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>59</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>60</td>
<td>27</td>
</tr>
<tr>
<td>24</td>
<td>65</td>
<td>19</td>
</tr>
<tr>
<td>32</td>
<td>67</td>
<td>31</td>
</tr>
</tbody>
</table>

4.4 Discussion

The greater benefit from the addition of TFS information for the modulated masker than for the steady masker, and the increased masking release as CO increased both suggest that TFS information is particularly important for listening in a modulated background. The same patterns of results were found when differences in performance were expressed as changes in SRT and as predicted differences in percent correct performance at a particular SNR. TFS information could be important for identification of signals in masker dips, so directing attention to the portions of a signal with the most favourable SNR.

Significant (though reduced) benefit was also gained from TFS information at low frequencies when listening in steady noise, suggesting that TFS information is not important only for identification of signals in masker dips. TFS information could also allow better fundamental frequency identification. Current pitch perception models are based on the assumption that TFS information is important for low, resolved harmonics, and it has been suggested that TFS information may also be important for pitch perception based on harmonics with intermediate harmonic numbers that are unresolved (Moore et al. 2006a). Fundamental frequency may be an important grouping cue, allowing separation of the target and background into different auditory streams. Information about the fundamental frequency of the target speech would mainly be present at low frequencies, where there are resolved harmonics. The improvement in performance as TFS information was added up to 548 Hz, but not for higher frequencies (for the steady noise), is consistent with this idea.
It is possible that the vocoder processing used in the present study may have had an adverse effect on the temporal envelope information available to the central auditory system, as well as the TFS. The widths of the analysis filters were chosen to be similar to those of the normal auditory filters, and so auditory filters centered at the analysis channel center frequencies should have had similar temporal envelope cues at their outputs for the vocoded and unprocessed signals. However, only 32 contiguous filters were used, which is somewhat different to the peripheral auditory system, which can be thought of as an array of many overlapping filters. Consequently, the temporal envelope information at the outputs of auditory filters centered between the analysis filters in the vocoder processing scheme would have been somewhat degraded. This could partly account for the higher SRTs when CO was low.

Previous studies measured masking release for cochlear implant users and found little difference in performance when listening in steady or modulated noise (Nelson et al., 2003). Similar results were found when using vocoder simulations of cochlear implant processing with normal-hearing subjects (Qin and Oxenham, 2003). This is in contrast with the present study, where 6.7 dB of masking release was measured when CO=0, and the signal was entirely vocoded. A possible explanation for this difference is the number of channels used in the vocoder processing. In this study, 32 1-ERB\(_N\)-wide channels were used to simulate the frequency selectivity of the normal auditory system. Cochlear implants have fewer effective channels, either because only a few electrodes are used, or because of current spread between electrodes. Similarly, previous vocoder simulations often used fewer channels than the current study. The current data suggest that masking release can occur with little or no TFS information, provided that frequency selectivity is normal. This is consistent with the results of Gnansia et al. (2008), who measured the intelligibility of VCV signals in steady and modulated noise, processed to remove TFS information. As in the current study, many channels were used, so frequency selectivity was similar to that of the normal auditory system. Gnansia et al. (2008) found that performance was about 24 percentage points higher when identifying VCV syllables in modulated noise than in steady noise when the TFS information was removed. A larger improvement was seen when the stimuli were unprocessed, but the difference in masking release was smaller than measured in the present study. Similarly, Qin and Oxenham (2003) measured a small amount of masking release for sentences processed with a 24-channel vocoder (2.1 dB lower SRT for a modulated than for a steady noise masker) and Fu and Nogaki (2005) measured a small amount of masking release for sentences processed with a 16-channel vocoder, but no masking release when the number of vocoder processing channels was reduced to four. The smaller masking release values reported in these previous studies probably reflect differences in the maskers used. The current study
used a noise that was sinusoidally amplitude modulated on a dB scale, whereas Gnansia et al. (2008) used noise that was sinusoidally amplitude modulated on a linear scale. Noise that is amplitude modulated on a dB scale has more of its energy concentrated at the peaks of the waveform, with longer effective dips. This may give more opportunity for dip listening and consequently more masking release.

4.5 Conclusions

Adding TFS information to tone-vocoder-processed speech improved performance more when listening in modulated noise than in steady noise, suggesting that TFS information contributes to masking release for normal-hearing subjects. A loss of ability to use TFS information in subjects with cochlear hearing loss and subjects with cochlear implants could partially account for the reduced masking release observed for such subjects.
Chapter 5

The importance of TFS information in speech at different spectral regions for normal-hearing and hearing-impaired subjects

5.1 Introduction

Temporal fine structure (TFS) information improves speech intelligibility for normal-hearing subjects when listening in background noise. Speech reception thresholds (SRTs) for a target talker with a competing-talker or noise background are raised when the combined signal is noise- or tone-vocoded, even when such processing uses similar analysis bandwidths as peripheral auditory filters so that the temporal envelope and spectral information available to the central auditory system are similar to those for unprocessed speech (Gnansia et al. 2008, Hopkins et al. 2008, Hopkins and Moore 2009). Nonsense syllables processed to preserve TFS but to remove, as far as possible, temporal envelope information are highly intelligible to normal-hearing subjects, following some training, suggesting that TFS may also have a role in speech intelligibility in quiet (Lorenzi et al. 2006a). However, the role of TFS information remains uncertain. A number of possibilities have been suggested, and these are described below.

Fundamental frequency (F0) coding

Considerable research has addressed the importance of TFS information in pitch perception, particularly in coding the frequencies of pure tones and resolved harmonics in complex tones. Psychophysical data support the hypothesis that TFS information is
important for the perception of pure tones for frequencies up to at least 5000 Hz. For example, Moore (1973) measured frequency difference limens (DLFs) for short-duration pure tones, which had a broader frequency spectrum than longer tones of the same frequency. DLFs were better than was predicted by a model that assumed that discrimination of pure tones was based on detection of a difference in the excitation pattern (Zwicker 1970). TFS information has also been suggested as a possible cue for coding F0 information for complex tones with components that are unresolved but intermediate in harmonic number (harmonic ranks 8-14) (Moore and Moore 2003b, Moore et al. 2006a), although this role remains uncertain, with other researchers suggesting that F0 coding based on these harmonic ranks can be accounted for by coding the repetition rate of the waveform using envelope information rather than TFS, or on the basis of partially resolved harmonics (Bernstein and Oxenham 2006). F0 information in speech has been shown to be important for conveying intonation, and semantic information in the case of tonal languages. In the presence of a background sound, F0 information may be important for segregating sound sources into separate auditory streams, particularly when the background is a competing talker (Brokx and Nooteboom 1982, Scheffers 1984, Bregman et al. 1990).

**Coding the frequencies of peaks in the short-term speech spectrum**

The coding of formant frequencies is important for vowel identification. Some animal studies have shown that a rate-place code carries insufficient information to account for psychophysical measures of vowel discrimination at all sound levels, as at high levels auditory filters broaden and peaks in the rate-place profile at the vowel formants become indistinct (Sachs and Young 1979). However, Young and Sachs (1979) described representations of vowel formants using rate, place and temporal information that were stable up to levels of at least 80 dB SPL. These results suggest that temporal information plays some role in coding the frequencies of peaks in the spectra of speech signals.

**Listening in the dips**

It has been proposed that TFS information is important when listening to target speech in a background that is modulated in amplitude. In such a background, subjects can take advantage of a short-term improvement in signal-to-background ratio (SBR) to ‘glimpse’ information about the target. Normal-hearing subjects have lower SRTs in modulated noise than steady noise (Duquesnoy 1983, Peters et al. 1998), and they show better word and syllable identification for signals presented at the same long term signal-to-noise ratio (SNR) in modulated than in steady noise (Festen and Plomp 1990). This improvement in performance has been termed ‘masking release’. Hopkins and Moore (2009, Chapter
measured SRTs for a tone-vocoded target and background signal in which unprocessed information was progressively introduced, starting in the low-frequency channels. They found that normal-hearing subjects benefitted more from TFS information when listening in a modulated-noise background than a steady-noise background. They suggested that subjects could use a change in TFS information at the dips of a modulated masker to identify the presence of a signal. The idea that TFS information is important for masking release is supported by the positive correlation found by Lorenzi et al. (2006a) between the ability of young hearing-impaired subjects to understand speech processed to contain primarily TFS information, and masking release for intact speech in noise.

Hopkins et al. (2008) measured the improvement in SRT as TFS information was added for a combined target and competing-talker signal. The signal was split into 32 channels that were 1-ERB\(_N\) wide (Glasberg and Moore, 1990). Above a cut-off channel, CO, channels were noise vocoded, and so contained only temporal envelope information. For channels up to and including CO, channels were not processed, and so contained both TFS and envelope information. For normal-hearing subjects, performance improved as CO was increased and more TFS information was included in the combined signal. The biggest improvement in performance occurred when CO was increased from zero to four, and the improvement as CO was raised further was gradual. No significant improvement in performance occurred when CO was increased beyond 24 (equivalent to a frequency of 4102 Hz). It may initially appear that this supports the hypothesis that TFS information is mainly important at low frequencies, and that TFS information is not at all important for frequencies above about 4000 Hz. However, this result could be influenced by redundancy in the TFS information that was added to the signal as CO was increased. If there was redundancy in the TFS information at different spectral regions, adding more TFS information would not necessarily result in an improvement in performance.

This hypothesis was tested in experiment one of this chapter. A similar method to that of Hopkins et al. (2008) was used, but as well as TFS information being added starting with the low-frequency channels (TFS-low conditions), TFS information was added starting with the high-frequency channels (TFS-high conditions). If benefit from TFS information is not affected by redundancy of information between channels, the change in performance between different values of CO for TFS-high and TFS-low conditions should be similar.
5.2 Method

5.2.1 Subjects and Materials

Seven normal-hearing subjects were tested. All had thresholds of 15 dB HL or better in their test ear at standard audiometric frequencies, and were aged between 19 and 24 years. All subjects were native British English speakers. Sentences from the co-ordinate response measure (CRM) corpus were used for testing (Moore, 1981; Kitterick and Summerfield, 2007). CRM sentences all have the same standard format: Ready callsign, go to colour, number now! Subjects were required to identify the callsign, colour and number that they heard from a list of four (in the case of colour) or eight (in the case of callsign and number) alternatives. Sentences were presented in a competing-talker background. The background was randomly selected from a passage of continuous prose, spoken by a different speaker. Silent intervals between sentences and pauses for breath were removed from the passage by hand editing. Natural-sounding pauses between words of less than 200 ms were left intact. Both target and competing speakers were male speakers of British English. The F0 of the target talker’s voice was estimated to range from 130 to 200 Hz, and the competing talker had a larger F0 range of about 130 to 280 Hz.

5.2.2 Processing and equipment

Target and competing-speech signals were added together and subsequent processing was carried out on the combined signal. The signal was filtered into 30 channels spanning a frequency range of 100 to 8000 Hz with an array of linear-phase, finite-impulse-response (FIR) filters. The filters had a variable order so that the transition bands of each filter had similar slopes when plotted on a logarithmic frequency scale. Each filter was designed to have a response of -6 dB at the frequencies at which its response intersected with the responses of the two adjacent filters. Each channel was 1-ERBN wide. Hence the processing was not expected to have a large effect on the spectral information available to the central auditory system. Channels were divided into two groups by a cut-off channel, CO. For TFS-low conditions, channels up to and including CO were not processed further and channels above CO were tone vocoded. For TFS-high conditions, channels above CO were not processed further, and channels up to and including CO were tone vocoded. For channels that were vocoded, a Hilbert transform was used to find the analytic signal for each channel (Hilbert, 1912; Bracewell, 1986), and the envelope was extracted by finding the absolute value of this analytic signal. Each channel envelope was used to modulate a sine wave with a frequency equal to the centre frequency of the channel. Each modulated sine wave was then filtered using the original channel analysis filters to remove side bands.
that fell outside of the channel bandwidth. Finally, signals from all channels were aligned in time and combined.

All signals were generated with a high-quality 16-bit PC soundcard (Lynx One) at a sampling rate of 22050 Hz, passed through a Mackie 1202-VLZ mixing desk and presented to the subject monaurally via Sennheiser HD580 headphones. Subjects were seated in a double-walled sound-attenuating chamber.

