Production and Perceptual Analysis of Speech Produced in Noise

Doctoral Thesis

Youyi Lu

Supervised by Prof. Martin Cooke

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY AT
THE UNIVERSITY OF SHEFFIELD
DEPARTMENT OF COMPUTER SCIENCE
February 2010

Abstract

When exposed to noise, speakers modify the way they speak, possibly in an effort to maintain intelligible communication. These modifications are collectively referred to as the Lombard effect. The work described in this thesis compares speech production changes induced by noise with various spectral and temporal characteristics, and explores the perceptual consequence of these changes. The thesis consists of a series of experimental studies, which involve the analysis of speech corpora collected under different noise conditions, with and without a communicative task. Intelligibility is also measured and predicted using a computer model.

The first study concerns the acoustic and phonetic consequences of *N*-talker "babble" noise on sentence production for a range of values of *N* from 1 (competing talker) to "infinity" (speech-shaped noise). The effect of noise on speech production increased with *N* and noise level, both of which act to increase the energetic masking effect of the noise. In a background of stationary noise, noise-induced speech was always more intelligible than speech produced in quiet, and the gain in intelligibility increased with *N* and noise level, suggesting that talkers modify their productions to ameliorate energetic masking at the ears of the listener.

The effect of low- and high-pass filtered noise on speech production was also examined to address the issue of whether speakers can compensate for energetic masking by actively shifting their spectral energy to regions least affected by the noise. Little evidence was found that speakers are able to modify their speech production to take advantage of those spectral regions clear of noise.

To evaluate the origin of the increased intelligibility of Lombard speech, the fundamental frequency and spectral tilt of speech produced in quiet were artificially manipulated to match those of speech produced in speech-shaped noise. A perceptual evaluation showed that spectral flattening made a larger contribution to Lombard speech intelligibility, but failed to find an influence of an increase in fundamental frequency. A computational modeling study indicated that durational changes could also play an important role in increasing intelligibility. These findings suggest that speech modifications which reallocate energy in time and frequency to introduce more "glimpses" of clean speech in the presence of noise are able to contribute to speech intelligibility.

An analysis of the effect of noise on speech production requires material recorded while undertaking realistic tasks. The effect of a communication factor was explored using conversational speech collected in the presence of maskers with differing degrees of energetic and informational masking potential. The size of speech production changes was found to scale with the energetic masking potential of background noise, extending the findings with read speech to a communicative task. In addition, relative to the non-communicative task, speakers exploited temporal planning to reduce the amount of overlap with a modulated background noise, an effect which was stronger when the noise contained intelligible speech.

In conclusion, the strategies used by talkers to promote successful speech communication under various noise conditions reported in this thesis could enable spoken output applications such as dialogue systems to adapt to communicational environment.

Acknowledgements

First special and the deepest thanks are to my supervisor, Professor. Martin Cooke, for his kindness, continuous support and encouragement throughout my PhD, especially during the most difficult time of my study. He has been a main source of my inspiration for this work. He has offered me many useful suggestions on my project. His "Glimpsing model" and "Grid corpus" have laid the foundation for many experiments reported in this thesis. His writing abilities also have helped me make this thesis a great deal better. It would be much more difficult for me to have this achievement without his help. It has been truly an invaluable leaning experience working with him.

My sincere thanks go to Dr. Jon Barker for his useful suggestions and advice. I am also very grateful to the anonymous reviewers for their suggestions on the work in Chapter 3, 4 and 5. I would like to thank Professor Hideki Kawahara for providing the MATLAB implementation of STRAIGHT v40 used in Chapter 5. Special thanks also go to the company of Tucker-Davis Technologies (TDT) for much technical advice on the TDT equipment used throughout the recording and listening experiments of this thesis.

I would like to thank Sheffield University and the Computer Science Department who offered me such a good opportunity of pursuing a PhD. Especially, the help from Professor Phil Green who was the Head of the Computer Science Department must be recognized.

I must acknowledge all the participants who have been recruited on the behavioral experiments throughout the thesis. Also, thanks to my parents Zhonghua Lu (陆忠华) and Meifang Wang (王梅芳) for their support and encouragement.

Contents

Chapter 1

Introduction.....1

1.1 Motivation.....1

1.2 Thesis outline.....6

Chapter 2

The effects of noise on speech production and perception.....8

2.1 Introduction.....8

2.2 Human speech production.....8

2.3 Effects of factors other than noise on speech production.....12

2.4 Effect of noise on speech production.....13

2.5 Effect of noise on speech perception.....19

2.5.1 Peripheral and central human auditory system.....19

2.5.2 Auditory masking by noise.....20

2.6 Possible causes of noise-induced speech changes.....23

2.7 Research questions.....25

2.7.1 Effects of background talkers on speech production.....25

2.7.2 Active strategies in Lombard speech.....27

2.7.3 Intelligibility of Lombard speech.....28

2.7.4 Role of communication in Lombard speech.....30

2.7.5 Summary.....32

Chapter 3

Speech production modifications produced by competing talkers, babble and stationary noise.....34

3.1 Introduction.....34

3.2 Speech production in noise.....36

3.2.1 Corpus design.....36

3.2.2 Sentence lists and maskers.....37

3.2.3 Talkers.....38

3.2.4 Procedure.....39

3.2.5 Postprocessing..40
3.3 Acoustic and acoustic-phonetic analyses 41
3.3.1 Utterance-level analysis 41
3.3.2 Phoneme-level analysis 46
3.3.3 Correlation analysis..51
3.3.4 Discussion 53
3.4 Intelligibility of noise-induced speech 54
3.4.1 Motivation 54
3.4.2 Experiment I: Sentences in stationary noise..55
A. Listeners 55
B. Stimuli 56
C. Procedure...56
D. Results...57
3.4.3 Experiment II: Sentences in competing utterances 58
3.4.4 Experiment III: Induced speech in matched and unmatched backgrounds 60
3.4.5 Discussion 61
3.5 Energetic masking analysis of noise-induced speech production 63
3.5.1 Motivation 63
3.5.2 Glimpse measures 64
3.5.3 Results...64
3.6 General discussion 68
3.6.1 Noise-induced speech and energetic masking..68
3.6.2 Basis for the increased intelligibility of noise-induced speech 70
3.6.3 Speech changes produced by a competing talker.......................................71
3.6.4 Task-dependence 73
3.7 Conclusion 74
3.8 Compensation for own-voice attenuation ………………………………………...74

Chapter 4

Speech production modifications produced in the presence of low-pass and high-pass filtered noise 78
4.1 Introduction...78
4.2 Speech corpus collection..80
4.2.1 Speech material and noise backgrounds..80
4.2.2 Talkers...81
4.2.3 Procedure 82
4.2.4 Postprocessing...83
4.3 Acoustic measurements and statistical analysis 84

4.4 Discussion 88
4.5 Conclusion 90

Chapter 5

The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise 91

5.1 Introduction 91
5.2 Intelligibility of manipulated speech 94
5.2.1 Speech stimuli and masker 94
5.2.2 Listeners 98
5.2.3 Procedure 98
5.2.4 Results 99
A. Effect of resynthesis procedure 99
B. Effects of manipulated speech on intelligibility 100
5.2.5 Discussion 103
5.3 Does manipulated speech offer more glimpsing opportunities? 104
5.3.1 Motivation 104
5.3.2 Results 105
5.4 General discussion 108
5.5 Conclusion 113

Chapter 6

The effect of task on noise-induced speech production 114

6.1 Introduction 114
6.2 Speech corpus collection 117
6.2.1 Talkers 117
6.2.2 Tasks and maskers 117
6.2.3 Recording setup 119
6.2.4 Transcription 121
6.3 Speech production analysis 122
6.3.1 Word-level analysis 122
6.3.2 Vowel space analysis 125
6.3.3 Temporal effects 128
A. Foreground-background overlap 128
B. Speaking rate 130
C. Mean pause duration 131
D. Simulated talkers 132
6.4 Discussion 134

6.5 Conclusion 139

Chapter 7

Thesis review, implications and future work..140

7.1 Review of the thesis 140

7.2 Implications..144

7.3 Future work 145

7.4 Final summary…………………………………………………………………… 148

Bibliography 149

List of Figures

1.1 The speech chain ……………………………………………………………………….2

2.1 Schematic view of human speech production …………………………………………….9

2.2 Block diagram of human speech production …………………………………………10

2.3 Spectrum of vowel /i:/ and voiceless fricative /s/ ……………………………………..11

2.4 Organization of thesis chapters ………………………………………………………….33

3.1 Long-term average spectrum for the Grid corpus …………………………………….38

3.2 Changes in acoustic parameter values for noise conditions compared to quiet ………43

3.3 Phoneme-specific acoustic differences in noise and quiet conditions 48

3.4 Formant frequencies and energies for vowels produced in quiet and noise …………..49

3.5 Differences for correlation values of acoustic properties in quiet and noise …………52

3.6 Intelligibility gain of Lombard speech in stationary noise background ………………58

3.7 Intelligibility gain of Lombard speech in competing talker background ……………..60

3.8 Glimpse area and proportion for speech produced in noise and quiet ………………..66

3.9 Relation between increase in glimpses and intelligibility of Lombard speech ……….67

3.10 Excitation patterns and glimpses for a sentence produced in quiet and noise ………68

3.11 Frequency response of the attenuator ………………………………………………….75

3.12 Acoustic modifications produced by noise for the two recording methods ……...…...76

4.1 Spectra of speech-shaped noise, and speech produced in quiet and noise ……………80

4.2 Frequency responses of the low- and high-pass digital filters ………………………..82

4.3 Acoustic parameter values for quiet, full-band and filtered noise conditions ………...85

4.4 Acoustic parameter values for individual talkers in quiet and noise ………………….....86

5.1 F0 and spectrum of a processed sentence, the original and the Lombard one ……….97

5.2 Intelligibility of the re-synthesized and original speech in noise ………………………….100

5.3 Intelligibility gain of speech with acoustic manipulations in noise ………………………..101

5.4 Glimpse area and proportion for speech with acoustic manipulations ………………………105

5.5 Relation between increase in glimpses and intelligibility of manipulated speech …………. 107

5.6 Long-term average spectrum of F0-shifed sentences and the originals ……………………..109

6.1 Acoustic parameter values as a function of task and background …………………………123

6.2 F0 differences across tasks in quiet and noise conditions for individuals …………………125

6.3 Between-category and within-category vowel space dispersion …………………………….127

6.4 Overlap percent as a function of task and background ……………………………………129

6.5 Speaking rate as a function of task and background ………………………………………131

6.6 Mean pause durations ………………………………………………………………………..132

6.7 Pause and contribution length densities in quiet and fluctuating backgrounds ……………..133

List of Tables

2.1 Summary of Lombard speech studies and the types of noise employed ……………………...14

2.2 Summary of tasks used in Lombard studies …………………………………………………..15

3.1 Statistics comparing the acoustic parameter values for quiet and noise conditions …………..44

3.2 Phoneme categories ………………………………………………………………………...46

Chapter 1

Introduction

1.1 Motivation

A common experience in today's mobile-phone-dominated world is finding oneself having a conversation in the presence of various background sound sources. Imagine talking over a mobile phone when walking on a street. You may confront traffic noise when cars are passing by, passengers' speech when walking past a bus stop and music emanating from cafes and bars. All these sounds could interfere with your communication with the talker on the other side of the phone. Yet, without realizing it, you may have been able to carry on a normal conversation with your partner the whole time. This raises the question of how speakers can maintain a normal conversation when interfering noise is present.

Denes and Pinson (1973) illustrated the typical situation of a speaker talking to a listener, as shown in figure 1.1. The speech produced by vocal organs of the speaker reaches not only the ears of the listener but also those of the speaker himself. It was suggested by Denes and Pinson that in the simple speaker-listener situation, there are really two listeners, not one, because speakers not only speak, but also listen to their own voice.

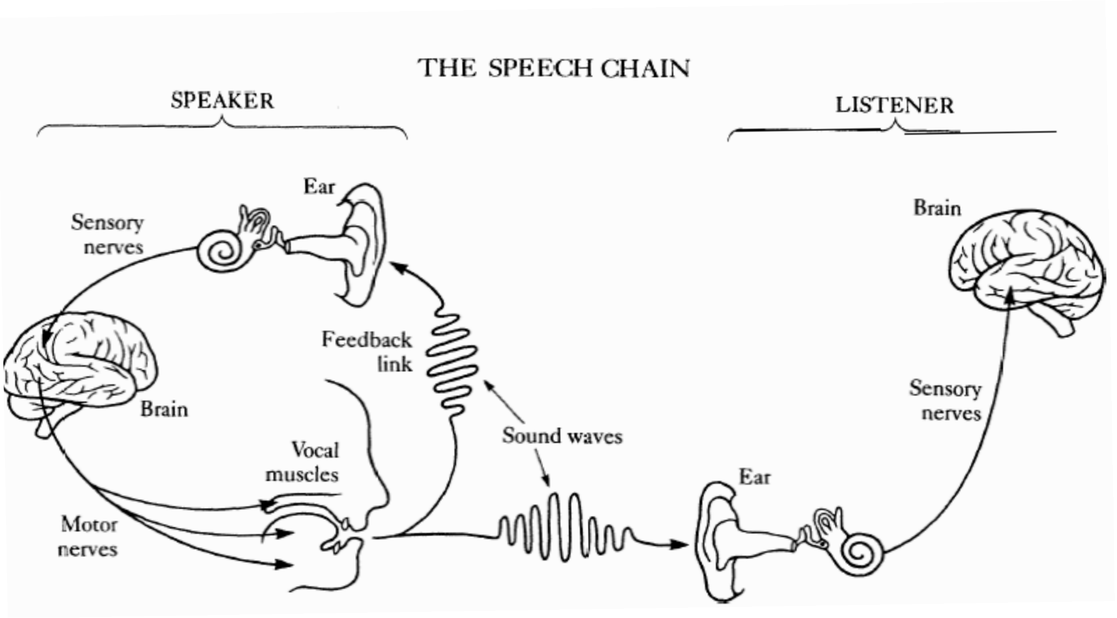


Figure 1.1: *The speech chain (Denes and Pinson, 1973): the progress of a spoken message by the speaker.*

It is notable that the scenario described by Denes and Pinson does not consider possible external factors, in reaction to which changes in neutral speech production could occur. For instance, it has been found that a speaker modifies his vocal output while speaking in the presence of background noise. This phenomenon is called the Lombard effect after the French oto-rhino-laryngologist Etienne Lombard, who first described the impact of noise on speech production (Lombard, 1911). Speech production changes also occur when speakers talk to infants and foreigners who have inadequate experience and knowledge of the language being spoken, known respectively as infant-directed speech (IDS) (Burnham et al. 2000; Kitamura and Lorenzo, 2004) and foreigner-directed speech (FDS) (Knoll and Uther, 2004; Uther et al., 2007). In addition, speech modifications have been observed when auditory feedback is altered in respect of fundamental frequency (F0) (Stuart et al., 2002; Xu et al., 2004), speech level (Fletcher et al., 1918; Howell, 1990) and speaking rate (Hain et al., 2001; Stuart et al., 2002). However, this thesis focuses primarily on speech modifications in response to noise, in the widest sense of the term.

Interfering noise is typically present in many everyday situations e.g. when a speaker is being engaged in a conversation in a crowded party, talking over a mobile phone on a busy street or communicating with a partner in an airport lounge while exposed to a broadcast announcement in the background. Lombard's original treatment (Lombard, 1911) of the effect of noise on speech production has since been extended by a range of studies. Although a variety of experimental conditions have been employed to elicit the Lombard effect across studies, it has been commonly observed that, compared to speech produced in quiet, Lombard speech contains changes in primary acoustic parameters such as increased speech level and F0 and prolonged word duration (Summers et al., 1988; Junqua, 1993; Patel and Schell, 2008). In addition, an increase in the first formant frequency (F1) as well as a flattening of spectral tilt (more spectral energy at higher frequencies) have been reported (Junqua, 1993; Pittman and Wiley, 2001; Varadarajan and Hansen, 2006).

Modifications to normal speech production have a variety of origins. One of the effects of noise on speakers is noise-induced physiological stress. Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity, and deterioration of performance (Gaillard and Wientjes, 1994), which could be provoked by speaking in high noise environments. As suggested by Steeneken and Hansen (1999), among the physiological consequences of stress are respiratory changes, e.g. increased respiration rate, irregular breathing and increased muscle tension of the vocal cords and vocal tract. All of these can lead to alterations to normal speech production.

Noise is also known to yield masking effect on human auditory system. There are two types of masking produced by noise, namely energetic masking (EM) and informational masking (IM). EM results in a reduced audibility of the target sound due to dominance of the noise energy relative to that of the target in certain spectro-temporal regions. One example of IM is the failure to attend to the target in the presence of a competing masker. This could result from the similarity between target and masker (Brungart, 2001; Kidd et al., 2002) or an increased cognitive load due to limited attentional resources (Cooke et al., 2008). EM and IM are described in more detail in section 2.5.2. The presence of these two masking effects might make it difficult for speakers to monitor their own speech via auditory feedback or for listeners to decode the target signals reaching their ears.

Although many studies have demonstrated the Lombard effect, the origin of noise-induced speech modifications is still unclear. Some studies, including the original work of Lombard and others (Lombard, 1911; Pick et al., 1989) suggested that the effect is a kind of reflex rather than a conscious response to a particular noise background. If that is the case, changes in speech production that speakers make in the presence of noise might be caused by, for instance, noise-induced physiological stress. However, there have been studies that suggest otherwise (Junqua, 1993; Hansen, 1996). These studies hypothesize that in a noisy environment, speakers adjust the way they talk in an attempt to maintain intelligible communication. Since the effectiveness of speech communication is likely to be adversely affected by the masking effect of noise, the adjustment to speech production could be a conscious reaction to compensate for the masking effect. Indeed, there is evidence that in the presence of noise, Lombard speech is more intelligible than speech produced in quiet even when speech intensity levels are equalized across the two conditions (Summers et al., 1988; Pittman and Wiley, 2001; Garnier, 2007).

The origin of the intelligibility advantage of Lombard speech is not yet clear.

Pittman and Wiley (2001) suggested that it could be the consequence of combined effect of speech modifications in such as vocal level, spectral slope and word duration. Since it has been found that noise maskers with differing spectral and temporal densities can lead to different amount of EM and IM (Brungart 2001; Simpson and Cooke, 2005), the Lombard effect could be modulated by the differing maskers as a result of speakers' attempts to reduce the masking effect. To test this hypothesis, one goal of this thesis is to analyze the effects of noise with various spectral and temporal characteristics on speech production changes, and to explore how the changes contribute to Lombard speech intelligibility.

In many studies of the Lombard effect, speakers were asked to read prepared texts alone in the presence of noise (e.g. Junqua, 1993; Varadarajan and Hansen, 2006). However, the typical situation of speech communication involves an interlocutor. Summers et al. (1988) and Lane and Tranel (1971) indicated that the incentive to communicate with a speech partner might lead to differences in the Lombard effect compared to while talking alone. Although the Lombard effect when a communication factor is present was studied in e.g. Mixdorff et al. (2007) and Bořil (2008), the differences in noise-induced speech changes between tasks with and without a communicative factor have rarely been examined. A further aim of this thesis is to discover how the effect of noise on speech production is affected by the presence or absence of communicative intent.

Investigating the Lombard effect is not only important for a better understanding of the perception-production link in human speech communication. Behavioral and computational studies of the Lombard effect have technological relevance. Improvements in automatic speech recognition performance under noisy conditions have been reported when Lombard effects have been incorporated into the recognizer

(Hansen, 1994; Chi and Oh, 1996; Hansen, 1996; Bořil, 2008). In speech synthesis, there is an increasing need to cater for changing background conditions. This gives rise to a requirement for spoken output technologies to be able to enhance the intelligibility of synthesized speech in response to dynamic and adverse environments. The current generation of speech synthesis and live speech technologies such as talking in-car GPS is incapable of adapting to the listener's context. The findings of this thesis could contribute to the scientific foundations of adaptive speech synthesis technologies.

1.2 Thesis outline

The remainder of this thesis is arranged as follows. Chapter 2 reviews previous work on the effect of noise on both speech production and perception, and defines the main research questions addressed in the thesis. The joint behavioral-computational studies described in chapters 3 to 6 constitute the original work of the thesis. Chapter 3 investigates the effects of noise with differing numbers of background talkers on speech production, and measures the perceptual consequences. Chapter 3 also attempts to explain the origin of the increased intelligibility of Lombard effect using a computational model based on glimpses of speech which survive the noise. Chapter 4 examines the effect of high and lowpass noise on speech production to explore the question of whether speakers can shift important speech information to those spectral regions least affected by noise. The perceptual influence of changes in spectral tilt and fundamental frequency – two of the most robust effects observed in Lombard speech - is studied in Chapter 5 using artificial Lombard speech. Chapter 6 evaluates the effect of a communication factor on noise-induced speech changes, with a particular focus on the temporal changes in foreground speech produced by different types of noise backgrounds. Chapter 7 concludes with a summary of the main findings and suggestion for future work.

Chapter 2

The effects of noise on speech production and perception

2.1 Introduction

This thesis is concerned with how speech production is affected by different types of noise which yield differential auditory masking effects, and the perceptual consequences of modifications to speech production. The aim of this chapter is to review the factors which are known to influence speech production, focusing on those studies which have investigated the effect of noise on both speech production and perception, and to define the research questions tackled in this thesis. First, the physiological mechanisms for normal speech production are described in section 2.2. Next, some of the alterations to speech production caused by factors other than noise are outlined in section 2.3. The effect of noise on speech production and speech perception is reviewed in sections 2.4 and 2.5 respectively. The potential origin of speech production changes in the presence of noise is discussed in section 2.6. Finally, the research questions to be addressed in this thesis are proposed in section 2.7.