5.2.3 Procedure

Six values of CO were used (0, 6, 12, 18, 24 and 30). Note that the TFS-low condition with CO=0 was equivalent to the TFS-high condition with CO=30 and the TFS-high condition with CO=0 was equivalent to the TFS-low condition with CO=30. These were the conditions for which the whole signal was vocoded (the former) or unprocessed (the latter). All other values of CO were tested for both TFS-low and TFS-high conditions, giving ten conditions in total. SRTs corresponding to the 70.7% correct point on the psychometric function were measured for each subject and condition. At the start of each run, the SBR was set to a value for which the subject was expected to identify all of the words in the target sentence correctly. A sentence was considered correct if the callsign, colour and number were correctly identified and incorrect if any mistakes were made. A two-up, one-down adaptive procedure was used to set the SBR for each trial. If the sentence was correctly identified twice in succession, the SBR for the next sentence was set k dB lower; if a sentence was incorrectly identified, the next sentence was presented at a SBR that was k dB higher. For the first two turnpoints, k was equal to 4 dB and for the remaining six turnpoints, k was equal to 2 dB. The SRT was taken to be the mean of the SBR at the last six turnpoints. For SBRs of -16 dB or higher, the level of the target talker was fixed at 65 dB SPL, and the level of the competing talker was varied. For SBRs lower than -16 dB, to prevent the level of the combined signal becoming uncomfortably loud, the level of the competing talker was not increased further; instead the level of the competing talker was fixed, and the level of the target talker was reduced to give the correct SBR. Subjects completed three adaptive ‘runs’ for each condition, and the order of the presentation of conditions was randomised for each subject. The experiment was run alongside experiment two (see later), and conditions for the two experiments were interleaved. Subjects completed a ‘set’ of all the conditions before any of the conditions were repeated.
5.3 Results

SRTs as a function of CO are plotted in Figure 5.1 for both the TFS-low and TFS-high conditions. For the TFS-low condition performance improved as CO increased and more TFS information was presented. A one-way, repeated measures, analysis of variance (ANOVA) showed that there was a significant effect of CO [F(5,30)=12.3; p<0.001]. A Fisher’s least significant difference (LSD) procedure was conducted to test whether SRTs for different values of CO were different from each other. The results are shown in the top panel of Table 5.1. SRTs improved for values of CO up to 12, but there was no significant improvement for higher values of CO.

For the TFS-high condition, performance improved as CO was reduced. The effect of CO was significant [F(5,30)=10.03; p<0.001]. LSD test results are shown in the bottom panel of Table 5.1. The pattern of the results broadly mirrored that for the TFS-low condition, with most change in performance when TFS was added to the low-frequency channels. However, performance significantly improved even when TFS information was added to the highest-frequency channels, a result that was not found for the TFS-low condition or by Hopkins et al. (2008).

![Figure 5.1: SRTs plotted as a function of CO for TFS-high and TFS-low conditions. Error bars are plotted at ± one standard error of the mean.](image-url)
Table 5.1: Differences between mean SRTs measured with different pairs of values of CO for TFS-high and TFS-low conditions. Equivalent cut-off frequencies (in Hz) are also shown. The LSDs calculated using Fisher’s procedure were 2.3 (TFS-low condition) and 2.1 (TFS-high condition). Differences equal to or above this value are shown in bold.
5.4 Discussion

For both the TFS-low and TFS-high conditions, performance improved as more TFS information was included in the combined signal. For the TFS-low condition, performance improved as CO was increased. This is broadly consistent with the results of Hopkins et al. (2008), although Hopkins et al. (2008) found significant improvements for higher values of CO (up to CO=24 rather than up to CO=12), and the improvement in SRT from adding TFS across the whole spectrum was much larger (15.5 dB compared with 6.5 dB). A number of factors could account for these differences. Firstly, Hopkins et al. (2008) used a noise vocoder, whereas a tone vocoder was used here. Whitmal et al. (2007) compared the intelligibility of noise and tone-vocoded stimuli and found that the noise-vocoded stimuli were less intelligible, particularly when very narrow bands were used, similar to those used here and by Hopkins et al. (2008). Similarly Stone et al. (2008) found that intelligibility was significantly higher for a tone vocoder than a noise vocoder, when listening to speech in a competing talker background. Whitmal et al. (2007) suggested that this difference in performance arose because the narrow-band noise carriers used for the noise vocoder introduced random modulations, which could be confused with the envelope modulations in the speech signal. A tone carrier has a flat envelope, and so does not introduce these spurious modulations in the envelope of the vocoded signal. Distortions introduced by the noise vocoder could partly account for the poorer performance when CO was low, and a large proportion of the signal was vocoded.

For both the TFS-low and TFS-high conditions, the greatest changes in SRT occurred when TFS information was added to channels with centre frequencies below 1000 Hz, which is consistent with the idea that TFS information is important for coding F0 information. However, performance was significantly better when CO=24 than when CO=32 for the TFS-high condition. For the TFS-low condition, there was no significant difference in performance when CO=24 and CO=32. This suggests that there is some redundancy in TFS information across spectral regions; adding TFS information to the highest frequency region only improved performance when the rest of the channels were vocoded (in the TFS-high condition). When the other channels were unprocessed (as in the TFS-low condition), no improvement occurred, perhaps because the TFS information in that region was similar to the TFS information in lower-frequency channels. This could be the case if TFS information was important for 'listening in the dips'. TFS information in a narrow spectral region could be used to identify signal portions in the dips of a fluctuating masker. Envelope information in speech is highly correlated across frequencies, so this information could be used to identify speech portions with a favourable SBR across the broadband signal. If this were the case, a lot of redundancy in the TFS information would be expected across frequencies.
5.5 Experiment two

5.5.1 Rationale

The results of experiment one suggest that there is some redundancy in TFS information across spectral regions. Additionally, the improvement in performance when TFS information is added to a particular spectral region could be affected by the distribution of speech information across the frequency spectrum. For example, if TFS was added to a region that contributed little to speech intelligibility, a smaller improvement in performance might be expected than if TFS information was added at a spectral region that contributed a lot to speech intelligibility, even if TFS information was equally important in the two cases. This is a particular concern when testing hearing-impaired subjects, as access to TFS information at a particular spectral region could be limited by the audibility of the signal in that region, even when frequency-dependent gains are applied to try to ensure audibility, as was the case here (see below). In experiment two, the importance of TFS information in different spectral regions was assessed for normal-hearing and hearing-impaired subjects. The importance of envelope information within each spectral region to speech intelligibility was also assessed.

The frequencies at which TFS information is important would be expected to be different depending on the type of information that is coded by TFS. If TFS information is mainly used for coding F0, TFS information should mainly be important for frequencies below about 1000 Hz, where resolved harmonics are present. If TFS information is used for coding information about the frequency spectrum of a sound, or for listening in the dips of a fluctuating masker, then TFS information could be important for frequencies above 1000 Hz, provided that the TFS information at these frequencies is available to the central auditory system.

5.5.2 Methods

Subjects

The seven normal-hearing subjects who took part in experiment one also took part in experiment two. Ten hearing-impaired subjects also took part. The audiometric thresholds of the test ears of the hearing-impaired subjects are shown in Figure 5.2. The age of each subject is also shown. The better hearing ear was used as the test ear for all subjects apart from HI 10. The worse hearing ear was used as the test ear for HI 10 because this ear had previously been used for other psychoacoustic experiments in the laboratory. For the hearing-impaired subjects, the TEN (HL) procedure described by Moore et al. (2004) was used to diagnose cochlear dead regions. The results for the hearing-impaired subjects
gave no evidence for cochlear dead regions between 500 and 4000 Hz in their test ears. None of the hearing-impaired subjects had air-bone gaps greater than 15 dB between 500 Hz and 4000 Hz, indicating that their hearing-losses were cochlear in origin.

Figure 5.2: Audiometric thresholds of the test ears of the hearing-impaired subjects. The age of each subject is shown in brackets.

**Stimuli, processing and equipment**

As for experiment one, the target and background talker signals were summed prior to further processing. The combined signal was filtered into 30 channels, as before, and the channels were divided into five spectral regions, each 6-ERBN wide. For each spectral region (1, 2, 3, 4 and 5), two conditions were tested: one where all of the information from that region was discarded (‘-’ conditions, 1-, 2-, 3-, 4- and 5-), leading to a spectral gap, and one where there was unprocessed information in that spectral region (‘+’ conditions, 1+, 2+, 3+, 4+ and 5+). For both + and - conditions, channels in other spectral regions were tone vocoded (as described for experiment one). For the hearing-impaired subjects, two comparison conditions were included: a condition where all channels were tone vocoded (allvoc), and a condition were all channels were unprocessed (allunproc). These conditions had already been included for the normal-hearing subjects, as part of experiment one. For the normal-hearing subjects, as in experiment one, the target speech was presented at a level of 65 dB SPL for SBRs of -16 dB or higher. The level of the competing talker was set to give the desired SBR. For SBRs lower than -16dB the level of the competing talker was not increased further to prevent the combined signal becoming uncomfortably loud. Instead the level of the target talker was reduced to give
the desired SBR. For the hearing-impaired subjects, gains were applied to the combined signal according to the ‘Cambridge formula’ hearing aid fitting prescription (Moore and Glasberg, 1998). Gains were specified at audiometric frequencies between 250 and 6000 Hz. The Cambridge formula gains are designed to ensure speech audibility between these frequencies. The sentence material and test equipment was the same as that used as for experiment one.

**Procedure**

The procedure was the same as for experiment one, except that the hearing-impaired subjects completed five runs per condition rather than three, so that the results for each subject could be compared more accurately, as previous studies have shown large individual differences among hearing-impaired subjects in their ability to use TFS (Lorenzi *et al.*, 2006a; Hopkins *et al.*, 2008). For the normal-hearing subjects, experiments one and two were run simultaneously. Conditions from experiments one and two were interleaved during testing, and in cases where conditions for the two experiments were the same (for example the TFS-low CO=6 condition in experiment one and the 1+ condition in experiment two), testing was not duplicated.

### 5.5.3 Results

The mean results for the normal-hearing subjects are shown in Figure 5.3. The benefit (improvement in SRT) from adding envelope information (open circles) was calculated as the difference in the SRT for the allvoc condition and the - condition for each spectral region. Similarly, the benefit from adding envelope and TFS information together (filled circles) was calculated as the difference between SRTs for the - and the + conditions for each spectral region. A two-way within-subjects analysis of variance (ANOVA) on the mean benefit scores for the normal-hearing subjects showed significant effects of both spectral region (1, 2, 3, 4, 5) and information type (envelope and envelope + TFS) [F(4,24)=4.65; p=0.006 and F(1,6)=6.46; p=0.04, respectively], but no significant interaction between the factors [F(4,24)=0.66, p=0.63] suggesting that the relative benefit of TFS and envelope information was similar for each spectral region.

Mean results for the hearing-impaired subjects are shown in Figure 5.4. Data are plotted in the same way as for Figure 5.3. A two-way within-subjects ANOVA on the mean benefit scores for the hearing-impaired subjects showed that there was a significant effect of spectral region [F(4,36)=6.41; p<0.001], but no significant effect of information type [F(1,9)=0.03; p=0.87], suggesting that as a group, there was no significant extra benefit of TFS information over that provided by envelope information. There was no
Figure 5.3: Mean results for the normal-hearing subjects. Error bars are plotted at ± one standard error of the mean.

significant interaction between spectral region and the type of information that was added \[ F(4.36)=1.11; p=0.37 \]. However, there was a large amount of variability in the pattern of results across hearing-impaired subjects. The individual results are shown in Figure 5.5. Variability within individual subjects was also high, despite five SRT measurements per condition. Mean standard errors for the SRT measurements for individual subjects ranged from 0.6 to 1.5 dB. Some subjects showed a similar pattern of results as the normal-hearing subjects (e.g. HI 6 and HI 3, except for region 4), whereas other subjects showed no benefit from addition of TFS in most spectral regions (e.g. HI 1 and HI 4). A two-way ANOVA was performed on the individual benefit scores for each hearing-impaired subject with factors spectral region and information type. The interaction term was used as an estimate of the error term for the ANOVA. The effect of spectral region was significant \((p<0.05)\) for subjects HI 1, HI 2, HI 4, HI 5, HI 6, HI 7 and HI 10, but the effect of information type was significant only for HI 6.

The mean benefit of adding TFS to the whole frequency spectrum (performance for the allunproc condition relative to performance for the allvoc condition) was 6.5 dB for the normal-hearing subjects, and 2.5 dB for the hearing-impaired subjects. A two-way, mixed design ANOVA with a between-subjects factor of subject type (hearing impaired and normal hearing) and a within-subject factor of condition (allenv and allunproc) was performed on the SRTs for each subject for the two conditions. The effects of sub-
ject type and condition were both significant \[ F(1,15)=16.7; p<0.001 \] and \[ F(1,15)=43.8; p<0.001 \], respectively], and there was a significant interaction between the two factors \[ F(1,15)=10.0; p=0.006 \]. The individual scores for the hearing-impaired subjects are presented, together with the results of experiment three, in Section 5.6.2.

![Figure 5.4: Mean results for the hearing-impaired subjects. Error bars are plotted at \pm one standard error of the mean.](image)

5.5.4 Discussion

The similar benefit of adding TFS information to the already-present envelope information for each of the spectral regions tested for the normal-hearing subjects suggests that TFS information is important for coding information in addition to F0. If TFS information was only important for coding the F0s of the target and competing talkers, the most benefit would be expected for spectral regions one and two, where resolved harmonics would have been present. The results are consistent with the idea that TFS information is important across the frequency spectrum. TFS could be important for coding the frequencies at which there are spectral peaks. For example, if there was a vowel formant at frequency \( f_x \), neurons at a place on the cochlea tuned to \( f_x \) would phase lock to that frequency, and this would convey information about \( f_x \). This is consistent with physiological evidence showing that there is insufficient information in the rate-place code to represent vowel formants at high levels. If information from the timing of spike discharge is included,
Figure 5.5: Individual results for the hearing-impaired subjects. The bottom-right panels show the mean results for the hearing-impaired and normal-hearing subjects, for comparison. Error bars are plotted at ± one standard error of the mean.
then the representation of vowel formants is stable over a wide range of sound levels.

Another possible role for TFS information is related to ‘listening in the dips’. Hopkins and Moore (2009) showed that normal-hearing subjects benefitted more from TFS information when listening in modulated noise than in steady noise. TFS information could have been used for dip listening at any frequency region, provided the TFS information was available to the central auditory system.

The benefit from TFS information at high frequencies (above 4000 Hz) is somewhat surprising, as it has traditionally been assumed that phase locking information above 4000-5000 Hz is unusable. This assumption comes from physiological data in mammalian species other than humans, showing that phase locking to pure tones above this frequency is much degraded. Additionally, human psychophysical measurements that are assumed to reflect temporal mechanisms at low frequencies often worsen above 4000-5000 Hz (for example, pure-tone frequency discrimination), suggesting that temporal information is badly preserved at such frequencies. However, the limit of phase locking in humans is unknown. Heinz et al. (2001a) used an auditory nerve model to investigate whether human psychophysical performance could be accounted for by rate-place information alone at high frequencies. They used filters that had the same properties as human auditory filters (measured psychophysically) and phase locking characteristics that were the same as those measured in the cat. They found that psychophysical performance was best predicted when both rate-place and temporal information were included in the model, for frequencies up to at least 10,000 Hz. This result suggests that even if phase-locking characteristics in humans are similar to those in other mammalian species, some TFS information could be useful even at very high frequencies. This hypothesis is supported by data of Moore and Sek (2008b), who showed that normal-hearing subjects could discriminate complex tones on the basis of their TFS up to frequencies of at least 8000 Hz. They used the same method as described in Section 5.6 of this chapter (see later), but with high fundamental frequencies (800 and 1000 Hz) and with the bandpass filter centred at 14F0. Moore and Sek (2008b) suggested that subjects may be able to use TFS in unresolved harmonics of complex tones at frequencies for which coding of pure tones using TFS is much reduced. They argued that, at high frequencies, the precision of phase locking may be limited by refractoriness in the responses of auditory nerve fibres because the refractory period of nerve fibres would be long relative to the time between peaks in the stimulus waveform. This means that the responses of many auditory nerve fibres would need to be compared to determine the frequency of a pure tone because a single nerve fibre would be in a refractory state for many periods of the waveform. The situation is different for a complex tone with the same centre frequency that contains only harmonics that are unresolved. In this case, nerve fibres probably only fire in response to a few TFS peaks that are close to maxima.
in the envelope of the waveform. The envelope repetition rate is equal to the F0 of the complex, which is a much longer time period. Measurement of this time period would be much less affected by refractoriness in auditory nerve fibres; the F0 of the waveform could be coded by a smaller population of auditory nerve fibres.

The results for the hearing-impaired subjects are consistent with those of Hopkins et al. (2008); the hearing-impaired subjects benefitted less than the normal-hearing subjects from TFS information in speech, both when it was added to the whole spectrum, or to a narrow spectral region. The benefit from the addition of envelope information alone was similar or larger to that found for normal-hearing subjects, which suggests that the gains that were applied to the signal according to the audiogram of each subject were successful in restoring audibility. The variability in the results across hearing-impaired subjects is also consistent with the findings of Hopkins et al. (2008), although the results for individual subjects were also rather variable, and so it was difficult to draw conclusions from the patterns of results for individual subjects. The lack of a significant extra benefit from TFS information for most of the hearing-impaired subjects could be due to the large standard errors associated with each benefit score. The differences across individual hearing-impaired subjects is discussed further in Section 5.6.3.

5.6 Experiment three: Measuring sensitivity to TFS

The results of experiment two showed a large amount of variability between hearing-impaired subjects’ ability to benefit from TFS information in speech when TFS information was added to the whole frequency spectrum. In experiment three, TFS sensitivity was measured for the same hearing-impaired subjects using a psychophysical procedure. The results of the two experiments were compared to investigate the relationship between TFS sensitivity in complex tones and the ability to use TFS information in speech in a competing talker background.

5.6.1 Methods

The procedure described by Moore and Sek (2008a) was used, which is based on the experiments described in Hopkins and Moore (2007, Chapter 2). The ten hearing-impaired subjects who took part in experiment two also took part in experiment three. Four different normal-hearing subjects were also tested. Normal-hearing subjects were aged between 21 and 24 years. None were musically trained. The task was designed as a quick method for measuring TFS sensitivity. The procedure took approximately one hour, including task familiarisation and practice runs. The same equipment was used as for
experiments one and two.

**Stimuli**

Subjects were required to discriminate harmonic tones (H) and frequency-shifted tones (S) in which each component was shifted upwards by the same amount in Hz ($\delta$). Such complexes have the same envelope repetition rate, but different TFS. Components were filtered using a fixed pass band to minimise differences in the excitation patterns evoked by the harmonic and frequency-shifted complexes. The passband of the filter was $5F_0$ wide, and the skirts of the filter had a slope of 30 dB/octave. Two fundamental frequencies ($F_0$s) were tested (129 Hz and 264 Hz) and the band pass filter was centred on the 11th harmonic in each case (corresponding to centre frequencies of 1419 and 2904 Hz, respectively). These centre frequencies were the same as the centre frequencies of spectral regions three and four in experiment two. Complexes were presented at a level of 20 dB SL, in threshold equalising noise (at a level that was 15 dB below the overall level of the complexes, specified in a 1-ERB$_N$ wide band centred at 1000 Hz) (Moore *et al.*, 2004), which was designed to mask any combination tones and components on the skirts of the filter that were well away from the pass band.

**Procedure**

A two-interval, two-alternative forced-choice task was used. Each interval contained four 200-ms tones. In one interval, all of the tones were harmonic (HHHH), and in the other the tones alternated between harmonic and frequency-shifted (HSHS). The subject was required to identify the interval that contained the frequency-shifted tones. $\delta$ was varied adaptively in a two-up, one-down procedure. $\delta$ was increased by a factor of $k$ following an incorrect response, and reduced by a factor of $k$ following two consecutive correct responses. For trials before the first turnpoint, $k$ was equal to 1.953 ($1.25^3$). Then $k$ was set to 1.5625 ($1.25^2$) until the second turnpoint had occurred; for trials after the second turnpoint, $k$ was equal to 1.25. The threshold corresponding to 70.7% correct was estimated as the geometric mean of the frequency shifts at the last six turnpoints.

The largest difference between a harmonic and frequency-shifted complex occurs for a frequency shift of $0.5F_0$. For some runs, a frequency shift larger than this was dictated by the adaptive procedure. If this happened three times in a run, then the procedure switched to a non-adaptive procedure. For these cases, the frequency shift was set at $0.5F_0$, 20 further trials were conducted, and the percentage of correct responses was measured.
Task familiarisation and testing

The absolute thresholds for pure tones at each centre frequency were determined at the beginning of the testing session using a two-down one-up adaptive procedure, so that the level of the harmonic and frequency shifted tones could be set accurately at 20 dB SL.

Hopkins and Moore (2007) reported that some hearing-impaired subjects performed no better than chance in a task similar to the one used here. To make sure that all subjects understood the task, a run was included that used the same procedure as the testing run, but different stimuli. Subjects were required to discriminate two harmonic tones with different F0s rather than harmonic and frequency-shifted tones. The same procedure and centre frequencies were used as for testing. All of the subjects could perform the adaptive procedure for this task. Moore and Sek (2008a) reported only very small training effects for the stimuli and procedure described here. Consequently, just one training run was included with the testing stimuli for each condition. Subjects then completed three testing runs for each condition.

Analysis

In order to compare thresholds obtained from the adaptive procedure and percent correct scores from the non-adaptive procedure, both measures were converted to \( d' \) values (Green and Swets, 1974) using the method described by Hopkins and Moore (2007). Conversion was by means of a table of \( d' \) values for m-alternative forced-choice procedures (Hacker and Ratcliff, 1979). The adaptive procedure tracked the 70.7% correct point on the psychometric function, which corresponds to a \( d' \) of 0.78 for a two-alternative forced-choice task. When the adaptive procedure was used, the \( d' \) value that would have been measured for a difference of 0.5F0 Hz was calculated by dividing 0.78 by the threshold measured in the adaptive procedure, and multiplying this value by 0.5F0. This is based on the assumption that \( d' \) is proportional to \( \delta \). The square root of the absolute value of the mean \( d' \) (with the sign of the values restored following transformation) was used in all of the statistical analyses, as this transformation results in roughly equal variance across conditions (Hopkins and Moore 2007).

5.6.2 Results

A two-way, mixed model ANOVA was performed to assess the between-subject effect of subject type (normal hearing or hearing impaired) and the within-subject effect of centre frequency (1416 or 2904 Hz) on performance. The effect of subject type was highly significant \( [F(1,12)=40.5; p<0.001] \) but there was no significant effect of centre frequency, or interaction between the factors \( [F(1,12)=2.5; p=0.19 \text{ and } F(1,12)=0.3; p=0.59, \text{ rese-} \)
Consequently, further analysis was carried out on the mean $d'$ scores for the two centre frequencies for each subject. In Figure 5.6, the benefit from adding TFS information to the whole frequency spectrum (measured in experiment two) and the $d'$ values for TFS sensitivity are plotted for each hearing-impaired subject. The mean results for the normal-hearing subjects are shown for comparison. The benefit from adding TFS information to the whole frequency spectrum (the difference in SRT for the unprocessed condition compared with the conditions where all frequency channels were vocoded) is plotted in the bottom panel. The results are plotted in order of performance and the results for the TFS1 test (experiment three) are plotted, with subjects in the same order in the top panel. The results of the two experiments were significantly correlated (Pearson’s correlation coefficient, $r=0.66$, $p=0.04$).