2.2 Human speech production

Figure 2.1 portrays a saggital section of the human speech production system. The description of the speech production system is summarized from Rabiner and Juang (1993). The main components of the system are the lungs, larynx (organ of speech production), pharyngeal cavity (throat), oral cavity (mouth), and nasal cavity (nose). The pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract. The vocal tract begins at the output of the larynx (vocal cords, or glottis) and terminates at the input to the lips, which forms a resonator shaped by various articulators such as tongue, jaw, lips, soft palate and teeth. The nasal tract begins at the velum and ends at the nostrils. When the velum (a trap door-like mechanism at the back of the oral cavity) is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.

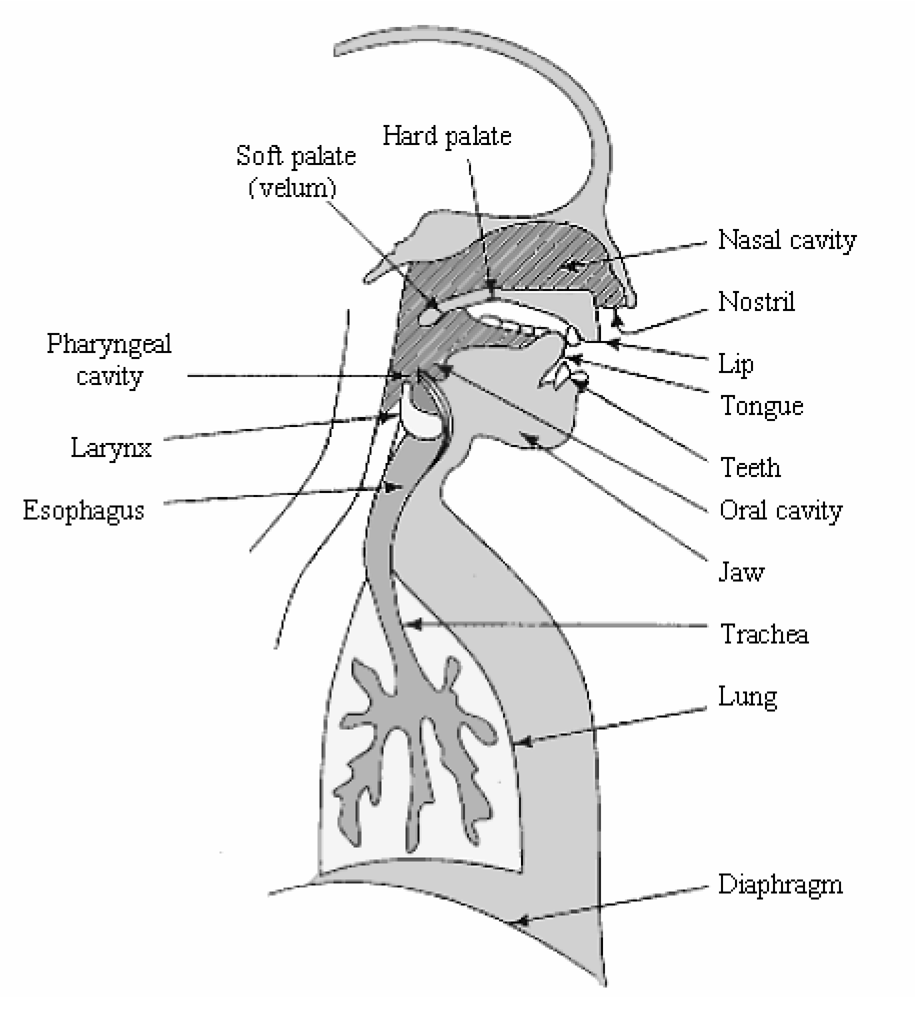


Figure 2.1: *Schematic view of human speech production (Rabiner and Juang, 1993).*

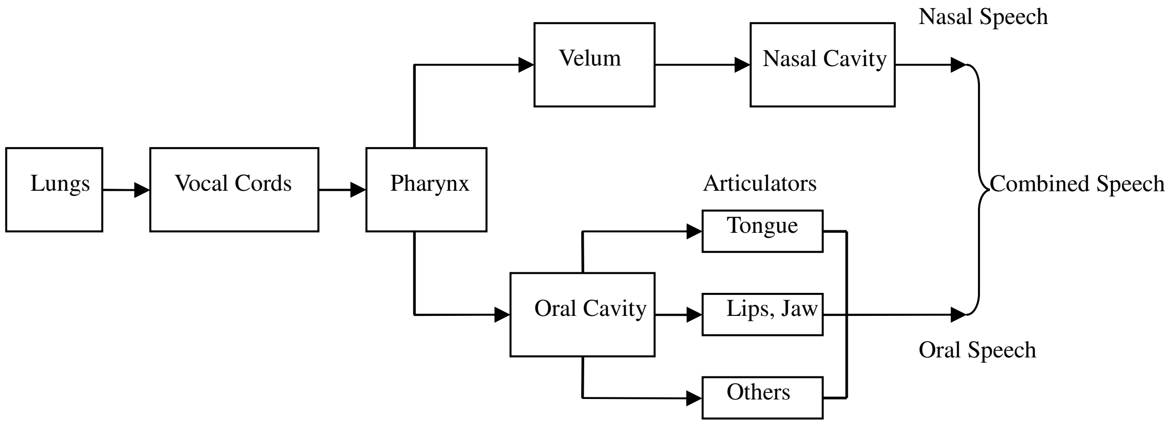


Figure 2.2: *Block diagram of human speech production (after Rabiner and Juang, 1993).*

A simplified representation of the physiological mechanism for creating speech is shown in figure 2.2. Air enters the lungs via the normal breathing mechanism. As air is expelled from the lungs through the trachea, the tensed vocal cords within the larynx are caused to vibrate (i.e. repeatedly fall apart and pull together again), producing quasi-periodic pulses of air flow. The rate of vibration of the vocal folds determines the fundamental frequency (F0) of the speech waveform, which contributes to the perceived pitch of the voice. The spectrum of the source signal originating from the vocal cords is then modulated in passing through the vocal tract, and possibly nasal tract. By using the articulators to change the shape of the vocal tract, hence modifying its resonant characteristics, different speech sounds are produced. In other words, different configurations of the vocal tract enhance some of the harmonics of the fundamental, resulting in formants, and suppress others of the source signal. Speech sounds produced in such a way that the vocal cords are tensed are so-called voiced sounds such as vowels. Alternatively, when the vocal cords are relaxed, in order to produce a sound, the sound source could be created via a constriction in the vocal tract as in the case for unvoiced fricative sounds, or via a sudden and abrupt release of air pressure built up behind a point of the total closure within the vocal tract as for unvoiced plosives. Examples of voiced and voiceless spectra are shown in figure 2.3. The sound produced by the vocal apparatus of the speaker then radiates through the air as a sound wave.

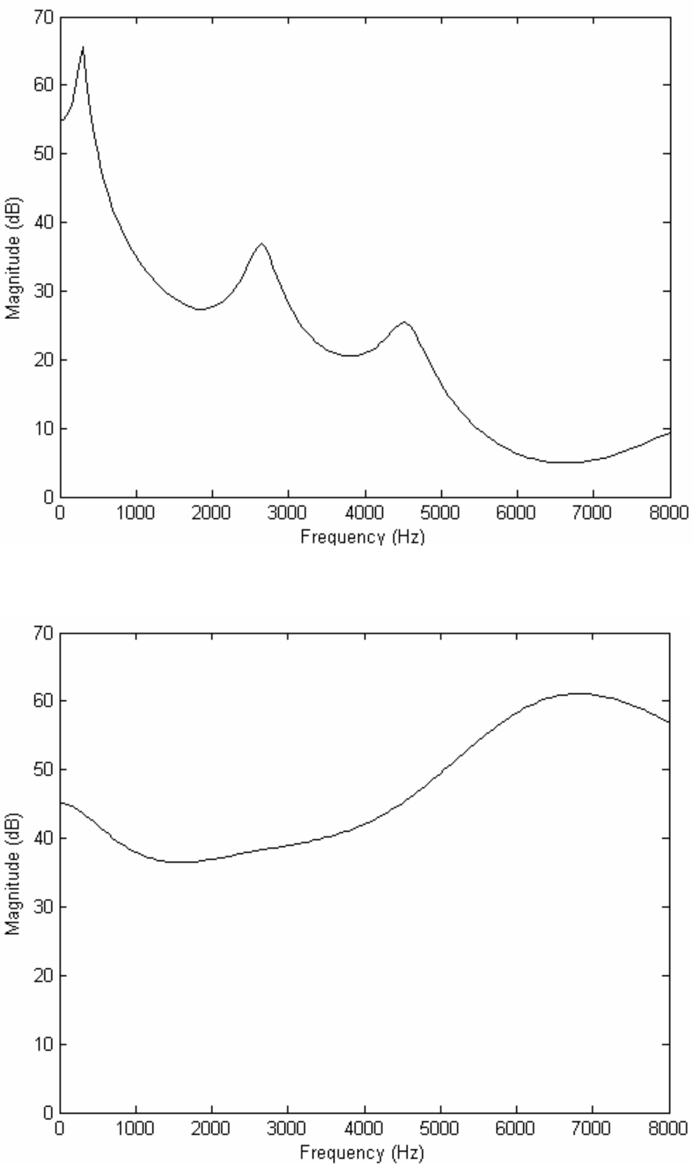


Figure 2.3: *Spectrum of vowel /i:/ (top) and voiceless fricative /s/ (bottom). Formant structure of the vowel is visible.*

2.3 Effects of factors other than noise on speech production

Before focusing on the effect of noise on speech production, this section gives a brief review of speech production modifications induced by non-noise factors. Among others, the effects of clear speech, and infant-directed and foreigner-directed speech are relevant because they frequently result in speech which is more intelligible, and it is of interest to compare their acoustic and phonetic consequences with those which result from speaking in the presence of noise.

When talking to hearing-impaired listeners, speakers tend to adjust their vocal output in order to speak clearly, producing what has been called “clear speech”. Clear speech has been investigated by a number of studies (Chen, 1980; Picheny et al., 1986; Payton et al., 1994; Bond and Moore, 1994; Bradlow et al., 2003; Liu et al., 2004; Krause and Braida, 2004; Smiljanić and Bradlow, 2005), though it has not been studied much outside the hearing impaired field. These studies consistently reported an increase in F0 and a decrease of speaking rate as well as an increase in speech level of clear speech compared to normal speech. A tendency of high-frequency emphasis of spectral energy has also been reported by Payton et al. (1994) and Krause and Braida (2004). Besides these acoustic modifications, a shift of energy from consonant to vowel and an expansion of vowel formant frequency space (i.e. greater discrimination between vowel categories) have been observed by Chen (1980), Picheny et al. (1986), Bradlow et al. (2003) and Krause and Braida (2004). Clear speech has also been found to have a larger pitch range (Bradlow et al., 2003; Krause and Braida, 2004; Smiljanić and Bradlow, 2005) and an increased depth of temporal amplitude modulation (Krause and Braida, 2004; Liu et al., 2004).

Similarly, the principal acoustic changes relative to normal speech such as an increasing F0 and a slower speaking rate have also been observed in infant-directed speech (Andruski and Kuhl, 1996; Kitamura and Burnham, 1998; Burnham et al. 2000; Trainor and Desjardins 2002; Kitamura and Lorenzo, 2004) and foreigner-directed speech (Knoll and Uther, 2004; Scarborough et al., 2007; Uther et al., 2007). In addition, infant-directed speech and, to a lesser extent, foreigner-directed speech also demonstrate an expansion of vowel space and an exaggerated pitch contour (Knoll and Uther, 2004; Biersack et al., 2005; Smith, 2007; Uther et al., 2007).

In addition to these "listener-oriented" speech production alterations, modifications to normal speech production have also been observed when a talker's auditory feedback is altered in respect of speaking rate (Howell, 1990), speech level (Fletcher et al., 1918) and F0 (Natke and Kalveram, 2001). The study of Howell (1990) showed that there was a raising of the voice level and of F0 and a decrease in speech rate when a feedback delay was introduced to speakers' own voice. Fletcher et al. (1918) found that when voice level is amplified, speakers reduce their voice level and when voice level is reduced, speakers increase their voice level, which is called the Fletcher effect. Natke and Kalveram (2001) also reported a change of fundamental frequency in the opposite direction of the frequency shift in voice feedback.

2.4 Effect of noise on speech production

In everyday situations, speech communication frequently takes place in the presence of ambient noise. In its narrowest sense, "noise" is often taken to mean stationary noise whose properties do not vary with time e.g. white noise, but here the term is used in a more general way to refer to the presence of sound which is not the perceptual target of a listener. This wider definition allows for noise to refer to common backgrounds such as a mixture of other talkers (sometimes called multi-talker babble, cocktail party of cafeteria noise), or even a single talker's voice. Lombard's seminal article (Lombard, 1911) describes how a patient with unilateral deafness was presented with an intense noise with a form of continuous crackling through a telephone receiver supplied by an electromagnetic vibrator. The noise was presented first to the impaired ear, then to the normal ear, while the patient was being engaged in ordinary conversation. In the first case, the patient raised his voice slightly or not at all. However, with his "good ear" subjected to the noise, he immediately increased the vocal effort and fundamental frequency (F0), and reduced the vocal effort and F0 to the former level once the noise stopped.

Noise type	**Studies**
White noise	Dreher and O'Neill (1957), Summers et al. (1988), Junqua (1993), Tartter et al. (1993), Garnier et al. (2006)
Pink noise	Bond et al. (1989), Hansen (1996), Junqua et al. (1998), Varadarajan and Hansen (2006)
Multi-talker babble	Rivers and Rastatter (1985), Pittman and Wiley (2001), Mixdorff et al. (2007), Patel and Schell (2008)
Speech shaped noise	Korn (1954), Lippmann et al. (1987)
Cocktail-party noise	Garnier (2007)
Traffic noise	Letowski et al. (1993)

Table 2.1: *Summary of Lombard speech studies and the types of noise employed.*

Tasks	**Studies**
Reading words/sentences	Charlip and Burk (1969), Pisoni et al. (1985) Junqua (1993), Castellanos et al. (1996), Pittman and Wiley (2001), Garnier et al. (2006)
Conversing	Korn (1954), Gardner (1964)
Completing interactive tasks	Rivers and Rastatter (1985), Mixdorff et al. (2007), Patel and Schell (2008)
Communicating words reproduced by listeners	Webster and Klumpp (1962)

Table 2.2: *Summary of tasks used in Lombard studies.*

Inspired by Lombard's experiments, over the past decades a large number of studies have analyzed the impact of background noise on speech production by asking talkers with normal hearing to speak while listening to noise. The conditions used to induce Lombard speech, as well as the analysis techniques, varied across the studies. The diverse noise types and tasks that have been employed are summarized in tables 2.1 and 2.2 respectively. In spite of methodological variety, the majority of studies have converged on a set of primary acoustic changes seen in Lombard speech relative to speech produced in quiet.

Specifically, Lombard speech demonstrates not only an increase in speech level and F0 but also an increase in word duration (or a decrease of speaking rate) and first formant frequency (F1) as well as a shift of spectral energy to higher frequencies (Hanley and Steer, 1949; Korn 1954; Dreher and O'Neill, 1957; Webster and Klumpp, 1962; Charlip and Burk, 1969; Pisoni et al., 1985; Stanton et al., 1988; Summers et al., 1988; Bond et al., 1989; Howell et al., 1992; Junqua, 1993; Letowski et al., 1993;

Steeneken and Hansen, 1999; Pittman and Wiley, 2001; Garnier et al., 2006; Varadarajan and Hansen, 2006; Garnier, 2007; Mixdorff et al., 2007; Bořil, 2008; Patel and Schell, 2008). The prolongation of word length has been found to result primarily from an increase in vowel duration (Stanton et al., 1988; Junqua and Anglade, 1990; Junqua, 1993; Garnier et al., 2006). The findings for consonant duration vary across studies. Stanton et al. (1988) and Junqua and Anglade (1990) found a slight decrease in consonant duration while Junqua (1993) and Garnier et al. (2006) failed to find a significant durational change for consonants. However, Junqua et al. (1999) and Bořil (2008) observed a slight increase in consonant durations. Increases in F0 and F1 frequency in Lombard speech have been found to be physiologically related to the raised vocal effort. During Lombard speech production, the raising of subglottal pressure and the increase of tension in the laryngeal musculature needed to create a louder voice contribute to an increase in F0 (Schulman, 1985; Gramming et al., 1988). Likewise, in order to increase speech level, the wider mouth opening, accompanied by lowering the jaw and the tongue, induces an increase in F1 (Lindblom and Sundberg, 1971; Stevens, 2000).

Changes in noise-induced speech relative to normal speech have also been observed in other acoustic parameters, although there is a notable lack of consistency across studies. For instance, although F1 frequency tends to increase in Lombard speech, no consensus has been reached on the change in F1 bandwidth (Hansen and Bria, 1990; Junqua, 1993; Bořil, 2008). Second formant frequency (F2) has been observed to increase (Hansen and Bria, 1990; Junqua, 1993; Mixdorff et al., 2006) or decrease (Pisoni et al., 1985). In addition, Junqua (1993) and Womack and Hansen (1996) reported a shift of energy from consonant to vowel while Hansen (1996) observed energy shifts from semivowel to vowel and consonant in Lombard speech.

The size of the acoustic changes observed in Lombard speech is influenced by many factors. Noise level affects changes in word duration, vocal intensity and F0 (Dreher and O'Neill, 1957; Webster and Klumpp, 1962; Gardner, 1964; Summers et al., 1988; Letowski et al., 1993; Tartter et al., 1993; Tufts and Frank, 2003; Mixdorff et al., 2007; Patel and Schell, 2008). Dreher and O'Neill (1957) reported that increasing the level of the masking noise from 70 to 100 dB SPL resulted in a steady increase in word duration from 15 to 31%, and a 6 to 9 dB increase in intensity, over speech produced in quiet. With the rise in noise level from 70 to 90 dB SPL, Letowski et al. (1993) found an increase in F0 of between 10 to 20 Hz. In addition, noise level affects the scale of changes to spectral tilt and formant frequencies (Summers et al., 1988; Letowski et al., 1993; Tartter et al., 1993; Tufts and Frank, 2003; Varadarajan and Hansen, 2006). The spectral tilt of the background noise also influences the Lombard effect. Junqua et al. (1998) reported that duration and fundamental frequency tend to increase with noise spectral tilt.

Other factors that appear to affect the size of the Lombard effect include the role of the word in a sentence, the language spoken and speaker gender. Patel and Schell (2008) observed larger effects of F0 and duration for information-bearing word types. Greater F0 variability was also reported for stressed words compared to non-stressed ones (Rivers and Rastatter, 1985). In addition, the size of the Lombard effect was larger for American English than French (Junqua, 1996). Junqua (1993) reported that the influence of the Lombard effect on vocal effort and F0 was greater for male speakers than for females, though Patel and Schell (2008) failed to find such an effect. Stanton et al. (1988) and Junqua (1993) found significant inter-speaker differences in the range of speech production modifications. Furthermore, the type of task employed to study the Lombard effect is another factor which can influence the size of the noise-induced speech modifications. In the study of Garnier (2007), individual talkers were asked to complete a non-interactive task alone or an interactive task with a speech partner. In both tasks the background was quiet or contained wideband noise. It was found that noise-induced speech modifications such as increases in speech level, F0 and vowel duration as well as an increase in F1 and more spectral energy in higher frequencies were larger in the interactive task.

Over the past decades, the typical way of testing noisy data with automatic speech recognition (ASR) system has simply been to add noise to normal speech (e.g. Gu and Mason, 1989; Mokbel and Chollet, 1991). Accordingly, many of the robust automatic speech recognition (ASR) systems have been developed for “additive” noisy speech. However, the aforementioned studies indicate that the problem of recognizing speech produced in noise is not just a simple matter of detection and recognition of signals additively mixed in noise since speech and noise are not independent due to the Lombard effect, as pointed out by Young et al. (1993). Although the ASR systems developed for artificially constrained conditions have reached high levels of performance (e.g. Boll, 1979; Gu and Mason, 1989; Lockwood and Boudy, 1991), they are easily degraded in the face of real world conditions. It has been found that the Lombard effect corrupts recognition performance considerably even if the noise is suppressed or not present in the speech signal (Hansen, 1996; Bou-Ghazale and Hansen, 2000), and the performance degradation due to the Lombard effect can be significantly stronger than that caused by additive noise (Rajasekaran et al., 1986; Takizawa and Hamada, 1990). In order to propose ASR algorithms more resistant to Lombard effect, a number of studies have attempted to integrate Lombard effects into the recognizer, and as a result improvements in automatic speech recognition performance under noisy conditions have been reported (Hansen, 1994; Suzuki et al., 1994; Lee and Rose, 1996; Bou-Ghazale and Hansen, 2000; Bořil, 2008).

2.5 Effect of noise on speech perception

2.5.1 Peripheral and central human auditory system

The human auditory system is a transducer for sound and a sensory-nervous system to support hearing. The system is frequently described as consisting of two components, the peripheral and the central auditory systems, with the boundary between the two at the level of the auditory nerve. The peripheral system can be broken down into outer, middle and inner ear. Information conveyed by acoustic pressure variations enters the outer ear, traveling through the auditory canal to hit the eardrum and causing it to vibrate. The middle ear converts the lower-pressure eardrum vibrations into higher-pressure fluid vibrations in the inner ear. In addition to the airborne sound entering the ears, the sound waves can alternatively reach the inner ear through the bones of the skull, via bone conduction. The cochlea within the inner ear then transforms the mechanical sound vibrations to electrical nerve impulses, which are transmitted to central auditory system via the auditory nerve. The original sound waves are encoded in both the rate of firing and in the time intervals between nerve impulses as a function of frequency and time, sometimes referred to as an excitation pattern. The central auditory system is responsible for higher level auditory processing. Thus encoded, signals from the cochlea are transmitted via the auditory nerve and through a number of intermediate neural regions such as the cochlea nuclei, superior olivary complex and inferior colliculus to the auditory cortex of the brain, which is responsible with the decoding and interpretation of auditory information.