Figure 5.6: Individual results for the hearing-impaired subjects and mean results for the normal-hearing subjects for experiments two and three. Benefit from adding TFS information to the whole spectrum measured in experiment two is plotted in the bottom panel, ordered by performance. Results from experiment three are plotted in the top panel, for the same subjects. Error bars show ± one standard error of the mean.
5.6.3 Discussion

The significantly better performance of the normal-hearing subjects than the hearing-impaired subjects in discriminating complex tones on the basis of their TFS is consistent with the results of [Hopkins and Moore (2007)], who showed that hearing-impaired subjects with moderate cochlear hearing loss were very poor at discriminating similar stimuli. The results for the normal-hearing subjects are consistent with the results of [Moore and Sek (2008a)], who used the same method as was used here. They found that, for an F0 of 200 Hz and centre frequency of 2200 Hz, normal-hearing subjects with little training had a mean threshold of 22 Hz, which is equivalent to a d’ of 3.5, calculated using the method described in Section 5.6.1.

The correlation between the benefit obtained from the addition of TFS to the whole speech signal and the sensitivity to TFS measured using the TFS1 test was modest, but significant. The benefit scores for the addition of TFS information to the whole speech signal were associated with large standard errors for individual subjects. The correlation might have been larger if the benefit scores had been measured more accurately, and if a larger number of subjects had been tested. However, the significant correlation between the results of the TFS1 test and the benefit from TFS information in speech for the hearing-impaired subjects suggests that the TFS1 test could become a useful tool for predicting the ability of hearing-impaired subjects to benefit from TFS information in speech.

It was hoped that the benefit of adding TFS information in speech in a given spectral region would be predicted by the sensitivity of a subject to TFS measured using the TFS1 test at the same centre frequency. However, the present results were too variable to test this hypothesis: there was no significant benefit of adding TFS information to a narrow spectral region either for the hearing-impaired subjects as a group, or for most of the individual subjects. It is not clear whether this lack of significant benefit reflected an inability to use TFS information in a narrow spectral region, or whether significant benefits would have been measured had the SRTs for each condition been determined more accurately. Also, [Moore and Sek (2008a)] showed that untrained normal-hearing subjects could only consistently perform the TFS1 test for centre frequencies from 1100-4400 Hz, if the bandpass filter was centred on the 11th harmonic. Consequently, sensitivity to TFS information could only be tested for centre frequencies corresponding to two of the five spectral regions that were tested. The TFS1 test cannot be used to assess sensitivity to TFS information for frequencies below 1000 Hz, where TFS information is expected to be important in coding F0 information.
5.7 Conclusions

- For normal-hearing subjects, the largest benefit (reduction in SRT) of adding TFS information to a vocoded target and competing-talker signal was obtained when TFS information was added to channels with centre frequencies below 1000 Hz. This is consistent with the idea that TFS is important for coding F0 information.

- Some benefit from adding TFS information was measured when TFS was added to the high-frequency channels only. The same benefit was not measured when TFS was added to the same channels but there was already TFS in the low-frequency channels, suggesting that there is some redundancy in TFS information across frequencies.

- For normal-hearing subjects similar improvements in SRT were measured when TFS information was added to each of the five isolated spectral regions that were tested (spanning 100-8000 Hz). This suggested that TFS information was not important only for coding F0 information.

- Hearing-impaired subjects did not show a significantly greater benefit when TFS information was added to an isolated spectral region than when envelope information alone was added. The benefit from adding TFS information to the whole frequency spectrum was correlated with a psychophysical measure of TFS sensitivity.
Chapter 6

Processing speech to remove temporal envelope information

6.1 Introduction

The normal human auditory system is remarkable in its ability to extract phonetic information from speech, even when the speech signal is very degraded by noise. It is not yet clear how this information is extracted, but a number of possibilities have been suggested.

The cochlea acts as an array of bandpass filters, filtering sound into its component frequencies (Helmholtz 1863). Information about the frequency content of the signal could be carried by the pattern of excitation along the basilar membrane. This information is often referred to as ‘spectral’ or ‘place’ information. It seems likely that this information is important for speech understanding, as phonetic features can often be characterised by their frequency spectrum. The signal arriving at a particular place along the basilar membrane also carries temporal information in its waveform. Such temporal information can be classified as temporal fine structure (TFS) - rapid oscillations in the waveform at a frequency close to the band centre frequency - and envelope - slower modulations superimposed on this TFS. Changes in the firing rate of auditory nerve fibres could carry information about the signal envelope, and the times between spikes could carry information about the waveform TFS (Young and Sachs 1979), provided that nerve spikes are ‘locked’ to a particular phase of the waveform, as is the case in mammals for frequencies up to at least 4000-5000 Hz (Palmer and Russell 1986).

The relative importance of these spectral and temporal cues can be inferred by manipulating waveforms to degrade one or more categories of information while leaving others intact. For example, many studies have assessed the importance of temporal envelope and spectral information in speech recognition using noise or tone vocoders (Dudley 1993).
For a review, see Stone et al. (2008). For such processing schemes, a broadband signal is filtered into a number of frequency channels and the envelope of each channel signal is extracted. This envelope is used to modulate a noise or tone carrier signal centred on each channel and the modulated carriers from each channel are combined. If the analysis filters are broader than the auditory filters of the subjects, the resulting signal has less available spectral detail than the original, and little or none of the original TFS information. Speech in quiet processed in this way is highly intelligible, even when the analysis filters are broad, meaning that much of the spectral detail is lost, illustrating that speech has much redundant information (Shannon et al. 1995). When listening in background noise, however, much more spectral detail is required for good speech intelligibility (Dorman et al. 1998; Qin and Oxenham, 2003), suggesting that redundant information in speech may be important in making speech perception robust to the effects of background noise.

Spectral smearing has also been used to assess the role of spectral (place) cues, while attempting to keep temporal cues intact. Baer and Moore (1993) used an overlap-add procedure to simulate the ‘smearing’ of spectral information that is associated with cochlear hearing loss. They found that smearing of speech to simulate frequency selectivity that was three or six times broader than normal had little effect on speech identification in quiet, but reduced the intelligibility of speech in background noise.

Vocoder processing with narrow analysis channels can be used to assess the effect of removing TFS information on speech recognition. If the analysis channels are similar in width to normal auditory filters, it is assumed that the spectral information that is available to the central auditory system is similar to that for unprocessed speech. Hopkins et al. (2008, Chapter 3) used this method to investigate the effects of removing TFS on speech intelligibility for normal-hearing and hearing-impaired subjects. They found that SRTs for speech in a competing talker background were much improved when TFS information was present for normal-hearing subjects, but the improvement in performance was much smaller for hearing-impaired subjects. Gnansia et al. (2008) also used this method with normal-hearing subjects, and found that TFS information improved intelligibility in noise, especially when the noise was modulated.

Processing speech to remove temporal envelope information has also been attempted, although this is technically challenging. Speech signals that have been processed to try to remove envelope cues, while leaving TFS information intact, will be referred to as ‘TFS speech’. If it were possible to create a signal that only contained TFS information, this could be an important research tool, as it would allow the contribution of TFS to speech intelligibility to be measured under different conditions. Licklider and Pollack (1948) were the first to attempt such processing. They showed that speech subjected to infinite peak
clipping was intelligible to normal-hearing listeners. Initially, this method may appear to remove envelope information and leave only TFS intact. However, while this method removes temporal envelope information from the broadband signal, envelope information can be re-introduced by filtering by the peripheral auditory system (Ghitza, 2001).

More recently, the Hilbert transform has been used to split a signal into envelope and TFS components (usually after bandpass filtering). The Hilbert transform is used to calculate the ‘analytic signal’ corresponding to a waveform. This analytic signal can be thought of as a vector that rotates at a rate corresponding to the instantaneous frequency of the original signal, and which has a magnitude corresponding to the signal envelope (Hilbert, 1912; Bracewell, 1986). Dividing the original signal by this envelope gives a signal with the original TFS, but with a flat envelope. Smith et al. (2002) used a similar method to produce complex sounds with the envelope of one signal and the TFS of another. Signals were split into channels before TFS and envelope extraction. For speech stimuli, the envelope cues appeared to dominate perception. This contrasted with melody stimuli, made up from a series of harmonic complex tones, where TFS appeared to be dominant. This study was criticised, however, as, although the analysis bands were not as wide as the single band used by Licklider (1948), they were in some cases still much wider than the width of the auditory filter, allowing envelope cues to be recovered in a similar way as for the original peak clipping method (Zeng et al., 2004).

Gilbert and Lorenzi (2006) attempted to quantify how narrow the analysis channels needed to be to prevent the reintroduction of usable envelope cues by filtering in the auditory system. They filtered nonsense vowel-consonant-vowel (VCV) syllables into a number of analysis channels, and then processed the signal in each channel to leave only TFS information. The channels were then combined. The resulting signal was passed through a bank of gammachirp filters designed to have approximately the same frequency selectivity as the normal auditory system (Patterson and Irino, 1998). The output of each filter was processed to extract the envelope and this was used to modulate a sine wave with a frequency equal to the centre frequency of the gammachirp filter. These modulated sine waves were combined and the resulting signal was presented to normal-hearing subjects. If envelope cues were not recovered at the output of the gammachirp filter bank, speech processed in this way should not be intelligible. This signal, with recovered-envelope cues from TFS speech, will be referred to as ‘recovered-envelope speech’. Gilbert and Lorenzi found that such recovered-envelope VCV syllables were intelligible when the number of channels used for TFS processing was small, as predicted by Ghitza (2001). However, when the number of channels was eight or more, identification of recovered-envelope VCV syllables was very poor, suggesting that envelope cues sufficient to allow VCV syllable identification were not recovered. Gilbert and Lorenzi (2006) also measured the intelligi-
bility of the same VCV syllables subject only to TFS processing. They found that the intelligibility of these TFS processed syllables was high, even when the number of analysis channels was equal to 8 or more (and performance for recovered-envelope speech was very poor), suggesting that identification of the VCV syllables did not rely entirely on envelope cues, but could be performed mainly on the basis of TFS information.

Lorenzi et al. (2006a) measured the intelligibility of VCV syllables that were TFS processed with 16 analysis channels. They found that VCV identification was initially poor, but improved dramatically with training. Final performance was around 90% correct. They concluded that TFS information in VCV syllables was sufficient to allow their identification, and they suggested that TFS information was important for conveying phonetic information.

A possible problem with the TFS processing schemes mentioned so far is that low-level recording noise is amplified as a result of the TFS processing, and this may adversely affect intelligibility. A signal with envelope information removed has a constant amplitude within a channel. TFS speech, within a channel, can also be thought of as speech that has been infinitely compressed - whatever the input amplitude, the output amplitude is always the same. When there is no speech signal in a particular channel, very low-level recording noise is amplified to the same level as the target speech. A modified method of TFS extraction has been developed here to attempt to reduce the deleterious effects of amplification of low-level noise.

A second objective of this chapter was to assess the intelligibility of TFS processed sentences; most previous studies measuring the intelligibility of TFS speech have used nonsense syllables (Gilbert and Lorenzi, 2006; Lorenzi et al., 2006a). Listeners may learn idiosyncratic features of a small set of VCV stimuli, and so measuring the intelligibility of sentences may give more information about the contribution of TFS information to speech perception in the real world.

6.2 Processing

Speech signals were separated into N channels spanning the range 100 to 5000 Hz, using an array of finite impulse response (FIR) bandpass filters. The order of each filter was chosen so that its frequency response was approximately -6 dB at the point where the response intersected with the response of adjacent filters. Channels above 5000 Hz were not included, as it is widely believed that TFS in this frequency region is unavailable even to normal-hearing subjects, due to the breakdown in phase locking (Palmer and Russell, 1986); see, however, Chapter 5.

Channel edges were chosen to be regularly spaced on an ERB_N-number scale (Glasberg, 1986).
and Moore, 1990), and the signals in each channel were time aligned by adding a delay to compensate for the phase shift caused by the initial filtering.

Within each channel, the envelope of the signal was calculated by finding the magnitude of the analytic signal, using the Hilbert transform. The TFS of the channel signal was then derived by dividing the original signal by the Hilbert envelope. After TFS extraction, the channel signals were re-filtered using the original analysis filters, and the outputs of all filters were time aligned and combined. This was necessary because the TFS extraction processing meant that the frequencies in the tail regions of the original filters were brought up to a level that might have been audible. Note that previous processing schemes did not include this step (Smith et al. 2002; Lorenzi et al. 2006a), so TFS information in each channel would not have been limited to frequencies from the passband of that channel and could include large frequency excursions.