2.5.2 Auditory masking by noise

Lane and Tranel (1971) suggested that it is the masking effect of noise presented to the human auditory system that led to modifications in normal speech production. One of the findings supporting this claim was observed in Lombard's original study (Lombard, 1911). While patients with unilateral deafness reacted only slightly or not at all when having monaural noise was fed to the impaired ear, but when the noise was fed to the healthy ear, they raised their vocal level to near shouting, presumably because their auditory feedback was masked by noise. In general, masking can be said to occur whenever the reception of a specified set of acoustic signals ("targets") is degraded by the presence of others ("maskers") and the magnitude of the masking is usually measured by the elevation in threshold for detecting the target caused by the presence of the masker (Durlach, 2006). In recent years, there has been an increasing body of evidence to support the theory that auditory masking consists of two separate components that originate at different physiological levels. These physiological levels are not precisely determined, but may roughly be divided into the categories of "peripheral" and "central".

Peripheral masking results from competition between target and masker at the periphery of the auditory system i.e., overlapping excitation patterns in the cochlea or auditory nerve (Durlach et al., 2003a). Audibility of the target signal is degraded by the presence of the masker. Because in general there is a good correspondence between the amount of energy in the masker falling near the target signal frequency and the amount of masking that occurs, peripheral masking is often called ''energetic masking'' (EM) (Watson, 1987).

While it is clear that the EM effect of noise on a target signal can be affected by global signal-to-noise energy ratio (SNR), global SNR is not, on its own, a good predictor of the amount of EM. For example, Festen and Plomp (1990) found that a competing talker creates far less EM at any given SNR than a stationary speech-shaped noise. One recent approach to estimate EM, which is employed in this thesis, is the glimpsing model (Cooke, 2006). In his study, it was found that the degree of EM in vowel-consonant-vowel tokens can be well predicted by the amount of "glimpses" of target signal available at the ears of listener, which was further demonstrated for the intelligibility of spoken letters and individual talkers for sentence material in Barker and Cooke (2007). Glimpses of a target signal are those spectro-temporal regions dominated by the target source in the excitation pattern of target-masker mixture. A more recent study (Li and Loizou, 2007) further suggested that the amount of EM is determined by not only the glimpse amount but also the location of the frequency regions glimpsed since the availability of glimpses in frequency regions containing the first and second formants was found to lead to more masking release than in higher frequencies. In addition to the spectro-temporal regions dominated by the target source, a number of studies e.g. Drullman (1995) argued that the weak target elements below the masker level may also help perceive the target in the presence of the masker.

In contrast, central masking is characterized as the inability to detect a target signal embedded in a context of other sounds at the central auditory system even when the target signal is clearly audible, and is usually equated to non-energetic masking, termed "informational masking" (IM) (Pollack, 1975; Watson et al., 1975; Kidd et al., 2002; Durlach et al., 2003a). It has been found that IM is strongly influenced by target-masker similarity. A similar target and masker will increase the difficulty for listeners to determine which signal or parts of the signal they are supposed to attend to, leading to a greater degree of IM. Kidd et al. (2002) and

Durlach et al. (2003b) reported a release from IM when the similarity of spectral and temporal characteristics between target and masker stimuli was reduced. Brungart (2001) and Cooke et al. (2008) also found a dependency of the degree of IM on the similarity of the target and masker voices. In their studies, the greatest IM was seen when the voices used in the target and masker came from the same talker or the intensity levels of the two voices are near the same. Culling and Darwin (1993) and Bird and Darwin (1998) reported an increasing speech-on-speech recognition performance with increasing pitch difference between target and masker up to about eight semitones (about 60%).

IM also refers to the attentional distraction effect of the masker, as studied by Watson et al. (1975), Neff and Green (1987) and Neff and Callaghan (1988), who observed that the masked thresholds for a tonal signal are greatly elevated when the frequencies and temporal positions of masker components are varied at random from trial to trial. The distracting effect of a masker was also suggested by Simpson and Cooke (2005) in a task of consonant identification in the presence of multi-talker babble noise that contained numerous onsets. In addition, an earlier experiment (Treisman, 1964) demonstrated that the linguistic content of the interfering message can greatly influence speech recognition. In that study, competing speech in the same language and similar in content as the target speech was found to be the most disruptive. The results of Garcia-Lecumberri and Cooke (2006) also showed that maskers with the same language as target speech are more disturbing. Further, the studies of Kidd et al. (1994), Kidd et al. (1998), Freyman et al. (2001) and Arbogast et al. (2002) have found that spatial separation of target and masker plays an important role in reducing stimulus uncertainty and thus leads to a release from IM.

2.6 Possible causes of noise-induced speech changes

It is well-accepted that the changes in human speech production observed in noise contain an element of reflex. Lombard (1911) noted that speakers' changes in voice production due to the presence of noise, the Lombard effect, seemed to be unconscious i.e. reflexive. Bořil (2008) suggested that this idea is supported by the results of several experiments. For example, Pick et al. (1989) found that speakers were unable to follow instructions to maintain constant vocal intensity across alternating periods of quiet and noise. In another experiment in the same study, speakers learned to suppress consciously the effect of noise via visual feedback displaying their vocal intensity. However, when feedback was removed, they tended to lower their overall vocal level in noise, rather than changing their specific response to the noise.

In contrast to this, other studies observed a larger increase in speech level due to the rise of noise level when speakers were communicating (Webster and Klumpp, 1962; Gardner, 1964) compared to just reading texts (Dreher and O'Neill, 1957; Lane et al., 1970), showing that the reaction to noise cannot solely be a reflex, but rather consciously driven by other factors such as the speaker's effort to maintain effective communication in noise. In a more recent study, Garnier (2007) extended these findings to other acoustic parameters such as F0, duration, and F1 frequency. In the communicative task, she also found articulatory and prosodic modifications, which were absent in the conditions without interaction with a speech partner. This also suggests an active contribution in addition to a purely reflexive interpretation of the Lombard effect. Thus, when people are talking to each other in the presence of noise, speech production might be influenced by a speaker's efforts to make the speech intelligible at the ears of their interlocutor, possibly by estimating the masking effect of noise at the listener's ears.

In addition, speakers not only speak but also listen to their own voice. Relevant here are studies which show that speech production modifications occur when there is alteration to a speaker's auditory feedback. Fletcher et al. (1918) and Howell (1990) demonstrated that speakers lowered their voices when their own speech was amplified while they increased voices when their speech feedback was attenuated. The studies of Jones and Munhall (2000), Larson et al. (2000) and Natke and Kalveram (2001) reported that speakers attempted to stabilize their F0 by shifting F0 in the direction opposite to changes in the pitch of voice auditory feedback. Pile et al. (2007) also found individuals compensated for the alteration of vowel formant space during instantaneous feedback by pushing their productions in the direction opposite to that of the perturbation. These studies collectively suggest a possible role of own voice monitoring in the regulation of speech production.

It has been speculated that feedback signals that differ from those expected under normal speaking, in respect of phonetic information as well as linguistic content, could lead speakers to change their vocal production in such a way to repair the discrepancies (Levelt, 1983; Brunett et al., 1998; Hain et al., 2000). Denes and Pinson (1973) claimed that in listening to his or her own voice, a speaker continuously compares the speech produced with that intended and make the adjustments necessary to match the results with their intentions. Levelt (1989) also suggested that speech could be transmitted to the monitoring system in a form of covert speech via an internal loop as well as in a form of overt speech via auditory feedback (external loop). Thus, when ambient noise is present, speech production modifications might also be the consequences of overcoming speakers' difficulty in monitoring their own productions due to both energetic and informational masking effects of the noise.

2.7 Research questions

2.7.1 Effects of background talkers on speech production

One aspect of noise-induced speech production changes which has received little attention is the effect of masking noise with different number of background talkers. As detailed earlier, many studies have used stationary noise such as white noise (e.g. Summers et al., 1988; Howell et al., 1992; Junqua, 1993) or pink noise (e.g. Bond et al., 1989; Tufts and Frank, 2003), though some have employed multi-talker babble (e.g. Pittman and Wiley, 2001; Patel and Schell, 2008). Junqua (1994) discovered that multi-talker babble noise led to a larger vowel duration increase as compared to white-Gaussian noise. Garnier et al. (2006) demonstrated that increases in voice intensity, spectral energy, and word duration were greater in white noise than in cocktail party noise while mean F0 increased more in cocktail party noise than in white noise. However, wideband noise and multi-talker babble did not appear to differentially influence the production of speech (Letowski et al., 1993; Pittman and Wiley, 2001).

Surprisingly, the effect of an independent single competing talker on speech production, which might be expected to cause different types of disruption, has not been investigated in depth although speech produced in the presence of other speech material has been studied in the limited sense of altered auditory feedback (Lee, 1950; Natke and Kalveram, 2001; Howell and Sackin, 2002; Stuart et al., 2002; Xu et al., 2004). In this regard, the study of Webster and Klumpp (1962) is relevant. In their study, talker-listener pairs were seated face to face and communicated word lists in conditions of quiet and ambient noise. When there was one background talker-listener pair, the speech level of the foreground talker increased by up to 9 dB, compared to the condition without the background pair. The speaking rate in words per second decreased slightly when the background pair was present. It was also found that the foreground pair made more communication errors when talking at the same time as the competing pair.

In speech perception studies, it is known that a competing talker generates masking effects which differ in two ways from stationary noise. First, at any given global SNR, speech is a far less effective energetic masker (EM) than stationary noise (Festen and Plomp, 1990; Simpson and Cooke, 2005). This could be due to the fact that, compared to stationary noise, whose amplitude remains constant over time, a speech masker contains temporal amplitude fluctuations, which lead to more unmasked fragments of the target signal. Second, speech-on-speech produces additional informational masking (IM) (Brungart, 2001; Cooke et al., 2008) over and above that caused by purely energetic factors, and indeed, this form of masking is the dominant effect in determining the intelligibility of a speech target masked by a competing utterance (Brungart, 2001). Since speech and noise maskers differ in the degree of EM and IM they produce in speech perception, it is of interest to discover whether they have differing effects on speech production. While the task of speech production in noise differs from speech perception in noise, production might be influenced by perceptual concerns in a number of ways. First, masking noise renders monitoring of a speaker’s own productions more difficult, both energetically via loss of information of potential use in feedback, and informationally, due to competing attention. Second, speakers may be able to predict the masking effect of noise in the communicative environment at the ears of their interlocutor in an attempt to maintain intelligible communication. In both cases, modifications to normal speech production might be expected.

Inspired by these issues, chapter 3 explores speech modifications provoked by competing talkers, babble and stationary noise, which yield differential EM and IM, and measure the intelligibility of those modified speech in the presence of noise.

2.7.2 Active strategies in Lombard speech

Speech production changes in the presence of noise causes, amongst other things, an increase in speech level and fundamental frequency (F0), a flattening of spectral tilt and a tendency for an upward shift of F1 frequency. While the scale of changes in acoustic parameters observed in Lombard speech appears to be related to the relative level of the masker (Summers et al., 1988; Tartter et al., 1993), noise maskers with differing spectral shapes and temporal fluctuations have led to consistent changes in speech level, F0 and spectral tilt (e.g. Junqua, 1993; Hansen, 1996; Garnier et al., 2006). One interpretation of the consistency with which various types of noise provoke speech production modifications is that the spectro-temporal properties of the noise may play little or no role in the Lombard effect. Under this view, speakers cannot, or do not, engage in active strategies which take into account the effect of noise at the ears of listeners.

However, other studies have raised the possibility that Lombard speech has an active component. Junqua et al. (1998) studied the influence of noise spectral tilt on Lombard speech, with a constant masker level of 85 dB SPL. Speech level and F0 increased relative to a quiet background when talkers spoke with noise in the background in all conditions of spectral tilt, supporting the notion of a passive Lombard component. On the other hand, the size of the increase in speech level varied with noise spectral tilt. Mokbel (1992) recorded speech in the presence of white noise which was presented either low- or high-pass filtered or without filtering, at a fixed level. An increase of speech energy in frequency regions where the noise energy was most concentrated was observed, suggesting a dependency of the Lombard effect on the noise frequency distribution. However, Mokbel's study involved only one single speaker and did not report detailed changes in acoustic parameters, so it is difficult to appreciate the precise pattern as well as the reliability of the results, given that significant speaker-dependency of speech produced in noise has been observed (Stanton et al., 1988; Summers et al., 1988; Junqua 1993). However, the studies of Junqua et al. (1998) and Mokbel (1992) raise the intriguing possibility that the Lombard effect may have an active component which depends on the spectral characteristics of the background noise. In other words, talkers might use information gained by listening-while-talking to affect purposeful modifications to their speech, perhaps with the goal of improving intelligibility at the ears of the interlocutor. One of the strategies for maintenance of speech intelligibility in noise would be to place spectral information in frequency regions least affected by the noise. Chapter 4 reports on a study of speech modifications produced in the presence of low-pass and high-pass filtered noise whose noise-free region of the spectrum differs.

2.7.3 Intelligibility of Lombard speech

Acoustic differences between speech produced in quiet and in noise lead to differences in intelligibility. It is clear that by increasing speech level, talkers can improve the intelligibility of their speech in the presence of noise due to an increased signal-to-noise ratio (SNR) (Pittman and Wiley, 2001), as long as they do not raise their voices greatly (i.e. shout). For instance, Pickett (1956) and Junqua (1993) found that the intelligibility of Lombard speech degraded when the speech intensity was increased by as much as 20 dB relative to the normal speech. One explanation was that what is gained in SNR is lost in the reduced sound quality of shouted speech (Pickett, 1956; Rostolland, 1985).

A number of studies have also reported the intelligibility gain of Lombard speech over quiet speech in the presence of noise when their speech intensity levels are equalized. Dreher and O’Neill (1957), Summers et al. (1988), Pittman and Wiley (2001) and Garnier (2007) reported that for the same signal-to-noise ratio (SNR) with isolated words or continuous speech, speech produced in noise is more intelligible than speech produced in quiet across different SNRs by up to 30 percentage points of recognition accuracy. The magnitude of these effects increased as the SNR decreased, i.e. the testing environment became more severe (Summers et al., 1988; Pittman and Wiley 2001). Dreher and O’Neill (1957) and Summers et al. (1988) also found that the Lombard speech with greater acoustic changes tended to result in larger intelligibility gain.

Dreher and O’Neill (1957) suggested that the changes in the spectral and temporal properties of speech which accompany the Lombard effect lead to an improvement in speech intelligibility. Summers et al. (1988) also reported that differences in the acoustic-phonetic structure of utterances produced in noise resulted in consistent increases in intelligibility across SNRs and talkers (although only 2 talkers were used). Pittman and Wiley (2001) attempted to address the issue of how noise-induced speech production changes contribute to the intelligibility advantage of Lombard speech in the presence of noise. It was suggested in their study that the intelligibility gain of Lombard speech is likely to be the result of complex interactions between vocal level, spectral composition and other acoustic characteristics, rather than a simple relation between each of these parameters and recognition. However, there is still no clear idea of the cause of the enhanced intelligibility of Lombard speech. One of the hypotheses is that it could result from the speech production changes in an attempt to reduce the masking effect at the listener's ears. Work reported in chapter 3 employed a computational model of energetic masking to test this idea. Using the same model, chapter 5 further explores the contribution of different types of acoustic modification to the increased intelligibility of Lombard speech.

2.7.4 Role of communication in Lombard speech

In many of those studies who investigated the effect of noise on speech production, stimuli were collected when speakers were reading a list of words/sentences without receiving any feedback about the success or failure of their communication (Dreher and O'Neill, 1957; Pisoni et al., 1985; Summers et al., 1988; Junqua, 1993; Letowski et al., 1993; Pittman and Wiley, 2001; Varadarajan and Hansen, 2006). Consequently, there would be little incentive for the speakers to consciously change their speech even with masking noise present in the headphones. Lane and Tranel (1971) indicated that the speaker does not change speech production to communicate better with himself, but rather with others. Summers et al. (1988) also supposed that much larger changes might have been observed in the acoustic-phonetic properties of the utterances produced in noise if some form of communicative task was undertaken by the talkers.

Efforts have been made to explore the Lombard effect when a communication factor was introduced by asking talker-listener pair to establish a conversation (Korn, 1954; Gardner, 1964), to communicate word/utterance lists (Webster and Klumpp, 1962; Bořil, 2008) or to complete interactive cooperative tasks in an instructor-follower manner (Rivers and Rastatter, 1985; Mixdorff et al., 2007; Patel and Schell, 2008). Those primary noise-induced speech changes observed in tasks with no communication were also reported in these studies, such as increases in F0, speech level and vowel duration as well as a flattening of spectral slope. However, in these studies, the difference of the Lombard effect between tasks with and without a communicative element was not evaluated.

Junqua et al. (1998, 1999) compared the speech produced when speakers were reading a list of phrases with those produced while talking to a voice dialing system. In both tasks the background was quiet or contained wideband noise. It was found that the noise-induced speech modifications in both tasks had the same tendency. Specifically, for both tasks, increases in utterance intensity, vowel F0 and phrase duration were observed in the noise conditions compared to the quiet. In addition, the communication factor led to decreases in utterance intensity and duration while F0 increased whether noise was present or not. However, since the dialing system was trained for neutral speech, in spite of listening to noise, speakers had to produce speech close to "neutral" to communicate efficiently with the system. The results from Junqua et al. (1998, 1999) suggest that speakers consciously modify their speech production when communicating. On the other hand, the scenario is not a good example of typical communication in noise. A task that reflects the real-world human communication situation needs to be employed in order to discover the effect of communication on noise-induced speech changes. Motivated by this, the study reported in chapter 6 evaluates the influence of a communication factor on the Lombard effect by asking pairs of talkers to complete an interactive and cooperative task.

2.7.5 Summary

Figure 2.4 demonstrates how the subsequent chapters are organized to address the research issues described above.

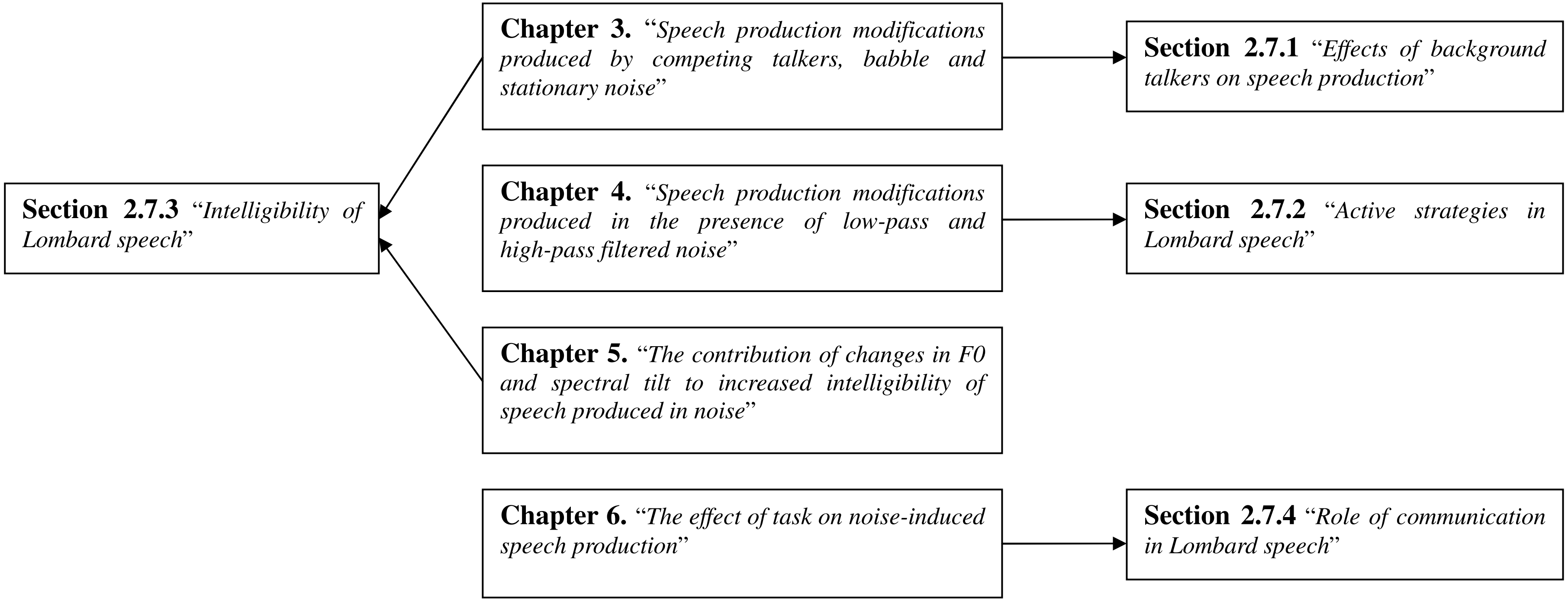

Section 2.7.3 "*Intelligibility of Lombard speech*"
Chapter 3. "*Speech production modifications produced by competing talkers, babble and stationary noise*"
Chapter 4. "*Speech production modifications produced in the presence of low-pass and high-pass filtered noise*"
Chapter 5. "*The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise*"
Chapter 6. "*The effect of task on noise-induced speech production*"
Section 2.7.1 "*Effects of background talkers on speech production*"
Section 2.7.2 "*Active strategies in Lombard speech*"
Section 2.7.4 "*Role of communication in Lombard speech*"

Figure 2.4: *Organization of subsequent chapters which will address the research questions raised in section 2.7.*

Chapter 3

Speech production modifications produced by competing talkers, babble and stationary noise [1]

This chapter examines the acoustic and phonetic consequences of noise with differing numbers of background talkers on speech production and measures the intelligibility of the resulting noise-induced speech in the presence of noise. This chapter also attempts to interpret the origin of the intelligibility gain of Lombard speech using a computational model of energetic masking.