Two processing schemes were tested. For the TFS-LNN scheme, to attempt to reduce the effects of amplification of recording noise, ‘low-noise’ noise (LNN) was added to each channel at a level that was 18 dB below the channel root-mean-square (rms) level, before TFS extraction. For the TFS-no-LNN scheme, no LNN was added to the channel signals. LNN is designed to have a peak-to-mean ratio that is much lower than that of Gaussian noise, meaning that it has a much flatter temporal envelope (Hartmann and Pumplin, 1988). LNN was synthesised to have the same long-term average spectrum as the target speech within each analysis channel for each value of N, by using an iterative procedure described in Moore et al. (2004). Firstly, a Gaussian noise with the same long-term average frequency spectrum as the channel signal was produced. The phases of the noise components were then adjusted using an iterative multidimensional minimisation routine to reduce the peak-to-rms level ratio. The noise was amplified, as a result of the TFS processing (envelope removal), to fill gaps in the envelope of the speech signal. Amplification of LNN in these gaps, rather than recording noise was preferable because LNN has a much more predictable structure, with smaller excursions in amplitude and instantaneous frequency. This is illustrated in Figure 6.1, which shows the amplitude and instantaneous frequency of the 6th channel signal from the 12-channel TFS processing scheme, before TFS processing. The bottom panels show the channel signal with LNN added at 18 dB below the channel rms level, and the top panels show the channel signal with no LNN added. There is little difference in the amplitude of the channel signals, except at very low levels, but the instantaneous frequency of the channel signal with no noise added has many more large frequency excursions, especially at points where the amplitude of the channel signal is low. For unprocessed speech, these rapid frequency excursions occur at minima in the envelope of the channel signal and probably have little effect on intelligibility. However, for TFS processed speech, these frequency excursions are amplified to the
same level as the target speech, and so could reduce intelligibility. The large frequency excursions were reduced by filtering the output of each channel with the original analysis filter before the signals from each channel were combined. This filtering restricted the frequency excursions to within the channel bandwidth.

![Figure 6.1: The amplitude and instantaneous frequency of the channel signal from the 6th channel of a 12-channel filter bank with channel edges equally spaced on an equal-ERB scale from 100 - 5000 Hz, without and with the addition of LNN at a level 18 dB below the channel rms level (top and bottom panels, respectively). The sentence used to produce the figure was ‘The little girl was staring’.](image)

6.3 Analysis of stimuli

To investigate the effect of the addition of LNN to the sentences before processing, sentences processed using the TFS-LNN and TFS-no-LNN schemes were analysed using a similar method to that described by Gilbert and Lorenzi (2006), to assess the TFS and envelope cues available to the central auditory system after peripheral filtering. Eighteen sentence lists from the adaptive sentence list (ASL) corpus (MacLeod and Summerfield, 1990) spoken by a British male speaker were used. The gaps between sentences were
reduced to 50-100 ms by hand editing and the lists were processed using each of the two schemes described above (TFS-LNN and TFS-no-LNN) for N=3, 6 and 12.

To assess the effects of peripheral filtering on the information available to the central auditory system, both the processed (TFS-LNN and TFS-no-LNN) and the original signals were passed through an array of bandpass filters that were designed to have similar frequency selectivity to normal peripheral auditory filters (Glasberg and Moore, 1990). Twenty-five 4th-order gammatone filters were used, spaced between 100 and 4600 Hz. Each was 1-ERBN wide. At the output of each filter, the envelope and TFS were decomposed using the Hilbert transform, in the same way as described in Section 6.2. The correlation, renv, between the envelope produced by the original signal and that produced by the processed signal was found for each channel for the TFS-LNN and TFS-no-LNN schemes. Similar correlations were obtained for the TFS, giving rTFS. The mean results for the 18 ASL lists are shown in Figure 6.2. Recovery of envelope information by simulated peripheral filtering decreased as N increased, consistent with the predictions of Ghitza (2001) and the results of Gilbert and Lorenzi (2006). The value of renv was generally higher in the low-frequency channels than in the high-frequency channels, perhaps because auditory filters at low frequencies are somewhat sharper than filters at high frequencies, when plotted on a linear frequency scale. The sharper the filters are, the more envelope information is expected to be recovered (Ghitza, 2001).

The effect of the addition of LNN on renv depended on the centre frequency of the simulated auditory filters. For filters with a low centre frequency, the addition of LNN before processing decreased renv. However, for high-frequency channels, the addition of LNN slightly increased renv. The reason for this difference is unclear. It is possible that differences in the characteristics of the speech and recording noise across frequencies may have led to the differing effects of the addition of LNN for channels with different centre frequencies.

Similarly, rTFS depended on N and the processing scheme that was used (TFS-LNN or TFS-no-LNN). Overall, correlations increased as N increased. rTFS was higher at frequencies that were close to the centre frequencies of the original analysis filters. For all values of N, addition of LNN increased rTFS.

Overall, the values of renv were somewhat higher than those reported by Gilbert and Lorenzi (2006), especially when N=6 and 12 (for which the channels had similar frequency spacing as for the 8- and 16-channel conditions reported by Gilbert and Lorenzi 2006). Gilbert and Lorenzi (2006) reported correlations that were close to zero for most channels when N=8 and 16. Here, for similar channel spacings, correlations were clearly above zero. A possible explanation is that Gilbert and Lorenzi (2006) did not include a second filtering stage following TFS extraction to limit the processed channel signal to the
Figure 6.2: Mean correlations between the envelope (renv, left-hand panels) and TFS (rTFS, right-hand panels) of the TFS-processed and original speech signals at the output of 25 1-ERB_N-wide gammatone filters, for TFS speech processed using the TFS-LNN and TFS-no-LNN schemes, and for three values of N. Error bars show ± one SD of the mean.
frequencies of the source channel. The second filtering stage used here could have increased envelope recovery following simulated peripheral filtering. To test this hypothesis, 10 ASL sentence lists, with the gaps between sentences reduced to 50-100 ms by hand editing, were processed with the TFS-no-LNN scheme (as this scheme was most similar to the processing described by Gilbert and Lorenzi 2006). The same 10 lists were also processed with the same scheme, but without the second filtering stage following TFS extraction. The recovery of envelope cues was then assessed in the same way as for Figure 6.2 and the results of this analysis are presented in Figure 6.3. The second filtering stage had little effect on renv. For some channels, renv was slightly higher when the second filtering stage was included, but for other channels renv was slightly lower. Correlations were still much higher than those reported by Gilbert and Lorenzi (2006), and so the re-filtering of the channel signals following TFS extraction is unlikely to account for the generally higher correlations that we report.

Gilbert and Lorenzi (2006) found the correlation between the logarithm of the envelopes at the output of simulated auditory filters in response to unprocessed and TFS processed speech. So far, we have shown correlations between the envelopes on a linear scale. To test whether this difference accounted for the generally lower correlations reported by Gilbert and Lorenzi (2006), the correlations between the logarithm of the envelopes at the output of simulated auditory filters in response to unprocessed and TFS-processed speech were determined. The same procedure was used as for producing Figure 6.2, except that a logarithmic transformation was applied to the envelope signals before the correlations were calculated. The results are shown in Figure 6.4. The logarithmic transform means that more weight is given to differences between the envelopes at low levels than for the correlations based on linear envelopes. This transformation generally reduced renv for the TFS-no-LNN scheme, and increased renv for the TFS-LNN scheme. Large excursions in instantaneous frequency would reduce envelope correlations, because components rapidly come in and out of the passbands of the simulated auditory filters, causing spurious envelope fluctuations. If adding LNN to the channel reduces spurious excursions in low-level signal portions, is it also expected to reduce spurious envelope fluctuations, and so increase renv. This result suggests that the low values for renv reported by Gilbert and Lorenzi (2006) were partly due to spurious fluctuations in envelope that arose because of distortions introduced by large excursions in instantaneous frequency.

Another possible explanation is that Gilbert and Lorenzi (2006) used gammachirp filters for their simulated auditory filterbank, whereas we used gammatone filters. Gammachirp filters are affected by level - they are broader at higher levels. It is possible that the gammachirp filters used by Gilbert and Lorenzi (2006) were mostly broader than those used here, which may partly account for the generally lower envelope recovery that
Figure 6.3: Mean correlations between the envelopes of unprocessed and TFS-processed speech at the outputs of 25 1-ERB\textsubscript{N}-wide gammatone filters. TFS speech was processed using the TFS-no-LNN scheme, with and without re-filtering after TFS extraction. Error bars show ± one SD of the mean.
Figure 6.4: Mean correlations (renv) between the logarithm of the envelopes of the TFS-processed and original speech signals at the output of 25 1-ERB_N-wide gammatone filters, for TFS speech processed using the TFS-LNN and TFS-no-LNN schemes, and for three values of N. Error bars show ± one SD of the mean.
they measured.

The analyses so far suggest that the addition of LNN before TFS processing reduced the large excursions in instantaneous frequency in the processed signal, especially when the level of the original signal was low. To assess this more quantitatively, the same lists of concatenated ASL sentences were used. These signals were processed using the TFS-LNN and TFS-no-LNN schemes described in Section 6.2 for three values of N (3, 6 or 12). Each channel signal was divided into 10 ms frames. Frames in which the rms level was lower than 15 dB below the overall channel rms level were identified. For these frames, the instantaneous frequency of the channel signal was calculated by using the Hilbert transform to find the analytic signal corresponding to the channel signal, and then finding the derivative of the instantaneous phase of the analytic signal. The standard deviation (SD) of the instantaneous frequency for each channel signal was found and the mean SD across channels was calculated for each list. ‘Grand mean’ SDs (for the 18 ASL lists) for the TFS-LNN and TFS-no-LNN processing schemes are shown in Table 6.1. A repeated-measures ANOVA was performed on the mean SDs across channels for the 18 lists, with factors N (3, 6 or 12) and processing scheme (TFS-LNN and TFS-no-LNN). The effects of N and processing scheme were both highly significant \( \text{F}(2,34) = 22243.1; p < 0.001 \) and \( \text{F}(1,17) = 944.8; p < 0.001 \) respectively and there was a significant interaction between the two factors \( \text{F}(2,34) = 682.4; p < 0.001 \). A Fisher’s least significant difference (LSD) test was used to assess whether there was a significant effect of processing scheme for each value of N. The LSD (5% level) when comparing means with the same value of N was 0.0013; the effect of processing scheme was significant for all values of N. In summary, the results confirm that the addition of LNN did reduce excursions in instantaneous frequency during the low-level portions of the channel signals.

<table>
<thead>
<tr>
<th>N</th>
<th>TFS-no-LNN</th>
<th>TFS-LNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.169</td>
<td>0.138</td>
</tr>
<tr>
<td>6</td>
<td>0.095</td>
<td>0.085</td>
</tr>
<tr>
<td>12</td>
<td>0.064</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Table 6.1: Mean SDs of the instantaneous frequency in N channels in low-level signal portions of concatenated ASL sentences. Values are shown for the TFS-LNN and TFS-no-LNN processing schemes.

The results of these analyses suggest that adding LNN to each channel signal at a low level before processing reduces some of the artefacts introduced by TFS processing. The fidelity of TFS in the processed signal was improved by the addition of LNN, and large excursions in instantaneous frequency were reduced. The correlation between the
envelopes at the output of simulated auditory filters for the original and TFS-processed speech were also increased by the additions of LNN, particularly if the logarithm of the channel envelopes were compared, which meant that more weight was given to differences between the envelopes at low levels than for the correlations based on linear envelopes.

6.4 Experiment one

6.4.1 Rationale

Envelope cues may be recovered by the peripheral auditory system from TFS speech, especially when N is small, as discussed earlier. Here we used a similar approach to that of Gilbert and Lorenzi (2006) to assess the usability of recovered-envelope cues from TFS speech for different values of N, with sentence stimuli, and with the TFS-LNN processing scheme described above. Only the TFS-LNN scheme was used, as the recovered-envelope cues measured in Section 6.3 were generally higher for this scheme than for the TFS-no-LNN scheme.

Firstly, lists of sentences were subjected to TFS-LNN processing. Then, this TFS speech was processed using a tone vocoder with analysis filters that were designed to have similar frequency selectivity to peripheral auditory filters, to produce recovered-envelope speech. If the original TFS processing eliminates recoverable envelope cues, this recovered-envelope speech should be unintelligible. Gilbert and Lorenzi (2006) low-pass filtered the signal envelope at 64 Hz before using it to modulate sine waves at the centre frequency of each channel. This meant that recovered-envelope information at rates higher than 64 Hz was not included. Stone et al. (2008) showed that including higher rate envelope modulations in a noise or tone vocoder improved intelligibility when listening in a background talker. In this experiment we tested whether including envelope modulations above 64 Hz improves the intelligibility of recovered-envelope speech, by measuring the intelligibility of sentences processed with an envelope low-pass filter at 64 Hz (LPF conditions), or no low-pass filter (no-LPF conditions) for three values of N.

6.4.2 Method

Subjects

Six normal-hearing subjects were tested. They had thresholds of 15 dB HL or less at the standard audiometric frequencies, and were less than 25 years old. All were native speakers of British English. None of the subjects had previously heard the test sentences.
Processing

Sentence lists were processed to extract the TFS in N channels (where N was 3, 6 or 12), using the TFS-LNN method described earlier (see Section 6.2). Gaps between sentences were not removed prior to processing, as was the case for the analysis described in Section 6.3. The TFS speech signal was vocoded to assess the usefulness of any remaining envelope information. The vocoder had 26 analysis channels, implemented using an array of FIR filters with edge frequencies spaced at 1-ERB_N intervals from 100 to 5000 Hz, to reflect the frequency selectivity of the normal auditory system. The order of each filter was chosen so that its frequency response was approximately -6 dB at the point that the response intersected with the response of adjacent filters. Following filtering, channels were time aligned. The Hilbert transform was used to find the analytic signal for each channel signal. The envelope was extracted by finding the absolute values of each analytic signal. For the LPF conditions, the envelope was then low-pass filtered using a cut-off frequency of 64 Hz and slope of -24 dB/octave. For the no-LPF conditions, the envelope was not filtered. The envelope of the signal was then used to modulate sine waves with frequencies equal to the centre frequencies of each channel. After modulation, each modulated sine wave was refiltered with the original analysis filters so that the spectra of the modulated sine waves were restricted to the bandwidths of the original channel signals. The modulated sine waves from each channel were then time-aligned and summed. A flow diagram illustrating the processing used for experiment one is shown in Figure 6.5.

![Flow diagram showing the processing scheme used for experiment one.](image-url)
Materials

Signals were generated by a PC soundcard (Lynx One) at a sampling rate of 22050 Hz, passed through a mixing desk (Mackie 1202-VLZ) and presented at a level of 68 dB SPL monaurally via Sennheiser HD580 headphones. Subjects were tested in a double-walled sound-attenuating booth. Microphones were placed in the booth and the control room to allow the experimenter and subject to communicate, although the control room microphone was only routed to the booth headphones during gaps between stimulus presentations. The testing material was taken from the adaptive sentence list (ASL) corpus \cite{MacLeod1990}, and was spoken by a male speaker of British English. Training sentences were from the Bench-Kowal-Bamford (BKB) lists \cite{Bench1979}, which are of a similar style to the ASL lists, and were spoken by the same British English speaker.

Training

Subjects were trained for one hour prior to testing. The training procedure was the same as the testing procedure described below, except that subjects were corrected verbally if they repeated a sentence incorrectly, and allowed to listen to it again.

Testing procedure

Three values of N were tested (N=3, 6, and 12) for the two processing conditions (LPF and no-LPF), making six conditions in total. Subjects were told that they would hear sentences that sounded very distorted, and that they should try to repeat what they heard. They were encouraged to respond even if they were unsure, or only heard part of the sentence. Scoring was by key word. During the testing session, no feedback was given. The order in which the conditions were presented was counterbalanced across subjects. Three sentence lists were used for each condition, making 18 lists in total.

6.4.3 Results and discussion

The number of key words identified correctly was recorded for each list and converted to a percent correct score. The mean scores in percent correct are plotted as a function of N in Figure 6.6. The intelligibility of the recovered-envelope speech decreased as N increased, and there was little difference between the scores for LPF and no-LPF conditions. Scores were converted to rationalised arcsine units (RAU) before statistical analysis \cite{Studebaker1985}. RAU are designed to transform percent correct data so that scores are normally distributed. A two-way, within-subjects ANOVA was performed to assess the effect of N and processing condition (LPF or no-LPF). The effect of N was highly

126
significant \( F(2,10)=240.97; \ p<0.001 \), but there was no significant effect of processing condition \( F(1,5)=0.29; \ p=0.61 \). There was no significant interaction between the two factors \( F(2,10)=0.18; \ p=0.84 \).

Figure 6.6: Percent correct scores for speech processed using the TFS-LNN scheme (\( N=3, \ 6 \) and 12) and then vocoded with a tone vocoder to assess the usability of recovered-envelope cues. Error bars show ± one SD of the mean. Results are shown with the channel envelopes low-pass filtered at 64 Hz (open circles), or not filtered in this way (filled circles).

Performance was good when \( N=3 \), which is consistent with the predictions of Ghitza (2001), who showed by modelling that envelope information could be extracted by the auditory filters from waveforms processed to preserve only TFS information, when the analysis bands over which the TFS was extracted were wide. Ghitza (2001) also predicted that, as the analysis bands from which TFS was extracted were made smaller, the amount of envelope information that could be recovered would decrease. This is also consistent with our data, as intelligibility decreased as \( N \) increased.

The mean percent correct score for \( N=12 \) was 6.7%. It is difficult to assess what portion of this score can be accounted for by subjects guessing keywords, or listening only to the general rhythm of the sentence. The sentences from the ASL corpus have a very uniform, predictable structure. Subjects were asked to respond after each sentence, guessing words that they did not hear. A certain level of performance would be expected just by intelligent guessing, and it seems reasonable to assume that the mean score of 6.7% for \( N=12 \) can be accounted for in this way. For this condition, subjects scored an average of only three words correct per list, and these words were usually predictable (e.g., they’re, man, he, girl).
Gilbert and Lorenzi (2006) used a similar processing strategy to assess the role of recovered-envelope cues in VCV nonsense syllables. They found that subjects performed little better than chance for their 8-channel and 16-channel conditions, which had similar channel spacing to our conditions when N=6 and 12. The present results are consistent with these results when N=12. However, performance when N=6 was clearly better than chance, suggesting that usable envelope cues were recovered. There are two possible reasons for this discrepancy:

- Gilbert and Lorenzi (2006) made no attempt to prevent amplification of recording noise. The addition of LNN to each channel before TFS extraction could have improved intelligibility in the present study by reducing spurious envelope modulations in the recovered-envelope speech.

- Frequency modulation in the TFS speech in each channel was limited to the analysis bandwidth by re-filtering with the analysis filters after TFS extraction. This limit was not imposed by Gilbert and Lorenzi (2006), and could have contributed to the reduced envelope recovery in their study. However, in Section 6.3 we showed that envelope recovery (as measured by the correlation between the envelope at the output of simulated auditory filters in response to original and TFS processed signals) was affected very little by re-filtering after TFS extraction (see Figure 6.3), so this explanation seems unlikely.

6.5 Experiment two

6.5.1 Rationale

Other studies that have used TFS-speech stimuli have not considered the effect of amplification of recording noise in the gaps in the speech signal. Here we compared the intelligibility of TFS processed sentences with or without LNN added before TFS processing, as described in Section 6.2 (referred to as TFS-LNN and TFS-no-LNN processing schemes, respectively). Intelligibility was measured for three values of N for each scheme, making six conditions in total.

6.5.2 Methods

Subjects

Six normal-hearing subjects were tested. Selection criteria were the same as for experiment one. None of the subjects had taken part in experiment one.
Processing

Sentences were TFS processed as described in Section 6.2. N was equal to 3, 6 or 12. Gaps between sentences were not removed prior to processing, as was the case for the analysis described in Section 6.3. This meant that for sentences processed using the TFS-LNN scheme, each sentence was preceded by a short (approximately 500 ms) period of noise. The implications of this are discussed in Section 6.5.4.

Training

Subjects were trained for one hour prior to testing. The same conditions were used in the training session as in the testing session, and the training procedure was the same as that for experiment one. BKB sentences were used in the training session.

Procedure

Firstly, the intelligibility of ASL sentences was measured for the six conditions described above. The order of presentation of the conditions was counterbalanced across subjects, and the procedure for testing was the same as that described for experiment one. Three ASL sentence lists were used for each condition, and they were presented consecutively.

Secondly, the intelligibility of IEEE sentences (Rothauser et al., 1969) was measured for the same subjects and the same conditions, without further training. Subjects were shown a sample IEEE sentence list to allow familiarisation with the sentence structure. IEEE sentence lists are more complex than the ASL and BKB sentence lists. They have five keywords per sentence, and typically are much less intelligible. They were included because intelligibility for the ASL sentence lists for TFS speech processed with three or six analysis bands was very high, possibly leading to ceiling effects that might have hidden differences between conditions. Two IEEE sentence lists were used for each condition, and the order of presentation of the six conditions was counterbalanced across subjects.

6.5.3 Results

The results are shown in Figure 6.7. The intelligibility of TFS speech decreased as N increased. This was the case for both processing schemes and sentence types. Two within-subject ANOVAs were performed to assess the effect of N and processing scheme on intelligibility for ASL and IEEE sentences. The ANOVA was performed on the mean RAU scores for each subject and condition. For the ASL sentences the effect of N and processing scheme was significant \[ F(2,10) = 108.3; \ p < 0.001 \] and \[ F(1,5) = 18.5; \ p = 0.008 \], respectively, but there was no significant interaction between the factors \[ F(2,10) = 2.8; \ p = 0.1 \]. The effects of N and processing scheme were also significant for the IEEE sentence
lists, though in this case the intelligibility of sentences processed with the TFS-LNN scheme was significantly worse than the intelligibility of sentences processed with the TFS-no-LNN scheme $[F(2,10)=232.3; \ p<0.001$ and $F(1,5)=11.3; \ p=0.02$, respectively], and again there was no significant interaction between the factors $[F(2,10)=1.93; \ p=0.20]$. The results for both sentence types could have been affected by floor and ceiling effects. For the ASL lists for $N=3$ and 6, performance was above 90% for both processing schemes, so differences between the two processing schemes could have been hidden by ceiling effects. For the IEEE lists, when $N=12$, correct word identification was around 20% for both schemes. In this case, differences between performance for the two schemes could have been hidden by floor effects.

Figure 6.7: Percent correct scores for TFS speech processed using the TFS-LNN and TFS-no-LNN schemes for $N=3$, 6 and 12. The results for the ASL sentence lists are shown in the right panel, and the results for the IEEE sentence lists are shown in the left panel. Error bars show ± one SD of the mean.

6.5.4 Discussion

Experiment one showed that the intelligibility of recovered-envelope speech dropped as $N$ increased. For the ASL sentence lists, the intelligibility of recovered-envelope speech was low when $N=12$, yet, as shown by experiment two, the intelligibility of TFS speech processed with 12 analysis channels was high for the ASL sentences, suggesting that speech understanding was possible mainly based on TFS cues. However, although the intelligibility of recovered-envelope speech was close to zero when $N=12$, recovered-envelope cues
may still be used in conjunction with TFS information to understand TFS speech.

The intelligibility of TFS speech decreased with increasing N for both processing schemes for the ASL and IEEE sentences, despite the fact that correlations between the TFS of the original speech and the TFS speech at the output of simulated auditory filters increased with N (see Figure 6.2). One reason for this could be the decrease in the recovery of envelope cues as N increased (as illustrated in experiment one). Another possible reason for the reduction in intelligibility as N increased is that, as the number of channels increases, there is more masking between channels. The TFS processing is equivalent to infinite compression within a channel, and so the level of each channel signal does not vary over time. The short term spectrum of speech is often peaky, with speech information in a narrow frequency region, and with other frequency regions dominated by noise. The TFS processing means that all channels have the same level, and so as N increases the masking of speech information by noise in adjacent channels increases. This is illustrated in Figure 6.8, which shows the excitation patterns evoked by filtered white noise with the same frequency spectrum as each channel signal when N= 3, 6 and 12. The excitation patterns were calculated using a model for normal hearing described by Moore et al. (1997).