3.1 Introduction

Although the effect of noise on speech production i.e. the Lombard effect has been widely studied, the goal of noise-induced modifications to normal speech production is not yet clear. It has been suggested that by modifying their vocal effort, speakers attempt to maintain a constant level of intelligibility in the face of degradation of the message by the environmental noise source (Summers et al., 1988) and indeed some studies have reported intelligibility gains for "Lombard speech" presented in noise when compared to normal speech in noise (Dreher and O'Neill, 1957; Pittman and

[1] A version of the work reported in this chapter appeared in Lu and Cooke (2008).

Wiley, 2001). However, the issue of how noise-induced speech production changes lead to the intelligibility gain has not yet been addressed.

The primary purpose of the studies described in this chapter was to determine how noise-induced speech production changes are affected by the degree of energetic and informational masking potential of the noise. To measure the effect of differing amounts of EM and IM, *N*-speaker babble noises were employed for a range of values of *N* including *N*=1 (single speaker) and *N*=∞ (speech-shaped noise). While EM increases with increasing *N* (Bronkhorst and Plomp, 1992) and tends to level out at around *N*=16 talkers (Simpson and Cooke, 2005), the influence of IM for sentence material is strongest for small *N* (e.g. *N*=2, Freyman et al., 2004; *N*=3, Carhart et al., 1975) and for *N*=8 for vowel-consonant-vowel tokens (Simpson and Cooke, 2005). Consequently, a number of intermediate values of *N* were also used in this study, and in particular we were interested in the effect of varying *N* on utterance-level properties such as duration, intensity and fundamental frequency as well as formant frequencies, energies and bandwidths and spectral energy distribution at the phonemic level. A further aim was to investigate whether talkers could exploit temporal fluctuations in the noise which are particularly profound for small values of *N*.

The intelligibility of noise-induced speech is known to increase over speech produced in quiet, when noise is added at the same SNRs (Dreher and O'Neill, 1957; Summers et al., 1988). A secondary goal of this chapter was to measure speech intelligibility as a function of the number of talkers and level of background noise. There is still no clear idea of the origin of these intelligibility gains. The current chapter employed a computational model of energetic masking in an attempt to determine whether the acoustic changes produced by noise-induced speech result from an attempt to reduce the energetic masking effect at the listener's ears.

3.2 Speech production in noise

3.2.1 Corpus design

To determine how speech production changes in the presence of widely-differing maskers, a corpus of *N*-talker babble maskers for *N*={1, 2, 4, 8, 16, ∞} was produced. These values were chosen based on an earlier study which measured the masking effect of *N*-talker babble for a large number of values of *N* (Simpson and Cooke, 2005). Since one goal of the study was to investigate the role of informational masking on speech production, talkers produced sentences which were similar in form to those used to produce *N*-babble maskers. The Grid Corpus (Cooke et al., 2006) was used as the source of the masking material, and new speech utterances were collected by asking talkers to read sentences from this corpus. Grid consists of simple 6-word sentences such as “lay green with A4 now” or “set white at B8 again”. Grid has been used in speech-on-speech tasks and shown to produce large amounts of informational masking (Cooke et al., 2008) and the noise-intelligibility relation for speech-shaped noise has been measured (Barker and Cooke, 2007). Maskers for the 6 values of *N* were presented at 89 dB sound pressure level (SPL), a level in the middle of the range known to induce significant speech production changes (Stanton et al., 1988 used 90 dB; Summers et al., 1988 used 80, 90 and 100 dB; Junqua, 1993 used 85 dB). To examine the effect of noise level for the extreme values of *N*, the single speaker (*N*=1) and speech-shaped noise (*N*=∞) maskers were also presented at 82 and 96 dB SPL. Finally, a “quiet” condition was used to provide a reference against which noise-induced speech production modifications could be measured. In summary, talkers produced speech in a total of 11 conditions (6 x *N* values at 89 dB SPL, 2 x *N* values at 82 dB SPL, 2 x *N* values at 96 dB SPL and quiet). Symbols used to represent the 10 noise conditions here and elsewhere are in the form “*N*<number of

talkers>_<level>" so that, for example, "*N*1_89" refers to a competing talker background at 89 dB level, while "*N*inf_96" indicates a speech-shaped noise background at 96 dB level.

3.2.2 Sentence lists and maskers

To allow comparison of acoustic and acoustic-phonetic properties, talkers produced the same set of 50 sentences conforming to the Grid syntax in each of the 11 conditions. However, to introduce some variation, each talker produced a different set of 50 sentences. *N*-babble maskers for the finite values of *N* were generated by adding utterances drawn at random from the Grid corpus into a 60 second circular buffer until the required babble density was obtained. This approach avoids problems with uneven masking effects which would have occurred if utterances had been added with synchronised start times. One consequence of this strategy was that masking sentences were not synchronised with the talker's productions: background utterances would start at a random point in the sentence and there could be a change in talker during the time allotted to the production of a single utterance. Prior to incorporation into the buffer, leading and trailing silence was removed and utterances were scaled to have equal root-mean-square (rms) levels. Masking noise to accompany individual talker productions consisted of 3 s segments of babble drawn at random from the 60 s buffer. Speech-shaped noise was produced by filtering white noise with a filter whose spectrum equalled the long-term spectrum of the Grid corpus, as shown in figure 3.1. Again, a 60 s segment was generated for subsequent random selection.

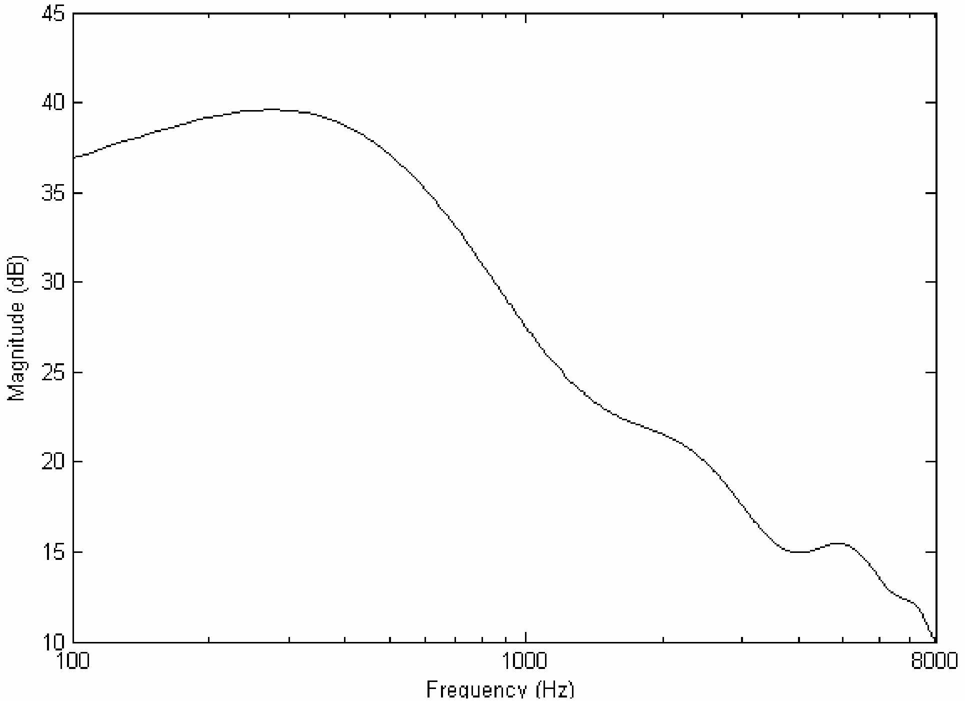


Figure 3.1: *Long-term average speech spectrum for the Grid corpus.*

3.2.3 Talkers

Eight native speakers of British English (4 males and 4 females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. All received a hearing test using a calibrated software audiometer which was used to test each ear separately at the 6 frequencies: 250, 500, 1000, 2000, 4000 and 8000 Hz. One participant had a slight hearing loss (23 dB hearing level) in one ear at the highest frequency (8 kHz) but was retained for the study. The remaining participants had normal hearing (better than 20 dB hearing level in the range of 250-8000 Hz). Ethics permission for the present study and all the other behavioral experiments throughout this thesis was obtained following the University of Sheffield Ethics Procedure.

3.2.4 Procedure

Corpus collection sessions took place in an IAC (Industrial Acoustics Company) single-walled acoustically-isolated booth. Speech material was collected using a Bruel & Kjaer (B & K) type 4190 ½ inch microphone coupled with a preamplifier (B&K type 2669) placed 30 cm in front of the talker. The signal was further processed by a conditioning amplifier (B & K Nexus model 2690) prior to digitisation at 25 kHz with a Tucker-Davis Technologies (TDT) System 3 RP2.1. Simultaneously[2], maskers were presented diotically over Sennheiser HD 250 Linear II headphones using the same TDT system. Speakers wore the headphones throughout, including for the quiet condition. Of course, the use of closed headphones to deliver masking noise can be expected to introduce frequency-dependent own-voice attenuation (Arlinger, 1986; Bořil et al., 2006). Since the current study involves comparison across masking conditions, the constant attenuation characteristics were not considered to be an important factor. However, the closed headphone setup was compared with a compensated transmission channel for a subset of conditions. The chief finding was that the recording method was not a significant factor in the speech production modifications measured (see section 3.8 for details).

Sentence collection and masker presentation was under computer control. Talkers were asked to read out sentences presented on a computer screen and had 3 seconds to produce each sentence and were allowed to repeat the sentence if they felt it necessary. All the repetitions were saved to allow analysis of the number of "false starts" in the different masking conditions. Prior to saving, signals were scaled to produce a

[2] Processing delays in the TDT System 3 processor meant that the noise output was slightly delayed (maximum 6 msec) with respect to speech input.

maximum absolute value of unity to make best use of the amplitude quantisation range. Scale factors were stored to allow the normalisation process to be reversed.

Talkers recorded the 11 conditions over two sessions of 30 minutes each on two days. They were familiarized with the type of sentences and the task before each collecting session. The three single-talker conditions were combined into a single block as were the three speech-shaped noise conditions, and both sentence order and masker level was randomized within the block. Thus, the 11 conditions were presented in seven blocks and block order was randomized for each talker.

3.2.5 Postprocessing

In order to measure acoustic parameters at the level of individual phonemes, a set of speaker-independent phoneme-level hidden Markov models (HMMs) was built from speech material (34, 000 sentences) in the Grid corpus (Cooke et al., 2006) using the HTK HMM toolkit (Young et al., 1999). The speech signals were parameterised into standard 39-dimensional Mel Frequency Cepstral Coefficients (MFCCs), i.e. 12 Mel-cepstral coefficients and the logarithmic frame energy plus the corresponding delta and acceleration coefficients. Each word of Grid sentence was split into phonemes using the British English Example Pronunciation dictionary (BEEP[3]), which resulted in 35 phonemes in total. Each phoneme model was represented using three states each of which had two transitions; a self transition and a transition to the adjacent state. A three state silence model was used to represent the silence period before and after the utterance, and a single state model was used to model optional short pauses between words. The short pause model had a transition between its

[3] Available at ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/

non-emitting start and end states allowing it to be skipped when it was not required. Each state was modelled using a Gaussian mixture model with 25 diagonal covariance components. Training proceeded from a 'flat start' where all model states were initialised with a single Gaussian whose mean and variance were computed across the complete data set. Mixture splitting was employed to increase the number of mixture components. These developed models were used to produce phoneme-level transcriptions of the collected utterances via forced alignment using the HVITE tool in HTK (i.e. to find phoneme boundaries in the speech signal given the known phonetic content). Leading and trailing silent intervals identified via the alignment process were removed. For each talker in each of the 11 conditions, transcriptions of a random selection of 10% of the utterances were manually inspected and found to be accurate.

3.3 Acoustic and acoustic-phonetic analyses

3.3.1 Utterance-level analysis

Eight acoustic properties were estimated for each utterance. Sentence duration, rms energy, mean fundamental frequency (F0) and spectral centre of gravity (CoG) were computed via PRAAT v4.3.24 (Boersma and Weenink, 2005). Mean energy was calculated from the averaged power amplitude of all the samples across time. F0 estimates were provided at 10 msec intervals using an autocorrelation-based method (Boersma, 1993) implemented in the PRAAT program. Mean F0 was obtained by averaging all the valid F0 estimates and expressed in semitones. Spectral centre of gravity was computed on the spectrum of an entire utterance by averaging the frequency spectrum weighted by its power magnitude.

Sentence start time (i.e. the onset of speech production relative to the onset of the interfering signal), and the number and duration of short pauses (> 20 msec) were computed using phoneme-level transcriptions. These latter measures were motivated by the possibility that talkers might avoid overlapping with the background signal, especially in the competing speech conditions. Finally, the voiced-to-unvoiced energy ratio (V/UV ratio) was estimated.

Differences between across-talker means in each background compared to the quiet condition are shown in figure 3.2 for each of the eight acoustic parameters. The number of talkers and noise level in each background is shown as is the baseline mean for the parameter (that is, the mean value in the quiet condition).

To aid the interpretation of figure 3.2, several statistical analyses were carried out for each acoustic parameter. A repeated-measure analysis of variance (ANOVA) analyzed the effect of the number of talkers (*N*) in the maskers at the 89 dB level. To determine any interaction effect between *N* and noise level, a two-way repeated-measures ANOVA with within-subjects factors of *N* (1, ∞) and masker level (82, 89, 96 dB) was computed. Two further single-factor repeated-measures ANOVAs examined the effects of noise level in the single talker and speech-shaped noise condition. Finally, paired-samples *t*-tests with Bonferroni-adjustment were employed to determine the significance of differences between each masking condition and quiet.

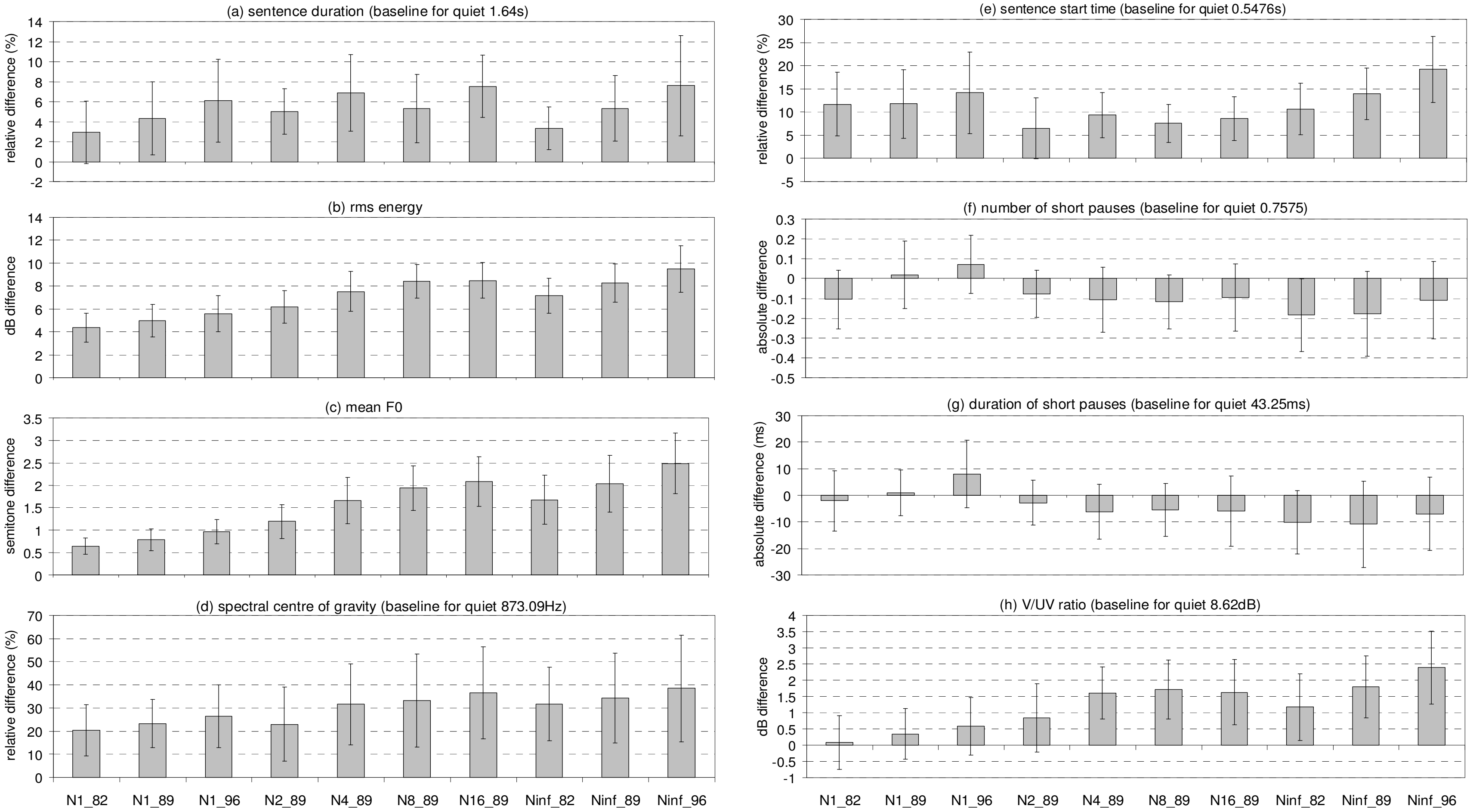


Figure 3.2: *Differences between acoustic parameter values for each noise condition compared to speech produced in quiet. Where meaningful, 'baseline' parameter values in quiet are given in order to provide an absolute reference. Values shown are means over talkers and error bars, here and elsewhere, indicate 95% confidence intervals. Noise conditions are indicated as N<number of talkers>_<level>.*

Table 3.1 summarizes the results of the statistical analysis for each of the utterance-level acoustic measurements. Many parameters demonstrated significant increases in most of the noise backgrounds compared to quiet (final 10 columns of table 3.1). The most significant effects were for energy (which increased by between 3 and 9 dB relative to quiet) and mean F0 (0.6 to 2.5 semitones). Spectral centre of gravity increased from the quiet baseline of 870 Hz by 20-38%. The mean sentence duration in quiet of 1.64 s rose by 2.4-7.6%, while the pause before speaking increased by 6-18% from a baseline of 0.55 s in quiet. The voiced-to-unvoiced energy ratio rose in most conditions, from 8.6 dB in quiet by up to 2.4 dB. No significant overall effect of the duration of short pauses or the number of short pauses was found.

	Repeated-measures ANOVA			**Paired-Samples *t*-test**									
	***N*_89**	***N*1**	***N*inf**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
sentence duration		* ↑	* ↑				** ↑	** ↑	* ↑	** ↑		* ↑	* ↑
rms energy	*** ↑	*** ↑	** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑
mean F0	*** ↑	** ↑	** ↑	** ↑	** ↑	*** ↑	*** ↑	*** ↑	*** ↑	*** ↑	** ↑	*** ↑	*** ↑
CoG	* ↑	* ↑		* ↑	* ↑	* ↑		* ↑	* ↑	* ↑	* ↑	* ↑	* ↑
sentence start time			** ↑	* ↑	* ↑	* ↑		* ↑	* ↑	* ↑	* ↑	** ↑	** ↑
number of short pauses	* ↓	** ↑											
duration of short pauses		* ↑											
V/UV ratio	* ↑		* ↑					** ↑	** ↑	** ↑	* ↑	** ↑	*** ↑

Table 3.1: *Summary of the results of statistical analyses comparing the values of acoustic parameters for speech produced in quiet with speech produced in noise. Column "N_89" represents 6 N-talker conditions (N={1, 2, 4, 8, 16, ∞}) at a noise level of 89 dB. Columns "N1" and "Ninf" represent 3 level conditions (level={82, 89, 96 dB}) for N=1 and N=∞ respectively. The final 10 columns represent the individual noise conditions as follows. (1-3) N=1, levels 82, 89 and 96 dB; (4-7) N={2,4,8,16} at 89 dB; (8-10) N=∞, levels 82, 89 and 96 dB. Symbols "↑" and "↓" represent significant increases or decreases in the parameter in noise over the quiet condition. Significance levels: *** <.001, ** <.01, * <.05.*

For some parameters, the difference between speech produced in quiet and that produced in the presence of noise increased with the number of talkers in the babble (column 2 of table 3.1). The strongest effect of *N* was seen for energy (F(1,7)=68.21, η^2=0.92) and F0 (F(1,7)=37.11, η^2=0.87), with a lesser effect of spectral centre of gravity (F(1,7)=9.38, η^2=0.35) and V/UV ratio (F(1,7)=5.92, η^2=0.42). However, the effect of *N* typically reached a plateau at around *N*=8 talkers. For duration, energy and mean F0, the effect of noise level was similar for the single talker and speech-shaped noise backgrounds (column 3 and 4 of table 3.1). Sentence start time (the delay before the talker started speaking following the onset of the noise) increased with level for the speech-shaped noise condition (F(1,7)=12.26, η^2=0.53), while the number (and to a lesser extent, duration) of short pauses increased with level in the single-competing talker condition (F(1,7)=22.85, η^2=0.74). Similarly, effects of centre of gravity (F(1,7)=6.77, η^2=0.44) and duration of short pauses (F(1,7)=4.19, η^2=0.37) were only found in the competing talker conditions, while V/UV ratio increased in the stationary noise conditions (F(1,7)=9.04, η^2=0.32). No interaction between the effects of *N* and noise level was found for any of the parameters.

Figure 3.2 also indicates the range of talker variation for each parameter and background. While all talkers showed similar scale of changes in energy and F0, significant cross-talker variability is present for the remaining measures.