![Excitation Patterns](image)

Figure 6.8: Excitation patterns evoked by white noise with the same frequency spectra as the channel signals when N=3, 6 and 12. (Moore et al., 1997).

The fact that the intelligibility of ASL sentences was higher for the TFS-LNN scheme than for the TFS-no-LNN scheme is consistent with the hypothesis that amplification of recording noise in channels with no speech signal reduces the intelligibility of TFS speech. The improvement in performance was greatest when N=12, which could reflect a ceiling effect in performance when N=3 and 6. However, the pattern of results appears somewhat different for the IEEE sentences than for the ASL sentences. For the former, performance
was better for sentences processed using the TFS-no-LNN scheme. There are a number of possible reasons for this discrepancy:

- The intelligibility of IEEE sentences was measured after subjects had completed the task for ASL sentences. It is possible that sentences processed using the TFS-no-LNN scheme are more intelligible for highly trained subjects, but not for subjects who have received a smaller amount of training. This explanation seems unlikely, as all of the subjects were trained for one hour prior to testing with the ASL sentences, which is a relatively long training period after which performance would be expected to be stable (Lorenzi et al., 2006a). The effect of training on performance for speech processed with the TFS-LNN and TFS-no-LNN schemes was investigated further in experiment three.

- For the condition where the largest difference occurred between processing schemes for the ASL sentences (N=12), performance for the IEEE sentences for both processing schemes was very low, so differences between processing schemes could have been hidden by floor effects.

- The level of recording noise in the IEEE sentence recordings is somewhat higher than for the ASL sentence recordings. This is illustrated in Figure 6.9, which shows the distribution of levels in channel signals from the 2nd, 6th and 10th channels, with N=12. A second peak in the level distribution is seen clearly for the IEEE sentences for the middle and high frequency channels, at around -45 dB. This second peak indicates the level of noise in the recording. For the ASL lists the second peak is smaller and at a lower level (around -50 to -60 dB). The higher level of recording noise for the IEEE lists may have had a positive effect on performance in the TFS-no-LNN conditions as large excursions in instantaneous frequency occur mainly in signal portions where the signal level is very low (see Figure 6.1).

- The recordings of both the ASL and IEEE sentences had been digitally processed (by those who had made the recordings) to set low-level frames to digital zero. This meant that, for the sentences processed with the TFS-no-LNN scheme, there was an additional cue that indicated where the sentence started. Between sentences, each channel signal had an amplitude of zero. The TFS was extracted by dividing the channel signal by the Hilbert envelope (which has non-zero values for arrays containing any non-zero elements), and so the amplitude of each channel signal processed with the TFS-no-LNN scheme was zero for portions between sentences. This was not the case for the TFS-LNN scheme; for this scheme LNN was added before TFS processing, so the channel signals were not at zero between sentences. The sentence onset cue might have been more important for the IEEE sentences than
the ASL sentences, because the IEEE sentences are generally much less predictable. The first word in the ASL sentences is often not a keyword (the first word is 'the' for over 70% of sentences), and if it is a keyword, it is often predictable (e.g. he, she, they). In contrast, for the IEEE sentences, the first word is often a keyword which is unpredictable (e.g. see, draw, hats, screen). This could account for the finding that performance was worse for the TFS-LNN scheme than for the TFS-no-LNN scheme when IEEE sentences were used. This idea could be tested by gating the sentences processed with the TFS-LNN scheme so that they have the same sentence onset cue as the sentences processed with the TFS-no-LNN scheme.

Figure 6.9: Distribution of levels for channel signals derived from ASL and IEEE sentence lists. The percentage of the total frames per dB is plotted relative to the channel rms level. Frames were either 10-ms long (dashed lines) or 125-ms long (solid lines). Four ASL and four IEEE sentence lists were used to create the figure, with the gaps between sentences reduced to 50-100 ms by hand editing.

6.6 Experiment three

6.6.1 Rationale

A previous study reported large training effects for TFS speech [Lorenzi et al., 2006a], and listening tests as well as observations from the training phase of experiment two suggested that speech processed using the TFS-LNN scheme may require less training to reach a stable level of performance than speech processed with the TFS-no-LNN scheme. Here,
the effect of training on the intelligibility of speech processed using the TFS-LNN and TFS-no-LNN schemes was tested in initially naïve subjects. Subjects were not given any feedback to lengthen the training phase so that differences between the two schemes could be measured, and allow the results to be compared to those of [Lorenzi et al., 2006a], who also gave their subjects no feedback and observed a large training effect. Only one value of N was tested (N=12), and ASL sentences were used, as this condition and sentence material gave the largest difference in performance between the two processing schemes in experiment two. Also, recovered-envelope cues were expected to be smaller for N=12 than for N=3 or 6.

6.6.2 Method

Subjects

Twelve normal-hearing subjects were tested. Selection criteria were the same as those described previously, and none of the subjects had previous been exposed to TFS speech. Six subjects were tested using speech processed with the TFS-LNN scheme, and six using speech processed with the TFS-no-LNN scheme.

Procedure

Eighteen ASL sentence lists were used. The number of keywords correctly identified for each list was recorded. Subjects were not given any feedback. The equipment and procedure were the same as described for experiment one.

6.6.3 Results and discussion

The results are shown in Figure 6.10. Performance improved markedly for both conditions with run number, but performance was always better for the TFS-LNN scheme than for the TFS-no-LNN scheme. The effects of run number and condition were assessed using a two-factor mixed model ANOVA with a between-subject factor of condition and a within-subject factor of run number. The ANOVA was performed on the RAU scores for each subject. The effects of processing scheme and run number were highly significant [F(1,10)=18.8; p=0.001 and F(17,170)=31.3; p<0.001 respectively]. There was also a significant interaction between the two factors [F(17,170)=2.2; p=0.006].

To assess the rate of improvement in intelligibility with run number for the two processing schemes, the percent correct scores were converted to RAU, and a linear regression analysis was performed separately for runs 1-5, 6-10 and 11-15. The results of these analyses are shown in Table 6.2. The slopes (B) of the regression lines fitted to the data...
Figure 6.10: Percent correct scores for speech processed using the TFS-LNN and TFS-no-LNN schemes for N=12 channels. Scores are plotted as a function of run number.

from runs 1-5 were significantly different from zero for both processing schemes, and the slope was steeper for the TFS-LNN scheme than for the TFS-no-LNN scheme, although this difference was not significant ($t(8)=2.2$, $p=0.06$). For runs 6-10, only the slope of the regression line fitted to the TFS-no-LNN data was significantly different from zero, and neither of the regression lines fitted to the data for runs 11-15 had slopes that were significantly different from zero.

These analyses show that initial performance and the rate of early learning were both greater for the TFS-LNN scheme than for the TFS-no-LNN scheme. The rates of learning for the two schemes became more similar later on, but performance remained superior for the TFS-LNN scheme.

The significant effect of processing scheme is consistent with the results of experiment two, which showed that ASL sentences processed with the TFS-LNN scheme were significantly more intelligible than sentences processed using the TFS-no-LNN scheme. The results also support the expectation that less training is required to reach a stable level of performance for the TFS-LNN scheme: more improvement occurred in the first five runs for the TFS-LNN scheme, and a plateau was reached after fewer runs.
Table 6.2: Results of linear regression analysis for the intelligibility of TFS-processed sentences with TFS-LNN and TFS-no-LNN processing schemes for three sets of five consecutive runs for each scheme. Percent correct scores were transformed to RAUs before the analysis. The slopes (B) for each fitted regression line are shown together with the standard error (s.e.) associated with each slope estimate. A t-test was performed to assess whether each value of B was significantly different from zero. The values of t (and corresponding p values) are shown.

<table>
<thead>
<tr>
<th>Runs</th>
<th>TFS-LNN</th>
<th></th>
<th></th>
<th>TFS-no-LNN</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>s.e.</td>
<td>t</td>
<td>p</td>
<td>B</td>
<td>s.e.</td>
</tr>
<tr>
<td>1-5</td>
<td>12.6</td>
<td>2.7</td>
<td>4.6</td>
<td>0.019</td>
<td>6.4</td>
<td>0.9</td>
</tr>
<tr>
<td>6-10</td>
<td>2.2</td>
<td>1.9</td>
<td>1.1</td>
<td>0.339</td>
<td>2.6</td>
<td>0.6</td>
</tr>
<tr>
<td>11-15</td>
<td>0.6</td>
<td>1.8</td>
<td>0.3</td>
<td>0.761</td>
<td>1.7</td>
<td>1.6</td>
</tr>
</tbody>
</table>

6.7 Discussion

We have presented a processing scheme for the production of TFS speech that aims to remove or substantially reduce envelope information that is useful for sentence intelligibility, and that reduces as far as possible distortions that are introduced as part of the TFS processing. Specifically, we have used LNN, added at a low level to each channel before processing, to reduce spurious fluctuations in amplitude and instantaneous frequency produced during low-level segments by the TFS processing. The results of experiments two and three show that the addition of LNN improves the intelligibility of ASL sentences. LNN has a predictable structure and has fewer wide frequency excursions than typical low-level recording noise. This is illustrated in Figure 6.1, which shows the amplitude and instantaneous frequency of the 6th channel signal from the 12-channel TFS processing scheme before TFS processing. There are large excursions in instantaneous frequency in signal portions where the amplitude of the signal is low. For TFS-processed speech, these frequency excursions are amplified to the same level as the target speech, and this may explain the reduced intelligibility of TFS-processed ASL sentences when no LNN was added before processing compared with when LNN was added.

We found that correlations between the envelopes of unprocessed and TFS speech at the outputs of simulated auditory filters were considerably higher than those reported by Gilbert and Lorenzi (2006). The reason for this is unclear, but it may reflect differences in the speech stimuli that were used. Extra care should be taken when interpreting TFS speech intelligibility data from hearing-impaired subjects. These subjects typically have broader auditory filters than normal, so they would be expected to be poorer at recovering
envelope cues at the outputs of their auditory filters. If recovered-envelope cues contribute to the intelligibility of TFS speech, then this auditory filter broadening would lead to a reduction in the intelligibility of TFS speech, independently of a subject’s ability to use TFS information.

Broadened auditory filters in hearing-impaired subjects also lead to an increase in masking across frequencies. As TFS processing is like infinitive compression within a channel, each channel signal has a constant amplitude. As previously mentioned, this means that the gaps between speech information within a channel are filled with either recording noise, or, if the processing described in the current paper is used, LNN. This noise would be expected to be more detrimental to hearing-impaired subjects than normal-hearing subjects because of their increased susceptibility to masking across frequencies. This could contribute to the reduced ability of hearing-impaired subjects to understand TFS speech, as reported by Lorenzi et al. (2006a).

6.8 Conclusions

Sentences processed to remove envelope information in 3, 6 or 12 analysis channels were highly intelligible, and this high intelligibility could not be fully accounted for by recovery of envelope cues at the outputs of auditory filters, as the intelligibility of recovered-envelope speech was very low when 12 analysis channels were used for the TFS processing. However, measuring the correlation between channel-signal envelopes for unprocessed and TFS-processed speech suggested that some envelope information was recoverable even when 12 analysis channels were used for the TFS processing. The very low correlations reported by Gilbert and Lorenzi (2006) for similar analysis channel spacing could have partly been due to distortions introduced by the amplification of low-level signal portions that had large excursions in instantaneous frequency.