Compared to quiet, increases in duration, mean energy and F0 in stationary noise with the rise of noise level were also found in previous studies on Lombard speech using words and short sentences (Pisoni et al., 1985; Summers et al., 1988; Bond et al., 1989; Letowski et al., 1993; Steeneken and Hansen, 1999; Pittman and Wiley, 2001; Garnier et al., 2006). The finding of an increased spectral centre of gravity in speech-shaped noise conditions is consistent with the results of Hansen (1988), Pittman and Wiley (2001) and Varadarajan and Hansen (2006), who reported a spectral energy migration to high frequencies for utterances and isolated words spoken in stationary noise. In addition, the V/UV ratio increased in most of the *N*-talker conditions, echoing the findings of Junqua (1993) and Womack and Hansen (1996) for stationary noise. However, a pattern of reduced sentence duration was found by Varadarajan and Hansen (2006), who also reported a decrease in short pause duration, while no such effect was found here. Varadarajan and Hansen suggested that decreases in sentence and short pause duration could be caused by a sense of urgency on the part of the speaker, which occurred due to the constant exposure to the background noise. Here, the fact that noise was presented only when the talker was due to speak might account for the differences.

3.3.2 Phoneme-level analysis

Prior to phoneme-level analysis, all of the post-processed utterances were normalized to have equal rms energy. Individual phonemes of the utterances were segmented via the phoneme-level transcriptions. Phonemes were grouped into the six categories {vowel, diphthong, liquid, fricative, plosive, nasal} as shown in table 3.2. Phonemes which had fewer than 500 instances were not used. On average, around 3000 instances of each phoneme were employed.

Vowel	i:, ɪ, e, u:, æ
Diphthong	eɪ, aɪ, aʊ
Liquid	w, l, r
Fricative	f, s, v, z, ð
Plosive	p, t, k, b, d, g
Nasal	n

Table 3.2: *Phoneme categories.*

Duration was measured for all phoneme instances, while spectral centre of gravity was measured for all apart from the plosives, whose more complex spectro-temporal development precluded a meaningful measurement. Spectral tilt was computed for all the vowels. It is important to note that due to the limited number of contexts present in Grid corpus, the phoneme instances used in this analysis should not be regarded as prototypical. For instance, in Grid, the /æ/ vowel can only be found in the word “at”, most of which were reduced to schwa in this context. As a consequence, a formant analysis (frequencies, energies and bandwidths) was undertaken solely for the vowels /i:/, /ɪ/, /e/ and /u:/ in the words “green”, “bin”, “red” and “soon” respectively. Frequency and energy values were computed as the average of the central 3 frames in each vowel instance. All of the measurements apart from spectral tilt were computed using the PRAAT program v4.3.24 (Boersma and Weenink, 2005). The Burg algorithm (Burg, 1975) implemented in PRAAT was used for the measurement of formant frequencies. For spectral tilt, the spectrum of an entire phoneme instance was divided into 10 energy bands following Stanton et al. (1988). Spectral tilt was estimated as the slope of the best linear fit to the 10 log energy values. Individual talker and overall measurements were computed for each phoneme. Measurements were obtained by averaging the differences between the phoneme instances of the utterances from each of the 10 *N*-talker conditions and the instances in the same position of the same speech sentences from the quiet condition. Individual and overall measurements for all the acoustic properties are expressed as relative percentage differences from quiet, apart from formant frequency and bandwidth, which were expressed as Hertz difference, and energy, which used difference in dB.

Figures 3.3 and 3.4 display the quantitative results of the phoneme-level analysis. To enhance the readability of the plots, results have been averaged across subsets of the 10 noise backgrounds. In general, changes in noise backgrounds over quiet were found, and stronger effects were observed for larger number of background talkers, and for higher noise levels.

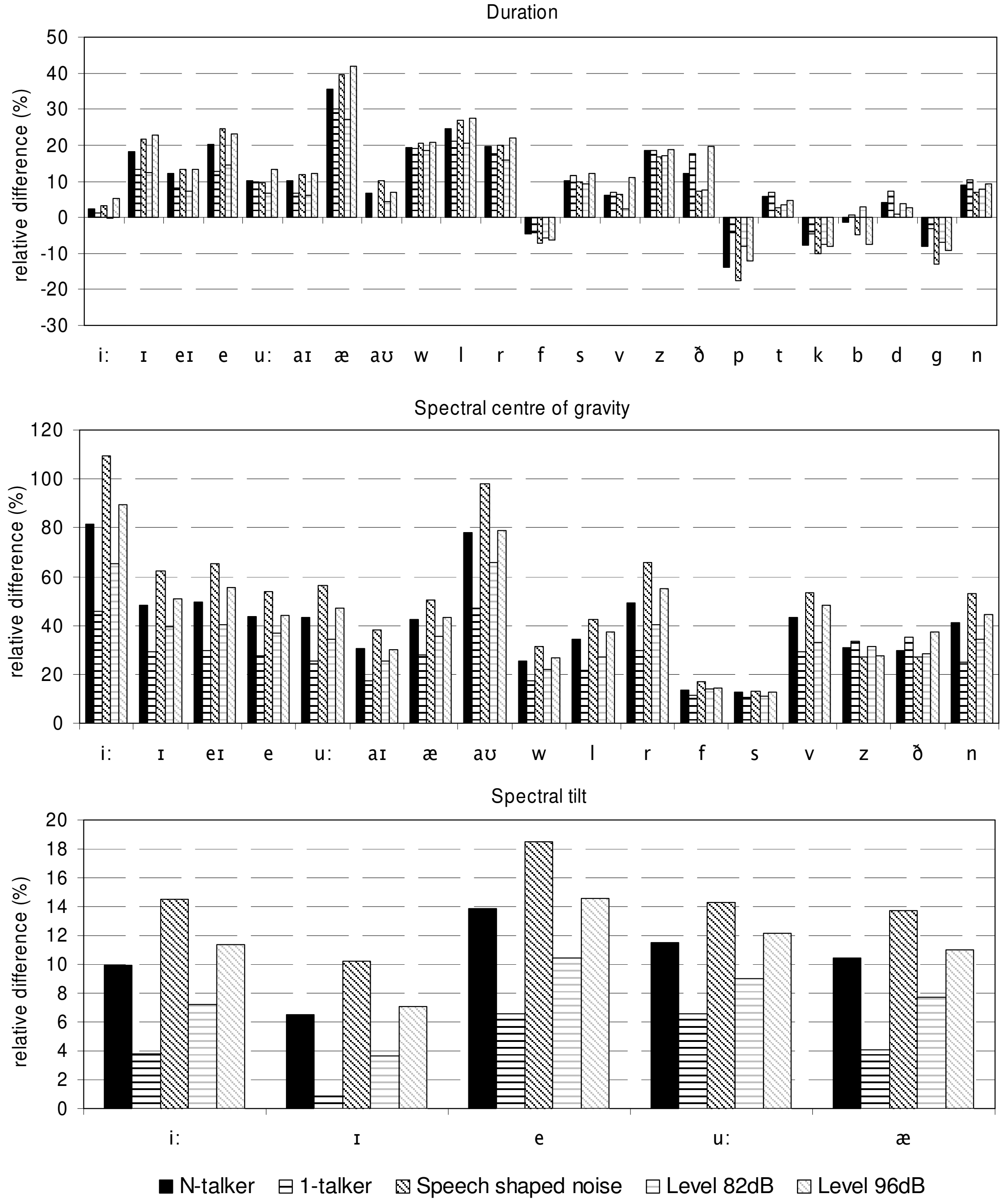


Figure 3.3: *Phoneme-specific differences in duration (top), spectral centre of gravity (middle) and spectral tilt (bottom) in noise and quiet conditions. For ease of display, the noise conditions are grouped into 5 subsets. N-talker: all 10 noise backgrounds; 1-talker: N1_82, N1_89 and N1_96; speech-shaped noise: Ninf_82, Ninf_89 and Ninf_96; level 82dB: N1_82 and Ninf_82; level 96dB: N1_96 and Ninf_96.*

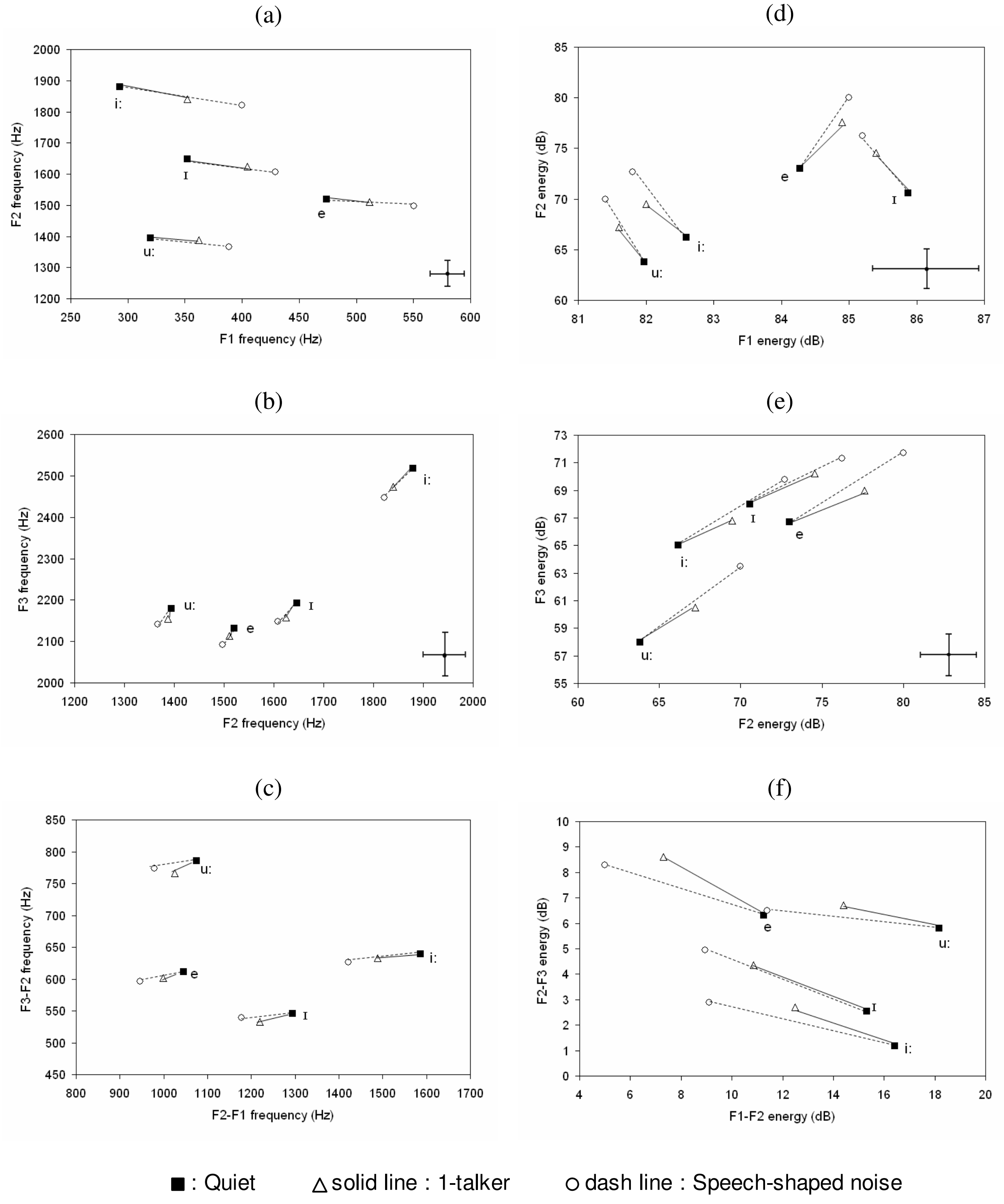


Figure 3.4: *Formant frequencies (left) and energies (right) for the vowels in "green", "bin", "red" and "soon", for speech produced in quiet and noise. In each case, values are averages taken from the central three frames over all instances of the vowels. For clarity, averages across the 3 single talker and 3 speech-shaped noise conditions are shown. Error bars in the lower-right corner indicate 95% confidence intervals.*

Compared to quiet, increases in *N* and masker level led to an increase in the duration of most sound types apart from the fricative /f/ and the non-alveolar plosives, for which a slight shortening was observed. Increases in spectral centre of gravity with the increase in *N* and the rise of noise level were seen for all sounds. For most, the increase was substantially larger than 25%, although the fricatives /f/ and /s/ showed only modest increases. Similar findings for the duration and centre of gravity of vowels have been reported for stationary noise (Junqua, 1993; Stanton et al., 1988; Garnier et al., 2006). Vowel spectral tilt became flatter in all conditions, with differences in degree between the vowels. Speech-shaped noise led to a flatter spectral tilt compared to competing talker background. In addition, the masker with a higher level yielded a larger change in vowel spectral slope. Such a pattern was also reported in Varadarajan and Hansen (2006) for stationary noise.

In addition, similar statistical analyses to those used for utterance-level parameters were carried out for formant frequencies, energies and bandwidths for each vowel. For speech produced in noise, F1 frequency increased significantly by up to 100 Hz. Such effects were stronger for speech-shaped noise, compared to a competing talker background, for all the vowels. F2 and F3 frequencies fell by as much as 60 Hz and 80 Hz respectively but these tendencies were only statistically significant for the vowels /iː/ and /ɪ/. For F2 and F3 frequencies, no significant differences were found between competing talker and speech-shaped noise. Increases in vowel F1 frequency were also seen in earlier studies (Pisoni et al., 1985; Summers et al., 1988; Bond et al., 1989; Takizawa and Hamada, 1990; Junqua, 1993; Garnier et al., 2006; Bořil, 2008). For F2, Junqua (1993) reported increases for females while Pisoni et al. (1985) found the opposite for both males and females. Other studies (Summers et al., 1988; Bond et al., 1989; Garnier et al., 2006) demonstrated a large amount of vowel and utterance-dependent pattern of F2 frequency change. Junqua (1993) suggested that the F3 frequency of vowels tends to remain constant in noise.

Significant increases in F2 and F3 energy for the vowels compared to the quiet condition were measured. Such effects were significantly stronger for speech-shaped noise compared to the competing talker background. F1 energy changed little. The overall effect of formant energy changes is consistent with observed changes in spectral tilt.

Furthermore, significant increases in F1 bandwidth and decreases in F2 and F3 bandwidths for all the vowels compared to the quiet condition were found. For most of the vowels, changes in F1 and F2 bandwidths tended to be significantly larger for speech-shaped noise compared to competing talker background while for F3 bandwidth, such tendencies were only significant for the vowels /i:/ and /ɪ/. The changes in F2 and F3 bandwidth for speech produced in noise are consistent with those reported in Hansen and Bria (1990) and Mixdorff et al. (2006). For F1 bandwidth, Hansen and Bria (1990) found an increase for /i:/ and /ɪ/ and a decrease for /e/ while Junqua (1993) suggested a decreasing tendency for most of the vowels.

3.3.3 Correlation analysis

The above analyses treat speech production changes as independent of each other, but it is possible that correlated changes exist in acoustic parameters such as F0 and F1 frequencies as a result of speech energy changes. Correlations between energy and both F0 and F1 frequencies were investigated. The Pearson correlation coefficient between energy and F0 and energy and F1 frequency was computed independently for all voiced segments. To arrive at a single correlation measure, the weighted mean of segment-based correlations was derived, with weights given by segment duration. As shown in figure 3.5, for energy versus F0, there was a slight but significant *decrease* (p<0.05) in correlation in most of the noise conditions compared to quiet (r=0.37). The correlation decreased significantly with an increasing number of background talkers (F(2.9,20.1)=3.98, p<0.05, η^2=0.46). For energy and F1 frequency, significantly *increased* (p<0.05) correlation was found in all noise conditions compared to quiet (r=0.23). Correlations also increased with the number of talkers (F(3.8,26.4)=6.50, p<0.01, η^2=0.78).

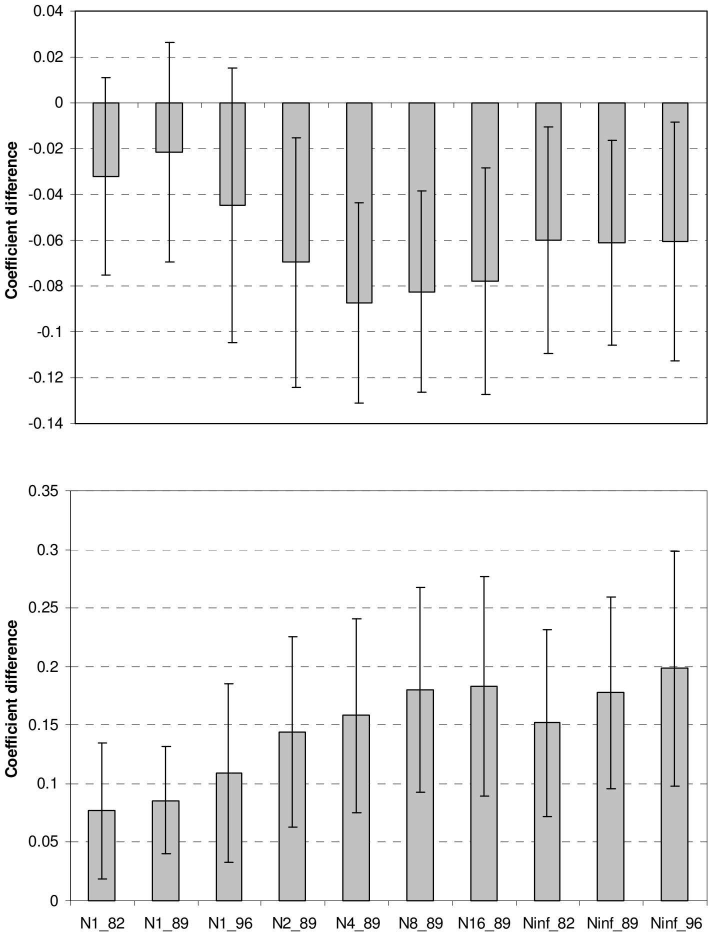


Figure 3.5: *Differences between correlation values (top: energy vs F0; bottom: energy vs F1) for each noise condition compared to speech produced in quiet.*

3.3.4 Discussion

The current results generally confirm the effects of stationary noise on speech production found in previous studies, both at the level of overall acoustic parameter values and for individual phoneme classes. More importantly, they demonstrate for the first time the effect of the number of talkers making up the background babble, including the case of a single talker. For nearly all of the parameters where there is a significant difference between speech produced in stationary noise and in quiet, there is a similar, but smaller, effect when a single talker speaks in the background while speech is produced. Similarly, changes in noise level which have an effect in the stationary noise case tend also to affect the single talker case. The effect of intermediate background conditions (i.e. multi-talker babble for more than one talker) usually falls somewhere between the two extremes. For all parameters, no interaction between the effects of noise level and the number of background talkers was present. One interpretation of these results is that the Lombard effect is influenced by both noise level and number of background talkers, acting independently.

For those parameters which might be expected to reflect the differences in information conveyed by the background, namely sentence start time and the statistics of short pauses, some small differences were found. There were more pauses longer than 20 msec in the single talker background than in the other conditions. The pause prior to speaking was longer in the single talker background than for most of the babble conditions, although the pause was slightly shorter than in the stationary noise case. It is possible that the non-communicative task limited the scope for such effects.

Some acoustic effects might be the consequence of intentional changes while others may be secondary, caused by articulatory constraints. For example, as pointed out by Schulman (1989) and Gramming et al. (1988), the raising of subglottal pressure in order to create a louder voice causes an increase in F0. On the other hand, it is also possible that in the production of high-pitched voice, SPL is raised due to a larger number of speech pressure cycles per time unit resulted from the increase in F0. In addition, the wider jaw opening in order to increase sound amplitude induces an increase in the first formant frequency (Lindblom and Sundberg, 1971). In the current study, correlations between F0 and energy actually decreased in noise, although F1 frequency and energy became more correlated. Thus, it is possible that speakers were using intentional changes in both energy and F0 in response to noise. It is likely that other factors such as physiological and semantic constraints on possible F0/F1 values and range also limit the extent to which speakers can manipulate these parameters independently.

3.4 Intelligibility of noise-induced speech

3.4.1 Motivation

Speech produced in the presence of noise can lead to increases in intelligibility over speech produced in quiet mixed with equivalent noise tokens at the same SNR (Dreher and O'Neill, 1957; Summers et al., 1988). The speech material collected in the current study employed a wider range of noise backgrounds, allowing several new issues to be explored. First, the general finding that the effect of noise on certain acoustic parameters tended to increase with both noise level and number of talkers (*N*) suggests that any intelligibility gains may also be influenced by noise level and *N*. Experiment I measured speech intelligibility as a function of noise level and *N* for noise-induced speech compared to speech produced in quiet with added noise.

When faced with the task of communicating in the presence of a single competing talker, talkers might adopt strategies to reduce both the energetic masking and informational masking components at the ear of the listener. Two further experiments explored these possibilities. In experiment II, listeners were presented with utterances masked by a competing talker. The intelligibility of utterances produced in quiet was measured and compared to that of the same set of utterances induced by a competing talker when presented in the background of the inducing competing talker maskers. Any intelligibility gains in the latter 'matched' case might be interpreted as resulting from a talker's awareness of the informational masking effect of the competing utterance. However, increases in intelligibility could also be derived from reductions in energetic masking due to acoustic changes in the competing speaker-induced utterances. Experiment III attempted to distinguish the two hypotheses by comparing the intelligibility of speech produced in the presence of a competing talker when presented in the matched competing talker background with the same utterances presented in an unmatched competing talker background. If talkers are sensitive to the informational masking potential of a specific competing utterance rather than the energetic masking properties of speech in general, listeners should produce higher scores in the matched condition.

3.4.2 Experiment I: Sentences in stationary noise

A. Listeners

Twelve native speakers of British English (9 males and 3 females) drawn from the undergraduate and postgraduate population of the Department of Computer Science at the University of Sheffield took part in experiment I. All subjects received a hearing test using the same software and procedure as described in section 3.2.3. All had normal hearing apart from one participant with a hearing level of 25 dB in one ear at 8 kHz. This subject was retained for the study.