Adding LNN to each channel signal before processing can improve the intelligibility of simple sentences by reducing the amplification of large excursions in instantaneous frequency in the low-level signal portions, although the same result was not found for more complex sentences. The effects of the amplification of low-level signal portions should be considered in future studies.
Chapter 7

Summary and suggestions for further work

7.1 Summary

This thesis has investigated the sensitivity of normal-hearing and hearing-impaired subjects to temporal fine structure (TFS) information, and the role of TFS information in speech perception for the two subject groups. We have presented evidence that TFS information is an important acoustic cue for normal-hearing subjects when listening to speech in background noise (Chapters 3, 4, and 5). We have also shown that hearing-impaired subjects are relatively insensitive to TFS information (Chapter 2) and that they are less able than normal-hearing subjects to benefit from TFS information in speech when listening in a competing talker background (Chapters 3 and 5).

Chapter 2 describes a measure of the ability of normally hearing and hearing-impaired subjects to use TFS information in complex tones. Subjects were required to discriminate a harmonic complex tone from a tone in which all components were shifted upwards by the same amount in Hz, in a 3-alternative forced-choice task. The tones either contained five equal-amplitude components (NON-SHAPED stimuli) or contained many components, but were passed through a fixed bandpass filter to reduce excitation pattern changes (SHAPED stimuli). Components were centred at nominal harmonic numbers (N) of 7, 11, and 18. For the SHAPED stimuli, hearing-impaired subjects performed much more poorly than the normally hearing subjects, with most of the former scoring no better than chance when N = 11 or 18, suggesting that they could not access the TFS information. Performance for the hearing-impaired subjects was significantly improved for the NON-SHAPED stimuli, presumably because they could benefit from spectral cues. It was proposed that normal-hearing subjects could use TFS information provided the spac-
ing between TFS peaks was not too small relative to the envelope period, but subjects with moderate cochlear hearing loss made little use of TFS information for unresolved components.

In Chapter 3, speech reception thresholds (SRTs) were measured with a competing talker background for signals processed to contain variable amounts of TFS information, for normal-hearing and hearing-impaired subjects. Signals (speech and background talker) were bandpass filtered into channels. Channel signals above a cut off channel (CO) were vocoded to remove TFS information, while channel signals up to and including CO were left unprocessed. Signals from all channels were then combined. As a group, hearing-impaired subjects benefitted less than normal-hearing subjects from the additional TFS information that was available as CO increased. The amount of benefit varied between hearing-impaired individuals, with some showing no improvement in SRT and one showing an improvement similar to that for normal-hearing subjects. The reduced ability to take advantage of TFS information in speech may partially explain why subjects with cochlear hearing loss get less benefit from listening in a fluctuating background than normal-hearing subjects.

We suggested that the benefit from TFS information in speech for the normal-hearing subjects measured in Chapter 3 could arise because subjects could use TFS information to identify signal portions when there were temporal dips in the masking signal, and so the short-term signal-to-background ratio (SBR) was high. This idea was tested in Chapter 4 using a similar method to that used in Chapter 3. SRTs were measured with steady and amplitude-modulated noise maskers for signals processed to contain variable amounts of TFS information. Subjects benefitted more from TFS information for the modulated than for the steady masker. For both maskers, addition of TFS information up to 548 Hz improved performance, though the improvement was greater for the modulated masker. The addition of TFS information at higher frequencies improved performance further for the modulated masker only. The results were consistent with the idea that TFS information is important for listening in the dips of a fluctuating masker.

In Chapter 5, the experiments described in Chapter 3 were extended to more accurately identify the frequency region where TFS information is most important for normal-hearing and hearing-impaired subjects when listening to speech in a competing talker background. In the first experiment, only normal-hearing subjects were tested. Target and background speech signals were combined and filtered into 30 1-ERB\(_N\)-wide channels between 100 and 8000 Hz. Channels were separated into two frequency regions by a cut-off channel (CO). Performance was measured when the high-frequency region was tone vocoded and CO was increased and when the low-frequency region was tone vocoded and CO was decreased. Six values of CO were tested in each case. Consequently, performance was
measured as TFS information was progressively added, starting at either the high or low end of the frequency spectrum. As TFS information was added starting at low frequencies, performance improved, consistent with the results of Chapter 3. There was no significant improvement in performance as TFS was added to channels with centre frequencies above 2000 Hz. As TFS information was added starting at high frequencies, performance also improved, but this improvement was small until TFS information was added to the frequency region around 400-1000 Hz. We concluded that TFS information may be particularly important in this frequency region.

In a second experiment the benefit gained from TFS information in isolated spectral regions was investigated for normal-hearing subjects and hearing-impaired subjects. The procedure was similar to that for the first experiment; target and background speech signals were mixed and filtered into 30 1-ERB\textsubscript{N} wide channels, and these channels were divided into five, 6-ERB\textsubscript{N} wide spectral regions. For each region, SRTs were measured when no information was present (the channel signals were missing), and when unprocessed information was present (with both envelope and TFS information). Other channels were tone vocoded so that only temporal envelope cues were preserved. Conditions where all 30 channels were tone-vocoded or left unprocessed were included for comparison. Additionally, for the hearing-impaired subjects, a psychophysical measure of sensitivity to TFS was used at two centre frequencies [Moore and Sek 2008a]. Normal-hearing subjects showed a comparable benefit from TFS information across the five spectral regions tested, suggesting that TFS information is important over a wide frequency range. As found in Chapter 3, hearing-impaired subjects benefitted much less than normal-hearing subjects from TFS information, and the benefit was variable across subjects. Subjects who scored poorly for the psychophysical measure of TFS sensitivity showed very little benefit from TFS information in speech. Subjects who scored well for the psychophysical measure of TFS sensitivity generally showed better performance, although this was variable, suggesting that other factors may be important in allowing subjects to benefit from TFS information.

Finally, a method for processing speech to remove (as far as possible) temporal envelope information while leaving TFS information nearly intact (TFS processing) was investigated in Chapter 6. Previous methods for TFS processing have not considered the distortions introduced by the amplification of low-level signal portions as a result of the processing. We investigated the effects of adding low-noise noise to the speech signal before TFS processing. The addition of low-noise noise reduced large excursions in instantaneous frequency, and improved the intelligibility of simple sentences. The opposite effect was seen when the same procedure was used for more complex sentences. The reason for this was not clear, and a number of possibilities were discussed.
7.2 Suggestions for further work

We have shown that hearing-impaired subjects have a reduced ability to use TFS information, both in complex tones and in speech. The reason for this reduced ability remains unclear, however, and a number of possibilities have been proposed (for a more detailed discussion, see Chapter 1). Hearing-impaired subjects typically have poorer frequency selectivity than normal-hearing subjects (Glasberg and Moore, 1986). The TFS at the output of a particular auditory filter in response to a broadband stimulus would therefore be more complex than for normal-hearing subjects with narrower filters, as a particular filter would respond to a wider range of frequency components. Such complex TFS could be un-interpretable by the central auditory system; the reduced ability to use TFS in hearing-impaired subjects could be a direct result of their reduced frequency selectivity. A change in frequency selectivity could also result in a change in the phase response properties of the basilar membrane. If temporal information is coded by comparing the phase response at two adjacent points on the basilar membrane (Shamma, 1985; de Cheveigné and Pressnitzer, 2006), then this could also lead to a reduced ability to use TFS information. Alternatively, it could be that hearing-impaired subjects have a deficit in phase locking to the stimulus waveform, even for frequencies at which phase locking is believed to be robust in normal-hearing subjects. Such a deficit could arise because of damage to inner hair cells, or to auditory nerve fibres.

Pinpointing the nature of the deficit that results in a reduced ability to use TFS information would be an interesting topic for future research. Auditory filter widths could be measured at the same centre frequencies as discrimination of complex tones based on TFS information using the TFS1 test (Moore and Sek, 2008a). If the reduced ability to use TFS observed in hearing-impaired subjects was a result of reduced frequency selectivity, a correlation between the results of the two measures would be expected.

The same hypothesis could also be tested by measuring the discrimination of complex tones on the basis of their TFS, using similar stimuli to those described in Chapter 2 but varying the number of components in the passband. If the ability to use TFS was reduced in hearing-impaired subjects because the outputs of broadened auditory filters were too complex, performance would be expected to improve when fewer components were included in the passband.

Whatever the reason for the deficit, the reduced ability of hearing-impaired subjects to use TFS information in complex sounds may partly explain some of the problems experienced by such subjects when listening in background noise, even when the target speech is amplified to a level that is audible. The individual differences in ability to benefit from TFS information in speech for people with similar audiometric configurations that we reported in Chapters 3 and 5 could account for the relatively poor correlations between
audiometric thresholds and performance, especially in noisy situations. However, most of the work investigating TFS sensitivity has used hearing-impaired subjects with moderate cochlear hearing loss and with a relatively flat audiometric configuration (Lorenzi et al., 2006a, Chapters 2, 3 and 5). These subjects are a poor representation of the hearing-impaired population, as flat audiometric configurations are relatively uncommon, and such subjects often have atypical aetiologies. It would be interesting to test a large number of subjects with a broad range of audiometric thresholds and aetiologies with the TFS1 test to establish how TFS processing ability is related to the pure-tone audiogram. If TFS processing ability is poorly predicted by the pure-tone audiogram, then a test for TFS sensitivity could become an important clinical tool.

The TFS1 test, developed by Moore and Sek (2008a) and based on the procedure described in Chapter 2, is designed to measure the ability to use TFS relatively quickly in a research setting, and ultimately, in the clinic (see Chapter 5, Section 5.6 for a description of the TFS1 test). The results of the TFS1 test could be used to counsel patients on the benefit that they might expect from a hearing aid; subjects who are unable to use TFS would be expected to have more problems when listening in noisy environments that those who are able to use TFS information, even if audibility was restored. Such patients could be advised to use strategies that improve the SBR of their environment, for example, by using directional microphones or assistive-listening devices. Ultimately the results of the TFS1 test could also be used to influence hearing-aid processing. For example, if a patient was insensitive to TFS in a particular spectral region, a hearing-aid could discard that information, and so reduce the processing power needed. This saving in processing power could be used to produce hearing aids with lower power consumption, so reducing the frequency with which batteries need to be changed, or making it possible to use smaller batteries. However, before such a radical processing scheme could be adopted, much more research would be required to validate the TFS1 test for hearing-impaired subjects and to understand the relationship between the results of the TFS1 test and the ability to use TFS information in speech. Even for subjects who do not benefit at all from TFS in terms of SRT, TFS may be important for sound quality.

A measure of the ability to use TFS might also influence the choice of compression strategy. Many authors have argued that TFS is important for dip listening (see Chapters 1 and 3). Moore (2008) suggested that patients who can use TFS information in speech may benefit most from fast-acting compression, which would make the dips in a fluctuating background (and so the signal portions with a high SBR) audible. Patients with little or no ability to use TFS information may rely more on temporal envelope information in speech, and so may benefit more from slow-acting compression, which better preserves this information at the expense of some low-level information being inaudible. This idea
could be tested using normal-hearing subjects. SRTs would be measured for a target speaker in a competing talker background for speech that is unprocessed, and speech that is tone vocoded using 1-ERB\(_N\)-wide analysis channels to preserve the temporal envelope and spectral information available to the central auditory system as far as possible. The signals would then be processed using either a fast or a slow compression system. The input level of the speech would be chosen so that some portions of the signal are inaudible without compression, and the maximum output of the compressor would be limited to simulate the smaller dynamic range of hearing-impaired subjects. If the ability to use TFS information is important for predicting which compression strategy will best suit an individual, SRTs should be lower for the fast-acting than the slow-acting compression scheme for the unprocessed speech, but lower for the slow-acting than for the fast-acting compression for tone-vocoded speech.

The role of TFS information in speech and pitch perception by normal-hearing subjects is still uncertain, and an interesting area for future research. A key question is whether TFS information is important for phonetic information in speech. So far, it has been demonstrated that TFS information is important for listening to speech in a noisy background, but it is not clear whether TFS information is important for listening to speech in quiet. Normal-hearing subjects can identify both unprocessed speech, and speech that is vocoded to remove TFS information almost perfectly when no background noise is present, and so this does not reveal whether TFS is an important cue for identifying phonetic features. The use of TFS processed speech (see Chapter 6) appeared to be a promising way to investigate the role of TFS cues for speech in quiet. However, this method is problematic and it is not clear that the problem of preventing the recovery of temporal envelope information by the peripheral auditory system can be overcome without also introducing distortions in the TFS.
Bibliography


147