B. Stimuli

Utterances collected in quiet and in the presence of noise were presented in a background of stationary speech-shaped noise. Five sets of 100 utterances balanced across the 8 talkers were used, corresponding to speech produced in quiet, in a background of a competing talker at levels of 82 and 96 dB SPL, and in a background of stationary noise at 82 and 96 dB SPL. In all five conditions of experiment I, utterances were mixed with a speech-shaped noise masker at an overall SNR of -9 dB, a value chosen on the basis of pilot tests to reduce ceiling and floor effects. Prior to mixing, target utterances were scaled to have the same rms level. Maskers were gated on and off with the endpointed utterances and the mixed signals were scaled to a level of approximately 68 dB SPL.

C. Procedure

Experiment I took place in an IAC single-walled acoustically-isolated booth. Stimuli presentation and results collection was controlled by a computer program. Stimuli were presented diotically over Sennheiser HD 250 Linear II headphones via a Tucker-Davis Technologies (TDT) System 3 RP2.1. Listeners were given instructions to identify in each noisy utterance the letter and digit keywords. This they did via a computer keyboard whose keys were selectively activated to minimise keying errors. For consistency with later experiments, in which the colour keyword was used to identify the target utterance, sentences within each condition were organised into 4 blocks by colour keyword. Condition order was balanced across listeners while both colour blocks and utterance order within blocks were randomised for each listener. The experiment took place in a single session which was preceded by a short practice. In addition, four practice tokens were added to the start of each condition. Listeners were unaware of these tokens and they were not scored. The entire session required around 30 minutes to complete.

D. Results

For utterances produced in quiet and presented in speech-shaped noise, listeners obtained a mean keyword identification score of 42%. However, for the 4 conditions involving the identification of utterances produced in a noise background, keyword scores were substantially higher. As shown in figure 3.6, the increase in scores for noise-induced speech ranged from 9 to 25 percentage points. These increases were statistically significant (p<0.01 in the single talker 82 dB condition; p<0.001 in the other 3 conditions) (by paired t-tests with Bonferroni-adjustment). The two single talker backgrounds led to the smallest improvements, and in both the single talker and stationary noise backgrounds, the gain in intelligibility increased with noise level. Among the four noise-induced speech conditions, a two-way repeated measures ANOVA with factors of $N=\{1, \infty\}$ and level={82, 96 dB} found a significant effect of N (F(1,11)=27.28, p<0.001, η^2=0.96) and noise level (F(1,11)=8.28, p<0.05, η^2=0.44). The N by noise level interaction was not significant (p=0.2).

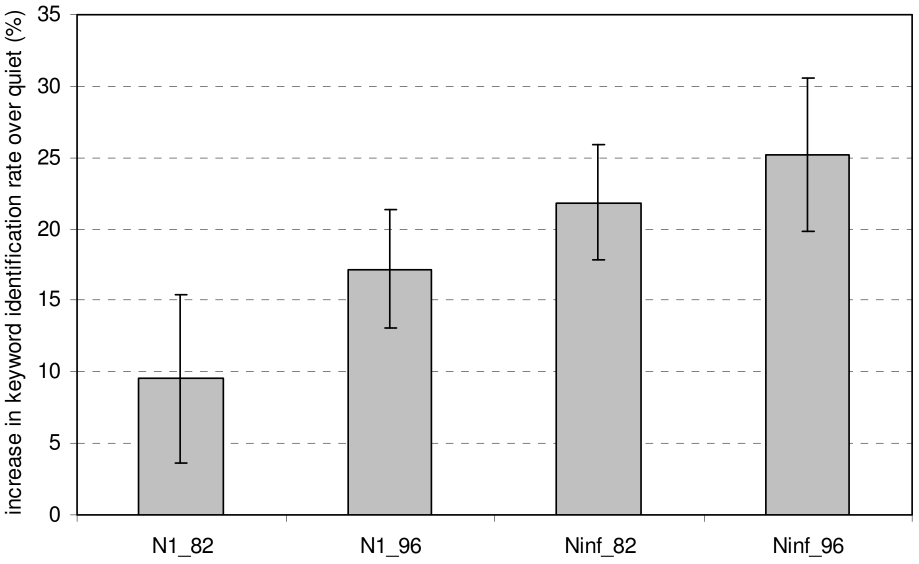


Figure 3.6: *Keyword identification rates for noise-induced speech over speech produced in quiet, when added to speech-shaped noise (experiment I). The baseline keyword identification score for utterances produced in quiet is 42%.*

3.4.3 Experiment II: Sentences in competing utterances

A. Listeners, stimuli and presentation

Listeners who took part in experiment I also took part in this experiment. Four conditions tested the identification of keywords in utterances when presented in a competing speaker background. In two conditions, listeners heard speech produced in quiet conditions added to other speech material produced in quiet, drawn from the same corpus (Cooke et al., 2006). In the other two conditions, listeners heard speech that was produced in a competing speech background added to that competing speech background. These "speech-induced" conditions were drawn from those collected as described in section 3.2, and corresponded to the 82 dB and 96 dB background levels. Both "quiet" conditions were identical apart from the choice of sentences used for the background. Two “quiet” conditions were used to enable the same set of speech maskers to be used in the “speech-induced” and “quiet” conditions.

As for experiment I, 100 utterances were used for each condition. For this experiment, sentences contained no keywords in common with those of the masker. Sentences were added so that the target to masker ratio was -9 dB, a value chosen on the basis of pilot experiments and known to be able to invoke informational masking effect in a speech-on-speech recognition task (Brungart, 2001; Cooke et al., 2008). Maskers were gated on and off with the endpointed target sentence. Due to the approach taken to the generation of competing speech maskers as described in section 3.2, the start of a sentence did not necessarily coincide with the start of a sentence in the masker. In this respect, the 2-talker scheme was different from those used in informational masking experiments (e.g. Brungart, 2001; Cooke et al., 2008).

The stimulus presentation setup was as described in experiment I. Since this task involved identifying a target in a very similar masker, listeners required information to distinguish the target and masker sentences. The color keyword was used to indicate which utterances listeners had to attend to. The corpus contains 4 color keywords, so stimuli were organized into 4 blocks within each condition. At the start of each block, listeners were instructed (via the computer screen) to identify the letter and digit in the sentence containing a given color.

B. Results

Figure 3.7 displays the difference in keyword identification rates between the “speech-induced” and “quiet” utterances for the two levels 82 and 96 dB. While the speech-induced utterances are more intelligible for both levels, only the 96 dB case reaches statistical significance (p<0.05) (with Bonferroni-adjustment), suggesting that speech produced in sufficiently intense backgrounds containing a single competing talker is more intelligible than speech produced in quiet when added to the same competing talker material. This finding extends that of experiment I to a highly non-stationary masker. However, the absence of an effect for speech produced in less intense backgrounds calls into question the extent to which this effect is due to an attempt by the speaker to minimize the degree of informational masking at the ear of the listener.

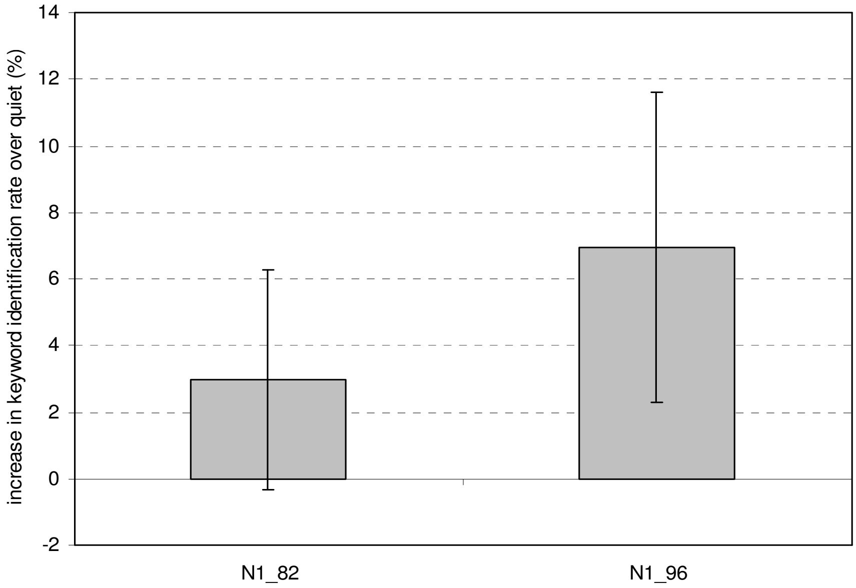


Figure 3.7: *Keyword identification rates for utterances induced by a speech background over utterances produced in quiet, when added to the inducing speech (experiment II). The baseline keyword identification scores for utterances produced in quiet are 81% and 78% respectively.*

3.4.4 Experiment III: Induced speech in matched and unmatched backgrounds

A. Listeners, stimuli and presentation

Listeners who participated in experiments I and II also took part in this experiment. Experiment III compared two conditions, one in which the target material consisted of speech induced by other speech was presented in the background of the inducing speech material ("matched"), and one in which the same target speech was presented in "unmatched" backgrounds. Target speech consisted of utterances collected as described in section 3.2 in the presence of a competing talker presented at 89 dB SPL. All other stimulus construction and presentation details were the same as for experiment II (100 utterances, -9 dB target-to-masker ratio, and presentation of targets blocked by colour keyword). Experiments II and III were performed in sequence in the same session, which lasted approximately 30 minutes.

B. Results

Keyword identification score in the "matched" condition was 1.4 percentage points higher than in the "unmatched" condition (score of 86%). However, this failed to reach statistical significance at the 95% level (*p*=0.08). This outcome suggests that, in this task, talkers do not modify their productions in response to the details of a specific competing utterance.

3.4.5 Discussion

The three perceptual experiments here explored the extent to which the presence of competing speech and stationary noise influences the intelligibility of speech productions. Experiment I confirmed previous findings on the increased intelligibility of speech produced in stationary noise backgrounds (Dreher and O'Neill, 1957; Summers et al., 1988; Pittman and Wiley, 2001) and extended these results to single talker maskers. The size of intelligibility gains was closely correlated with the extent of acoustic changes measured in section 3.3: stationary noise backgrounds and intense background level both resulted in larger intelligibility gains than single-talker backgrounds and less intense backgrounds. However, all backgrounds tested resulted in significant gains in intelligibility.

Experiments II and III employed maskers designed to invoke large amounts of informational masking to explore the possibility that talkers modify their production strategy dynamically in response to the presence of competing speech. Experiment II demonstrated that speech produced in an intense competing speech background was more intelligible than speech produced in quiet when presented in the same background. However, for speech produced in a less intense background, no such difference was found, suggesting that energetic rather than informational masking is dominant, since the lower background intensity during production (82 dB SPL) is still relatively strong and could be expected to produce informational masking effects. It seems likely that similar principles as those leading to modifications in production for speech produced in stationary noise backgrounds were operating in the competing speech condition.

The results of experiment III do not support the idea that talkers modify their productions in response to the details of individual competing utterances in order to improve intelligibility at the ear of the listener. There was no significant difference in identification scores between speech produced in the speech backgrounds for the maskers which induced the utterances compared to the same induced utterances presented with random speech maskers.

3.5 Energetic masking analysis of noise-induced speech production

3.5.1 Motivation

While the finding that noise-induced speech is often more intelligible when presented in noise has been reported in studies dating back many years (Dreher and O'Neill, 1957) and has been confirmed here, little effort has been directed toward an explanation of the intelligibility gain. Here, we test the hypothesis that the intelligibility of noise-induced speech is related to the availability of "glimpses" of speech at the ear of the listener in the presence of noise. Glimpses of a signal are defined as those connected regions in its spectro-temporal representation over a certain minimum "area" calculated from the number of spectro-temporal "pixels" and where each "pixel" has a local SNR larger than a threshold (Cooke, 2006). This hypothesis is grounded in the energetic masking produced by the masker, and differs from an explanation based solely on the energetic masking to be found in the auditory periphery in that glimpses incorporate the idea that listeners only have access to spectro-temporal regions which are sufficiently dominant in both local SNR and spectro-temporal extent to allow them to stand out above the masker (Cooke, 2006). Here, a spectro-temporal "pixel" corresponds to a single time frame and frequency channel in the spectro-temporal representation. "Pixels" of a region were deemed to be connected if they were part of the 4-neighbourhood (i.e. excluding diagonal neighbours) of any other pixel in the region.

3.5.2 Glimpse measures

Two glimpsing statistics were measured for the signal mixtures used in the intelligibility experiments described in the previous section. One, “glimpse area”, is the number of spectro-temporal points where the glimpse criteria described above hold. Since glimpse area will typically increase with signal duration, “glimpse proportion” was also computed, defined as the proportion of spectro-temporal points which meets the glimpse criteria. This latter measure is independent of duration, and helps to distinguish simple speech production processes which improve glimpsing opportunities by slowing speech rate from those which reallocate energy in time and frequency to improve glimpsing opportunities.

Computation of glimpse measures was based on a spectro-temporal excitation pattern (STEP) representation formed for the target and masker independently. A STEP is produced by first passing the time-domain signal through a 64 channel gammatone filterbank, smoothing the Hilbert envelopes, integrating the energy into 10 msec frames, followed by log compression. More details of the computation can be found in Cooke (2006). A minimum area of 5 and a local SNR of -5 dB were used here since Cooke (2006) suggested that it may be unreasonable that listeners can detect very small regions of favorable local SNR when surrounded by masker and the best fit to behavioral data came from a computational model that treated all regions with local SNR in excess of -5 dB as potential glimpses.

3.5.3 Results

Figure 3.8 shows the two glimpse measures for each of the conditions used in the experiments of the previous section. For the stationary noise conditions corresponding to experiment I, significantly more glimpses (as measured by both area and proportion) were produced by the noise-induced speech than for speech produced in quiet (p<0.01) (by paired t-tests with Bonferroni-adjustment). Stationary noise maskers produced more glimpses than the competing speech (F(1,7)=14.50, p<0.001, η^2=0.94 for area; F(1,7)=10.513, p<0.01, η^2=0.72 for proportion) while the effect of an increase in noise level was significant only for the glimpse area measure (F(1,7)=6.44, p<0.05, η^2=0.39). Regarding the two competing talker conditions of experiment II, both show significantly more glimpses than speech produced in quiet when measured in terms of glimpse area (p<0.05 for the less intense condition; p<0.01 for the more intense condition) (by paired t-tests with Bonferroni-adjustment) while there is a small increase in glimpse proportion for the more intense condition (p<0.05) (with Bonferroni-adjustment). Finally, as was the case for intelligibility, no significant effect was found for experiment III (p=0.1).

Overall, the results are strikingly similar to those for intelligibility, as illustrated by figure 3.9 which plots relative intelligibility gains for listeners against relative increases in the two glimpse measures. Both measures are highly-correlated with listener intelligibility gains, suggesting that noise-induced speech is more intelligible than speech produced in quiet because the articulatory manipulations lead to a release from energetic masking.

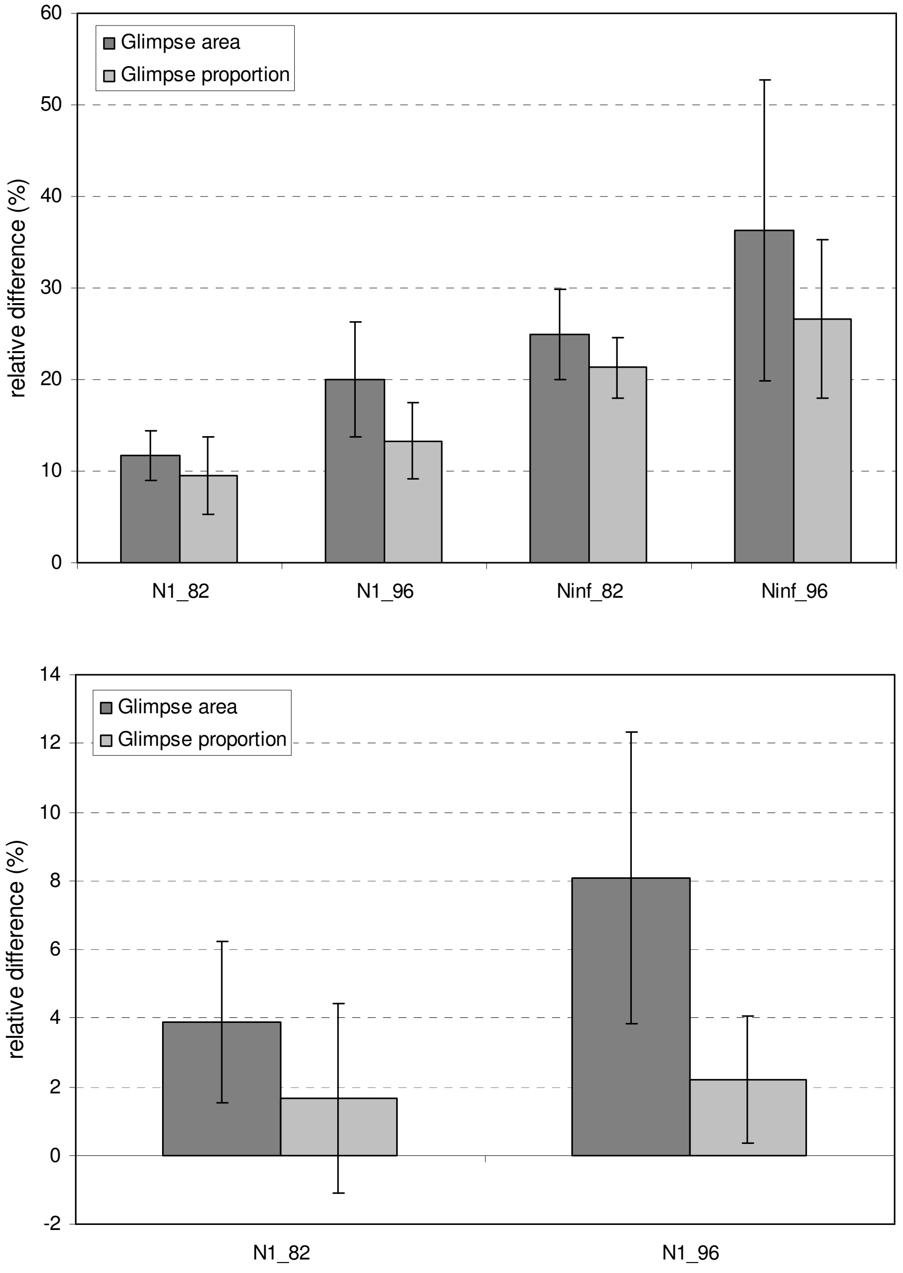


Figure 3.8: *Glimpse area and proportion for the listening conditions of experiment I (top) and II (bottom), expressed as percentage increase in area or proportion over speech produced in quiet.*

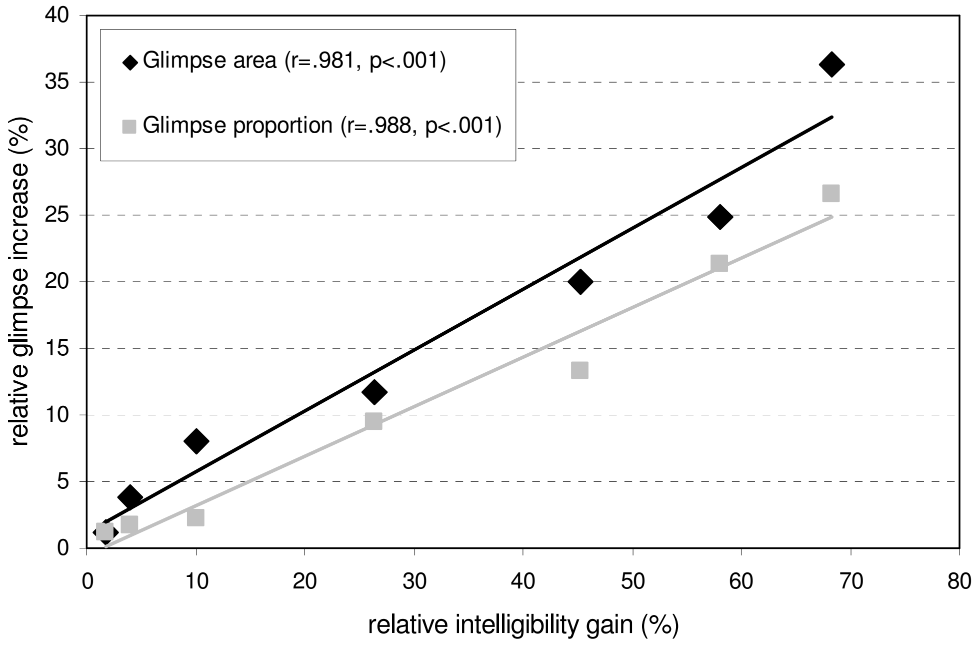


Figure 3.9: *Relation between increases in the two glimpse measures and increase in intelligibility for experiments I, II and III, together with least-squares fits.*

Of the two glimpse measures, significantly larger increases in glimpse area over glimpse proportion were found (F(1,7)=4.10, p<0.05, η^2=0.41; F(1,7)=7.39, p<0.05, η^2=0.47) for the conditions of both experiments I and II. This is presumably due to the tendency of noise-induced sentences to increase in duration. No significant correlation (p=0.1) was found between utterance-wise measures of duration and glimpse proportion in any of the noise-induced conditions, suggesting that speakers use both a slower speaking rate, to increase the overall number of glimpses, and other (mainly spectral) modifications in order to increase the proportion of glimpses available for the hearer. The pattern of an increased amount of the time-frequency plane glimpsed, as a result of spectral energy shift to higher frequencies, together with durational lengthening, is illustrated in figure 3.10 using a simple 6-word sentence, produced in quiet and in 3 Lombard conditions. However, in explaining listener performance, there is no clear basis to prefer glimpse area over glimpse proportion since there is a mixed evidence on the contribution of a slower speaking rate to intelligibility. For instances, while several researchers (e.g. Cox et al., 1987; Jones et al., 2007) have demonstrated that slower speaking rates led to increased speech intelligibility in noise, others (e.g. Sommers, 1997; Uchanski et al., 2002) have failed to find a perceptual correlate of speaking rate.

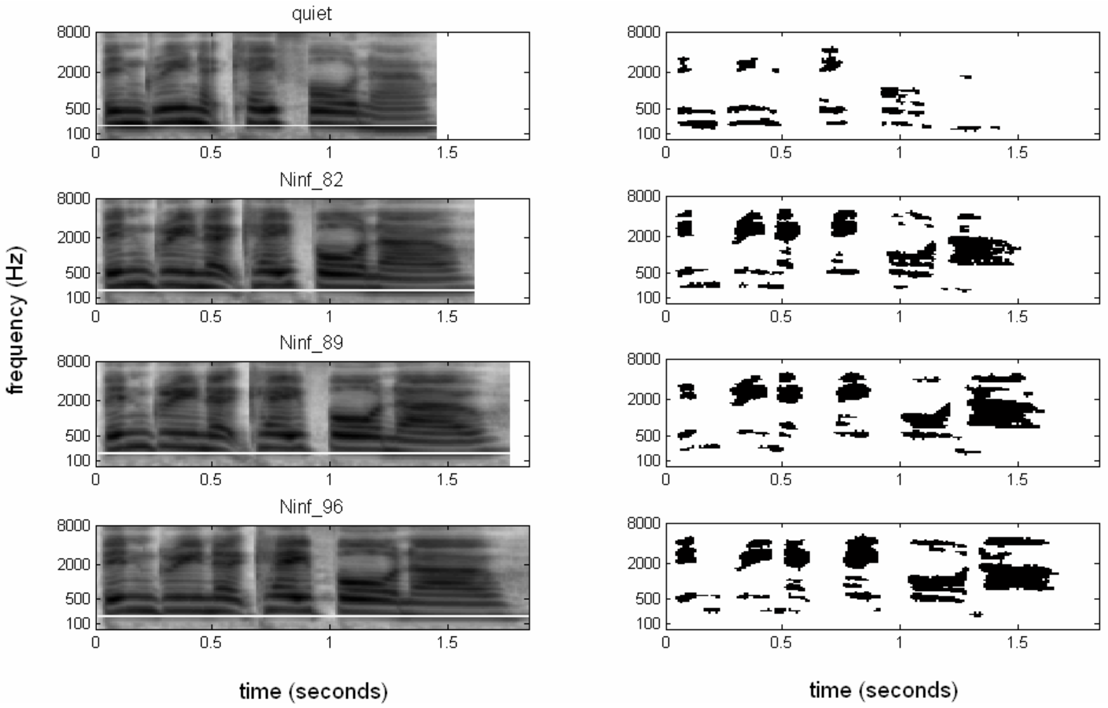


Figure 3.10: *Spectro-temporal excitation patterns (left column) and glimpses (right column) for the sentence "bin green at k 4 now" in quiet and 3 Lombard conditions, spoken by a female. Effects of spectral energy migration to higher frequencies and temporal duration lengthening on increasing glimpses are visible. Horizontal lines in the excitation patterns indicate a frequency of 200 Hz.*

3.6 General discussion

3.6.1 Noise-induced speech and energetic masking

A number of reliable and consistent acoustic modifications occur when speech is produced in the presence of noise. The main effects – increases in F0, energy and spectral centre of gravity – confirm those found in previous work using multi-talker babble and stationary noise. The current study extends the scope of noise-induced production effects to single-talker interfering speech, which was also found to be capable of producing significant acoustic changes compared to speech produced in a quiet background. Increases in F0, energy and centre of gravity grew as the number of talkers in the background increased, asymptoting at around 8-16 talkers. These results demonstrate that the extent of acoustic modifications is largely correlated with both the intensity of the background signal and number of background talkers. This suggests that noise-induced speech production changes are dependent on the overall energetic masking capacity of the background signal, since energetic masking is a function of both overall noise level and number of background talkers: a competing talker is a far less effective energetic masker than a broadband noise when both are presented at the same SNR (Festen and Plomp, 1990), and energetic masking increases with the number of background talkers in a task of consonant identification (Simpson and Cooke, 2005) as well as sentence recognition (Bronkhorst and Plomp, 1992).

Experiment I demonstrated that noise-induced speech was more intelligible when presented in stationary noise than speech produced in quiet, extending previous findings for stationary and multi-talker babble backgrounds. Interestingly, those backgrounds which resulted in the largest acoustic modifications also produced the biggest increases in intelligibility, suggesting that speakers modify their productions in response to the adversity of the background. Indeed, the result of production modifications is to increase the number and proportion of opportunities to glimpse the target speech in noise, and the increase in such opportunities is very closely correlated with listener keyword identification performance. Thus, the potential for energetic masking leads to articulatory modifications whose acoustic consequence is to cause a release from masking, and the more masking potential that exists, the greater the eventual release. These findings support Lindblom's suggestion that speakers compensate for environmental conditions (Lindblom, 1990), rather than a purely physiological and reflexive interpretation of the Lombard effect.

3.6.2 Basis for the increased intelligibility of noise-induced speech

It is not clear how the acoustic consequences of changes in speech production lead to increased intelligibility. While it is evident that the overall increase in intensity of noise-induced speech produces a release from energetic masking, this cannot account for the intelligibility gains observed here since all utterances were normalised to have the same SNR when presented alongside maskers. However, speakers can employ a number of other strategies to improve the SNR at the ear of the listener. For instance, a decrease in speaking rate provides more opportunities to glimpse acoustic information useful for phonetic distinctions. The largest increase in utterances duration of around 7% in the most adverse backgrounds might have contributed to the overall improvement in intelligibility, but it is unlikely to be responsible for the entire increase since the results of the glimpsing analysis showed that the *proportion* of the spectrum lying above the masker also increased for the noise-induced conditions.

Many of the acoustic consequences of noise-induced speech are compatible with an overall shift in the energy balance from lower to higher frequencies. For the vowels, increases in fundamental frequency, spectral centre of gravity, and energy for the second and third formants are reflected in a flattening of spectral tilt. One consequence of this shift to higher frequencies is a certain degree of masking release in the presence of the maskers employed in this study, whose mean spectrum was speech-shaped. However, vowel formant frequencies became more “central”, with increases in F1 and decreases in F3. Of course, there are articulatory limits to the range of speech production modifications possible, and some of the acoustic changes observed may be epiphenomena associated with other manipulations such as increased effort and vocal stress.

The issue of whether the shifting of spectral energy to high frequencies in the presence of speech-shaped noise results from speakers’ active attempt to place spectral information in locations where it is less likely to be masked merits further study since such a tendency is also reported when talkers are asked to speak clearly without any noise in the background (Picheny et al., 1986; Payton et al., 1994). Formant frequency changes are also found in clear speech compared to normal speech (Chen, 1980; Picheny et al., 1986; Krause and Braida, 2004; Smiljanić and Bradlow, 2005). These studies categorized vowels as tense or lax, corresponding to /i:, u:/ and /ɪ, e/ here. No consistent trends were found for the first three formant frequencies of tense vowels, although Chen (1980) reported that tense vowels clustered more tightly in vowel formant space in clear than in conversational speech. Picheny et al., (1986) and Krause and Braida (2004) reported increases in F1 and F2 of the lax vowel /ɪ/ of up to 50 and 200 Hz respectively.

3.6.3 Speech changes produced by a competing talker

One of the motivations for the current study was to determine how the presence of a competing talker affects speech production. One possibility is that competing speech material might disrupt the speech production process of the talker, resulting in false starts, hesitations and other dysfluencies. The speech material used in this chapter was deliberately chosen to be similar to that introduced in the background in order to provoke such effects. Some disrupting influence of the competing talker background was found: the number and duration of short pauses increased with intensity while no similar effects were seen for the stationary noise backgrounds. Further, the number of false starts was larger in the intense single talker background than in quiet ($t(7)=2.65$, $p<0.05$). However, these effects were small and the overall number of short pauses was not significantly greater than in a quiet background.

A second potential influence of competing speech is on the talker-listener communication process: the talker might anticipate the informational masking effect of two similar utterances at the ear of the listener and employ strategies to reduce the extent of informational masking. Experiment II demonstrated that utterances produced in the presence of an intense competing talker were more intelligible than utterances produces in quiet conditions when presented in speech backgrounds. For speech produced with a less intense talker, there was no significant gain over quiet. These findings suggest that it is primarily energetic rather than informational masking that leads to increased intelligibility since if the latter were at work, some effect in the less intense background would be expected since the production and background levels are closer and lead to more informational masking for the listeners (Brungart, 2001). Further, no evidence was found of speaking strategies which exploited the temporal fluctuations of specific competing utterances: there was no difference in the intelligibility of speech in the presence of the material which induced it when compared to speech in the presence of other speech material (experiment III). Talkers may be unable to attend to and track competing speech material sufficiently rapidly to modify their own productions in response.

3.6.4 Task-dependence

While few effects of a competing talker above and beyond energetic masking were found here, it is possible that other tasks might elicit more extensive speech production changes. The task employed in the study of this chapter was devoid of communicative intent, and it was possible for speakers to read the prompts on the screen with little regard for the specific information contained in the competing talker background. The possibility of a strong influence of the communication factor on speech production has been suggested by the studies of Lane and Tranel (1971), Summers et al., (1988) and Junqua et al., (1999). Further studies using two-way interactive task with single-talker maskers need to be conducted before ruling out the possibility of both positive effects of active strategies which are sensitive to the local masking conditions and negative effects of attentional deployment to processing an informative background source while speaking.

It is known that the greatest informational masking effects are found when the target signal and masker are similar in terms of linguistic content as well as talker characteristics (Treisman, 1964; Brungart, 2001; Cooke et al., 2008). Indeed, although the masking utterances were similar in form to those produced by the talkers, start times were not synchronized, so the chances of similar words overlapping were reduced and there could be a change in talker of competing background during the production of a single utterance. It is possible that tasks designed to produce large amounts of informational masking would give rise to more significant changes in speech production than those observed in the study of this chapter.

3.7 Conclusion

It was found in this chapter that speakers modified their productions in *N*-talker noise backgrounds across a wide range of values for *N*. This they achieved not only by increases in output level, but by changes to the fundamental frequency and formant frequencies and energies which result in an overall increase in spectral centre of gravity. The scale of acoustic modifications increased both with *N* and the level of the background noise, conditions which also result in increases in the energetic masking effect of the noise. Noise-induced speech was more intelligible when presented in stationary noise than speech produced in quiet, and the intelligibility gain increased with *N* and noise level. These findings, coupled with a computer model of energetic masking, suggest that speakers attempt to compensate for the energetic masking effect of the noise on their own speech. In contrast, no informational masking effects of a competing talker were found, perhaps because the task lacked a communicative element.

3.8 Compensation for own-voice attenuation

To determine whether own-voice attenuation caused by closed headphones was a factor in the work reported in this thesis, a compensation method was introduced. First, the spectral difference of a white noise signal with and without Sennheiser HD 250 Linear II headphones was measured using a Bruel & Kjaer (B & K) type 4100 head and torso simulator equipped with Bruel & Kjaer (B & K) type 4190 ½ inch microphones, as shown in figure 3.11. An order-32 IIR filter was designed to have a transfer function which was the inverse of the attenuation characteristic produced by the headphones. This filter was implemented on a TDT RP 2.1 processor and compensated for the headphone attenuation in real time.

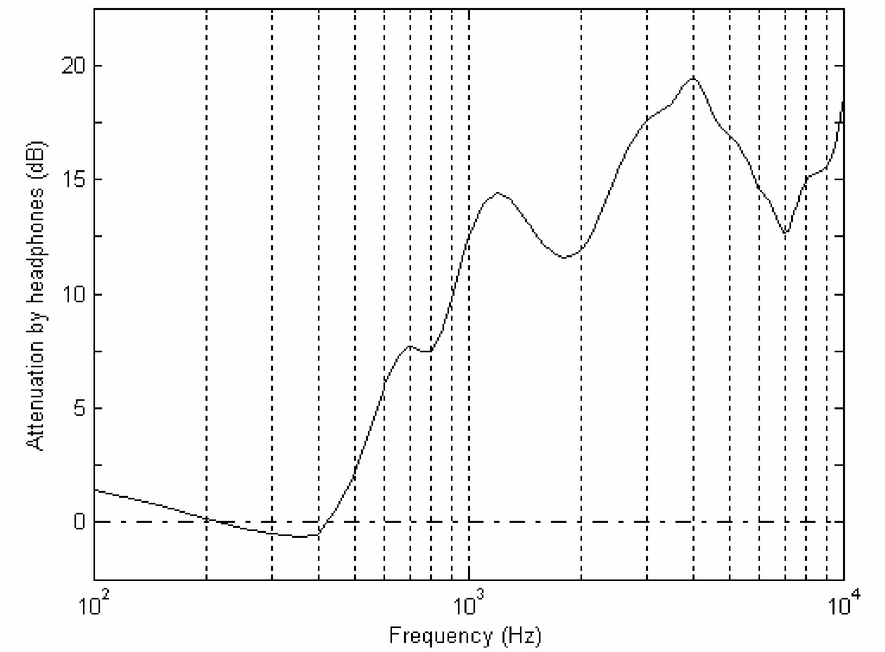


Figure 3.11: *Frequency response of the attenuator.*

In order to discover whether the original and the compensated recording method produced similar effects on speech production, a small corpus was collected using the two methods and analyzed at utterance level. Eight native speakers of British English (4 males and 4 females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. Eight recording conditions were employed which included quiet, competing talker, 8-talker babble and speech-shaped noise. Talkers produced the same set of 25 sentences in each of the eight conditions. Maskers for noise conditions were produced as described in section 3.2.2 and presented at 89 dB SPL. Condition order was randomized for each talker. For the collected utterances, leading and trailing silent intervals identified via the alignment process described in section 3.2.5 were removed.

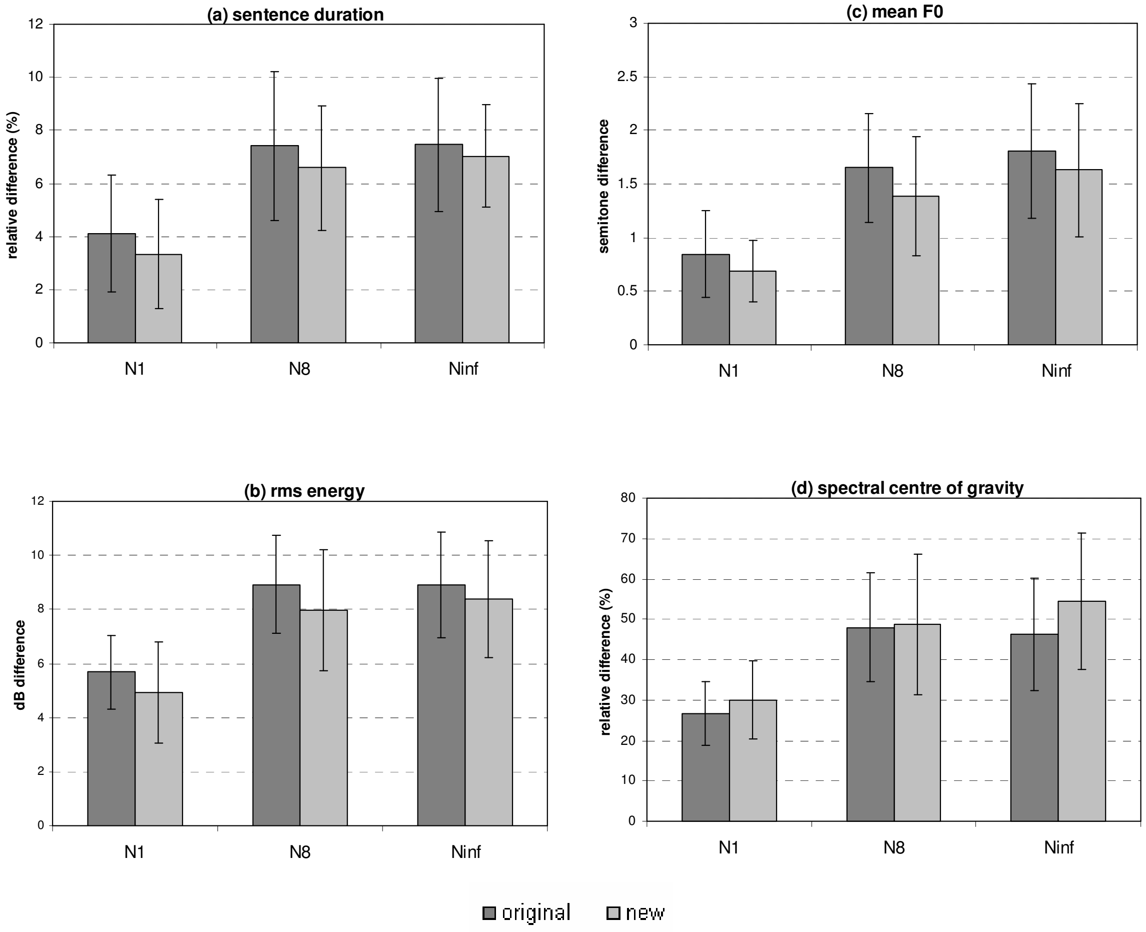


Figure 3.12: *Differences between acoustic parameters values for each noise condition compared to speech produced in quiet for both recording methods, that is the original method employed in section 3.2 and the compensated method used here. The noise conditions of competing talker, 8-talker babble and speech-shaped noise are indicated as "N1", "N8" and "Ninf" respectively.*

Four acoustic properties were estimated for each utterance in each of the 8 conditions. Sentence duration, rms energy, mean fundamental frequency (F0) and spectral centre of gravity (CoG) were computed as described in section 3.3.1. Differences between across-talker means in each background compared to the quiet condition are shown in figure 3.12 for each of the 4 acoustic parameters. A two-way repeated measure ANOVA (two recording methods × three noise conditions) was computed for each acoustic parameter. Post-hoc analysis showed that for all parameters and noise conditions, there was no significant effect of recording method (F(1,7)=0.89, p=0.38 for duration; F(1,7)=1.24, p=0.28 for energy; F(1,7)=0.92, p=0.37 for F0; F(1,7)=1.06, p=0.34 for CoG). For the quiet and competing talker conditions, short pauses (> 20 msec) within each utterance were manually identified and their number and duration computed. Again, the difference in recording setups led to no statistically-significant differences (F(1,7)=0.01, p=0.74 for the number of short pauses; F(1,7)=0.06, p=0.82 for duration of short pauses).

Chapter 4

Speech production modifications produced in the presence of low-pass and high-pass filtered noise [1]

This chapter tackles the issue of whether speakers shift their spectral energy distribution to regions least affected by the noise by examining speech modifications produced in the presence of low-pass and high-pass filtered noise.

4.1 Introduction

As mentioned in section 2.6, previous studies such as Lane et al. (1970), Hansen (1996) and Garnier (2007) have put forward the idea that the Lombard effect is not purely a reflex, but rather driven by the speaker's effort to maintain intelligible speech over noise. In this regard, an active speaking strategy in response to the differing spectral and temporal characteristics of background noise might be expected since one of the effective ways to make the produced speech intelligible at the ears of listeners in the presence of noise could be to take advantage of the regions least concentrated by the noise. Indeed, the findings of Junqua et al., (1998) and Mokbel (1992) have

[1] A version of the work reported in this chapter appeared in Lu and Cooke (2009a).

shown a dependency of the Lombard effect on the spectral shape of the background noise.

The study reported in chapter 3 investigated the effect of *N*-talker babble noise on speech production for *N* ranging from 1 (a single competing talker) to “infinity” (speech-shaped stationary noise), and taking in various multi-talker babble conditions for intermediate values of *N*. Consistent with other Lombard studies, an overall shift in the centre of gravity of energy from lower to higher frequencies was observed at all values of *N*. Further, listeners found Lombard speech substantially more intelligible than speech produced in quiet when both were presented in speech-shaped noise at the same signal-to-noise ratio. Since the long term spectrum of the noise was speech-shaped (for all *N*), an upward shift in centre of gravity causes a degree of release from energetic masking (figure 4.1). Thus, the improvement in intelligibility could be fortuitous, since noise-induced speech changes may coincidentally be in the right direction to be advantageous for the speech-shaped noise maskers. An alternative possibility is that the observed shifts were caused by speakers making an active attempt to place spectral information in locations where it was less likely to be masked. The purpose of the study in this chapter was to distinguish these two possibilities.

Here, changes in speech production were measured in conditions of low-pass, high-pass and full-band speech-shaped stationary noise, relative to quiet. If speakers adopt an optimal strategy in order to minimize the effect of noise on listeners, they would be expected to shift their spectral centre of gravity downwards for high-pass filtered noise condition compared to quiet, and in the opposite direction for low-pass noise condition. For each of the high and low-pass conditions, two noise bandwidths were used to investigate the effect of varying the size of the noise-free part of the spectrum. Again, a "listener-optimal" speaking strategy should lead to greater changes for the smaller noise-free regions because the shift in speech spectral energy would need to be larger to reach the clean parts of the spectrum.

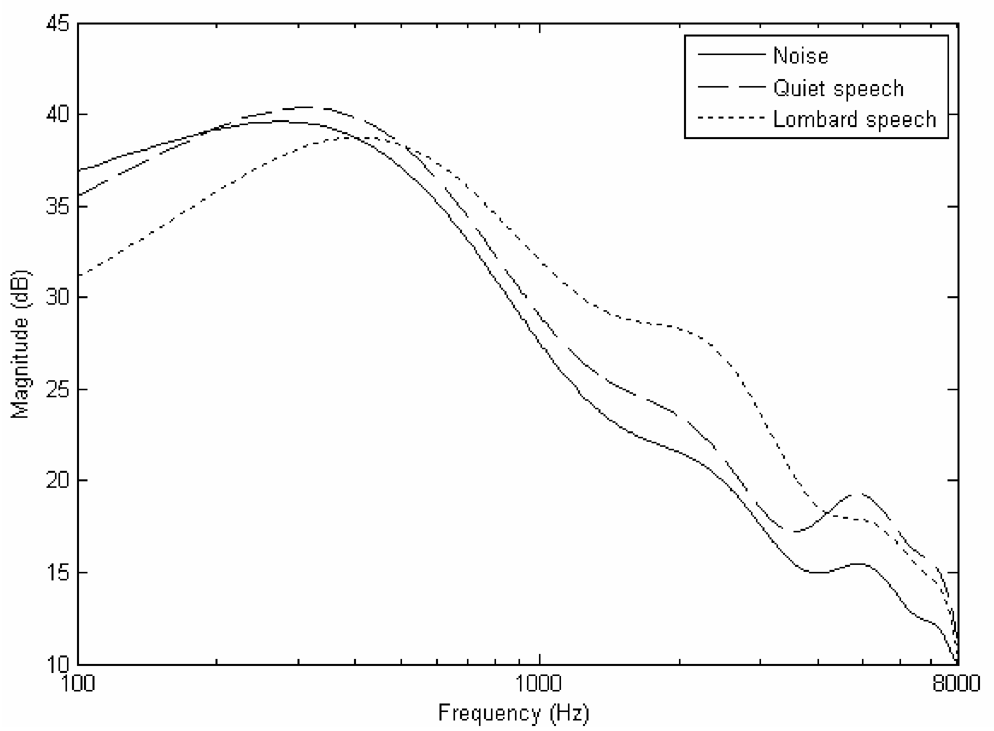


Figure 4.1: *Long-term average spectra of speech-shaped noise, and speech produced in quiet and noise in Chapter 3. Note that the signals have normalized rms energy. A clear Lombard effect of energy shift to higher frequencies relative to quiet speech is visible.*

4.2 Speech corpus collection

4.2.1 Speech material and noise backgrounds

Speakers produced sentences defined by the Grid structure used in previous collections of normal speech (Cooke et al., 2006) as well as Lombard speech (chapter 3). While Grid sentences are not representative of natural tasks, they control for differences in speaking style and syntax, and the existence of many keyword repetitions allows for cross-condition comparisons of acoustic properties. Talkers produced an identical set of 30 Grid sentences in each of the conditions (see next paragraph). To introduce some variation and remove any sentence dependency effect, each talker used a different sentence set.

Speech was collected in quiet and in the presence of 5 noise backgrounds, one full-band, two high-pass filtered and two low-pass filtered. The full-band stationary noise had a spectrum, shown in figure 4.1, equal to the long-term spectrum of utterances drawn from the 16 female and 18 male talkers of the Grid corpus. Low and high-pass noise was derived from full-band noise using Chebyshev filter implementations with 0 dB pass-band gain and 60 dB stop-band attenuation, with frequency responses illustrated in figure 4.2. To investigate the effect of the size of the stop-band on speech production in noise, narrow- and wide-band versions of both high- and low-pass noise were generated using cutoff frequencies of 1 and 2 kHz. Note that in the low-pass conditions, the 1 kHz cutoff results in a narrow-band noise while in the high-pass condition the same cutoff leads to a wide-band noise, and vice versa for the 2 kHz cutoff. All maskers were normalized to 89 dB SPL prior to presentation, as measured with a Bruel & Kjaer (B & K) type 2603 sound level meter and B & K type 4153 artificial ear.

4.2.2 Talkers

Eight native speakers of British English (4 males and 4 females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. All received a hearing test as described in section 3.2.3 of chapter 3. All the participants had normal hearing. Ages ranged from 24 to 48 years (mean: 29.8 years).

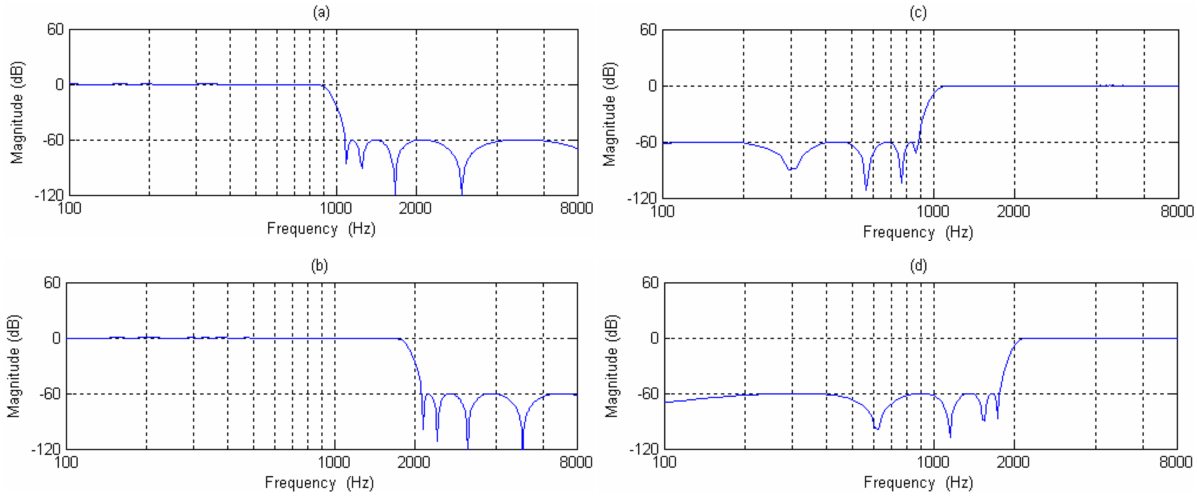


Figure 4.2: *Frequency responses of the low- and high-pass digital filters. Panel (a) and (b) represent the low-pass filters with cut-off frequency of 1 kHz and 2 kHz respectively. Panel (c) and (d) represent the high-pass filters with cut-off frequency of 1 kHz and 2 kHz respectively.*

4.2.3 Procedure

Corpus collection sessions took place in an IAC single-walled acoustically-isolated booth. Speech material was collected using a B & K type 4190 ½ inch microphone coupled with a preamplifier (B & K type 2669) placed 30 cm in front of the talker. The signal was further processed by a conditioning amplifier (B & K Nexus model 2690) prior to digitisation at 25 kHz with a Tucker-Davis Technologies (TDT) RP2.1 system. Simultaneously, maskers were presented diotically over Sennheiser HD 250 Linear II headphones using the TDT system. Talkers wore the headphones throughout, including for the quiet condition. In order to compensate for sound attenuation introduced by the closed ear headphones, the talkers' own voice was fed back via the TDT system and mixed with the noise signal prior to presentation over the headphones. At the beginning of the recording session, each talker was asked to speak freely into the microphone while wearing the headphones. The level of voice feedback was manually adjusted until the talker felt that the overall loudness level matched that when not wearing headphones. Voice feedback level was then held constant for all the recording conditions and talkers were unable to adjust the level.

Sentence collection and masker presentation was under computer control. Talkers were asked to read out sentences presented on a computer screen and had 3 seconds to produce each sentence. They were allowed to repeat the sentence if they felt it necessary, with the final repetition used for further analysis. In practice, talkers made only a few repetitions in any single condition with maximum of 4 out of 30 sentences and a mean of less than 2. Across-talker means of repetition in the 6 conditions were not statistically different (F(1,7)=0.86, *p*=0.44). Maskers were gated with the 3 seconds recording time. Condition and sentence orders within each condition were randomised. Talkers recorded all the 6 conditions (i.e. 5 noise conditions plus quiet) in one session of approximately 20 minutes.

4.2.4 Postprocessing

In order to identify and remove leading and trailing silent intervals of the collected sentences, a set of speaker-independent phoneme-level hidden Markov models (HMMs) was built from speech material in the Grid corpus using the HTK toolkit (Young et al., 1999). These models were used to produce phoneme-level transcriptions of the collected utterances via forced alignment using the HVITE tool in HTK. The leading and trailing silent intervals identified via the alignment process were removed. Transcriptions of the leading and trailing silent intervals for all the utterances were manually inspected and found to be accurate within approximately 15 msec relative to human judgements.

4.3 Acoustic measurements and statistical analysis

Four acoustic properties were estimated for each utterance. Root-mean-square (rms) energy, mean fundamental frequency (F0), spectral centre of gravity (CoG) and mean first formant (F1) frequency were computed via PRAAT v4.3.24 (Boersma and Weenink, 2005). The first three parameters were measured in the same way as described in section 3.3.1 of chapter 3. Mean F1 frequency was obtained by averaging all the F1 values estimated for voiced frames using the Burg algorithm (Burg, 1975) implemented in PRAAT. These parameters were selected since reliable changes in these properties have been reported in earlier Lombard studies, and, apart from rms energy, all these properties cue the location of spectral information, which allows the pattern of shifts in spectral energy distribution to be determined.

Across-talker means in quiet, speech-shaped noise and filtered noise conditions for each of the acoustic parameters are shown in figure 4.3. For all parameters and in both low- and high-pass conditions, noise resulted in increases in all parameters. In the low-pass case, little difference between the two filtered and full-band noises is visible, while for high-pass noise, filtered noise tended to result in smaller increases than in the full-band condition. While some variability among the individual talkers was present, similar patterns in each of the acoustic parameters and across backgrounds were observed (figure 4.4).

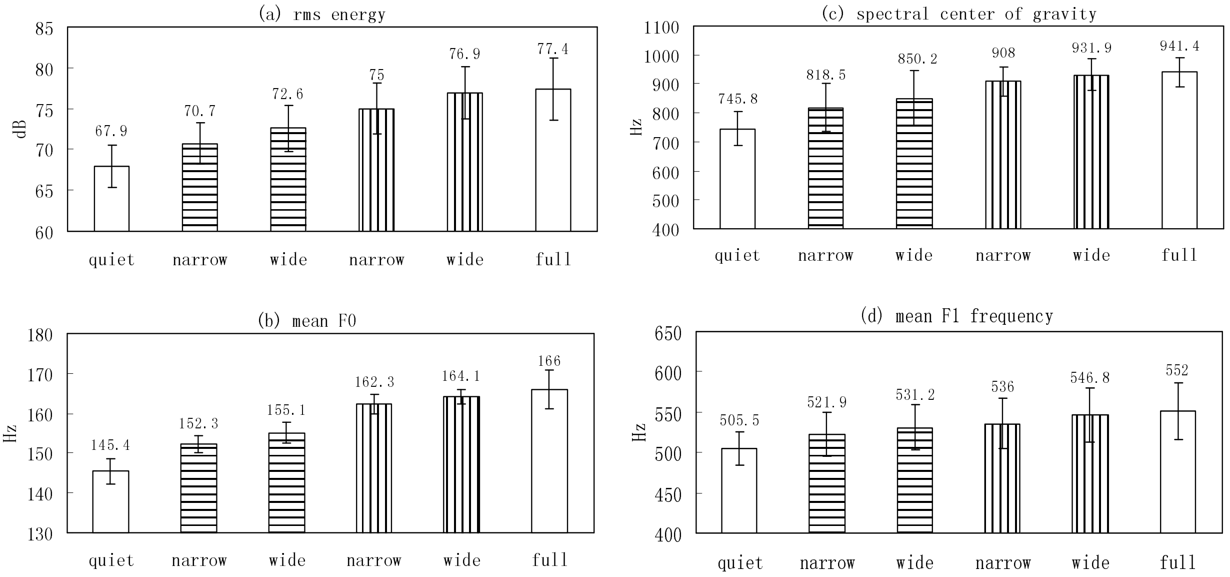
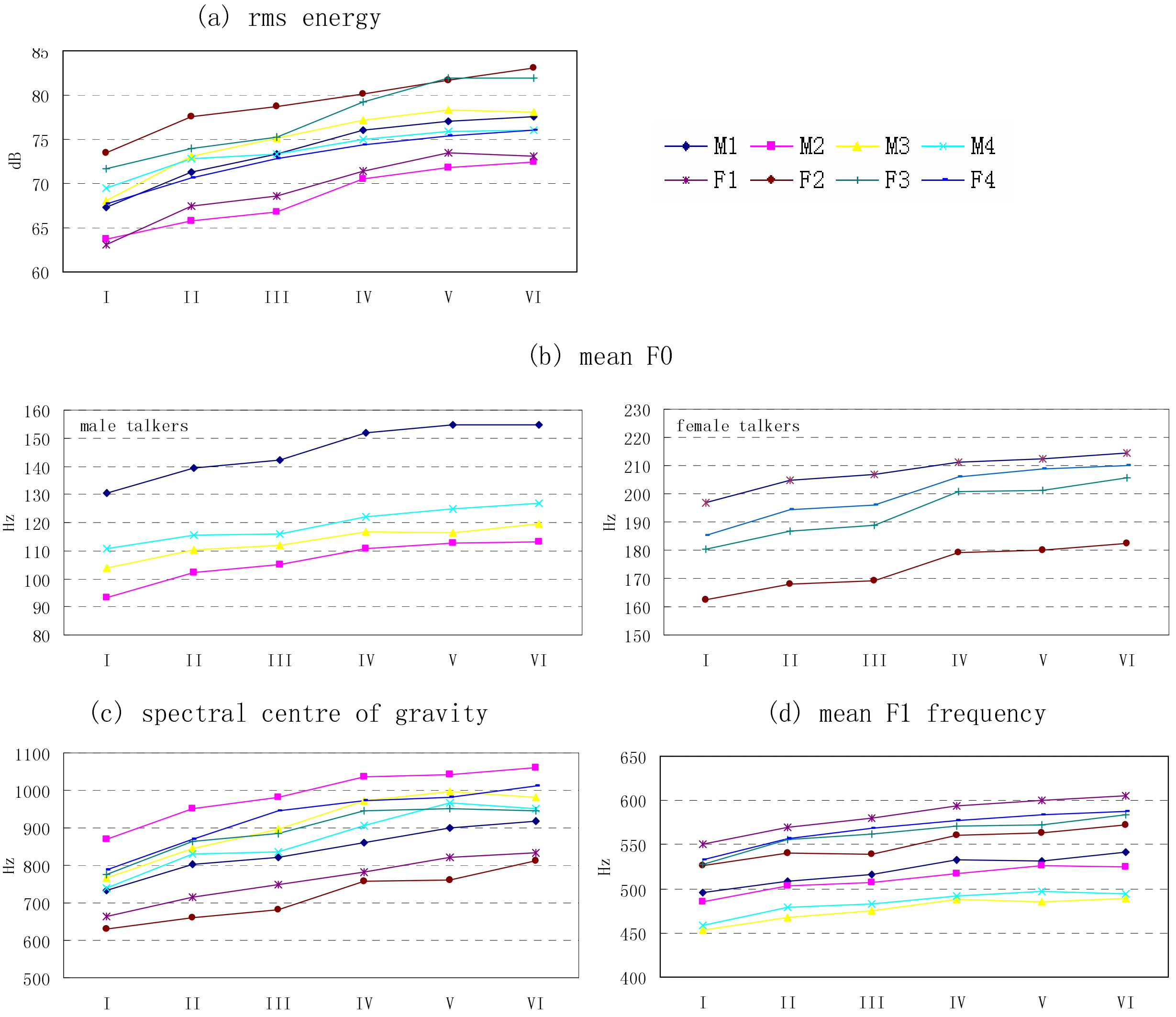


Figure 4.3: *Acoustic parameter values for quiet, two high-pass noise conditions (shaded bars with horizontal lines) with cutoff frequencies at 2 kHz ("narrow" bandwidth) and 1 kHz ("wide" bandwidth), two low-pass noise conditions (shaded bars with vertical lines) with cutoff frequencies at 1 kHz ("narrow" bandwidth) and 2 kHz ("wide" bandwidth), and speech-shaped noise condition ("full" bandwidth). Values shown are means over talkers and error bars indicate 95% confidence intervals.*


(a) rms energy

(b) mean F0

(c) spectral centre of gravity

(d) mean F1 frequency

Figure 4.4: *Acoustic parameter values for individual talkers in quiet (I) and 5 noise conditions (II. High-pass "narrow" bandwidth; III. High-pass "wide" bandwidth; IV. Low-pass "narrow" bandwidth; V. Low-pass "wide" bandwidth; VI. "Full" bandwidth). Mean F0 for male and female talkers are presented separately. Values shown are means over sentences.*

Due to the likelihood of moderate correlations between acoustic parameters such as speech level and both F0 and F1 frequency (Alku et al., 2002; Garnier 2007), multivariate analysis of variance (MANOVA) was used to examine the effect of noise background. Separate MANOVAs were computed for the low- and high-pass cases, with rms energy, F0, F1 and CoG as dependent variables. Initially, MANOVAs with one within-subject factor representing 4 types of background (quiet, narrow, wide, full) and one between-subject factor (gender) revealed that while gender differences were observed for F0 and F1, the pattern of results was the same for the male and female talkers since no significant interaction was found between gender and background type (p>0.05). In order to increase statistical power with the limited number of speakers used in the current study, data for male and female talkers were combined.

For the low-pass case, there was a significant multivariate effect of differences between the 4 backgrounds {quiet, two low-pass noise, speech-shaped noise} (F(12,47.9)=9.37, p<0.001, η^2=0.66), as well as for the four parameters individually (F(1.23,8.62)=49.15, p<0.001, η^2=0.88 for rms energy; F(1.38,9.65)=27.66, p<0.001, η^2=0.80 for mean F0; F(1.24,8.67)=21.87, p<0.01, η^2=0.76 for CoG; F(2.05,14.37)=97.64, p<0.001, η^2=0.93 for mean F1 frequency). Post-hoc pairwise comparisons (here and elsewhere in this chapter by paired *t*-tests with Bonferroni-adjustment) showed that the quiet condition was significantly different from the rest (p<0.01) for all four parameters. None of the differences between the three noise conditions was statistically significant.

As expected, given the difference between the quiet and full-band conditions, for the high-pass case, the multivariate effect of background type {quiet, two high-pass noise, speech-shaped noise} was also significant (F(12,47.9)=5.99, p<0.001, η^2=0.55). Of more interest is the confirmation by post-hoc pairwise comparisons that the high-pass conditions resulted in significant increases in all parameters relative to quiet (p<0.05), and, unlike in the low-pass case, increases were significantly smaller than the full-band condition (p<0.05) apart from the wide-band/full-band comparison for F1 (p=0.06). The tendency, visible in figure 4.3, for the wide-band high-pass noise to provoke larger parameter excursions than the narrow-band high-pass condition was not statistically significant except in the case of rms energy (p<0.05).

4.4 Discussion

The current study extends to both low- and high-pass filtered noise backgrounds the finding that talkers modify their productions when exposed to full-band noise. The low-pass conditions resulted in increases in F0 and F1 frequencies, and spectral centre of gravity. While these results are consistent with the hypothesis that speakers were actively avoiding the presence of noise whose spectrum was concentrated at low frequencies, two findings suggest otherwise. First, the full-band and low-pass filtered noise provoked statistically-identical increases in these parameters. One might expect to see a larger amount of shift in the low-pass condition to take advantage of the noise-free part of the spectrum relative to the full-band case. Second, there was no difference between the narrow- and wide-band low-pass conditions, where an active strategy would predict larger increases in the presence of wide-band low-pass noise in order to place spectral energy in the noise-free region.

High-pass filtering conditions also led to clear increases in F0, F1 and spectral centre of gravity, suggesting that speakers are unable to adopt the speaking strategy of adapting speech production to place information-bearing elements of speech in regions devoid of noise. Further, speakers reacted similarly to the wide- and narrow-band conditions, where optimality would suggest that a smaller noise-free spectral region would lead to differential shifts in acoustic parameters. The absence of the “optimal” response to high-pass noise may be attributed to articulatory side-effects of an increase in vocal effort, which was observed in all noise backgrounds. For example, the wider opening of jaw in an attempt to increase speech intensity level induces an increase in F1 frequency (Lindblom and Sundberg, 1971), and the raising of subglottal air pressure in order to produce a louder voice results in an increase for F0 (Schulman 1989; Gramming et al., 1988). The scope for active control of F0 and F1 frequencies might be limited by the stronger desire to increase output level in response to noise.

One surprising aspect of the current study is the fact that noise bandlimited to the region below 1 kHz produced an equivalent Lombard effect as full-band noise. This might result from the upward spread of masking into higher frequencies produced by the 1 kHz low-pass noise, a phenomenon first reported by Egan and Hake (1950). In addition, since all noises employed were presented at the same level, the little difference of Lombard effect between the low-pass filtered and full-band noise conditions appears to support the idea that noise level is the dominant component of the Lombard effect given that the scale of changes in acoustic parameters observed in Lombard speech has been found to be related to the relative level of the masker (Summers et al., 1988; Tufts and Frank, 2003; Patel and Schell, 2008). However, the high-pass filtered noise conditions led to a significantly smaller increase in parameters such as rms energy (2.8 and 4.7 dB compared to 7.1 and 9 dB in the low-pass conditions, a difference which probably also accounts for the lower scale of increases in other acoustic parameters given the articulatory constraints), suggesting that noise level is not the only factor in the Lombard effect. It is possible that the difference in response to high- and low-pass noise reflects the relative importance that these frequency regions have in speech perception or in own-voice monitoring. F0 information is more clearly masked in the low-pass conditions, for instance.

Overall, these findings do not support the idea of an active response to noise. However, there are several aspects of the current task which may have limited the scope or motivation on the part of talkers to exploit noise-free spectral regions. First, noise was gated on and off to coincide with the three second recording period. It is possible that speakers were not exposed to noise for long enough to learn about the potential benefit of re-allocating spectral energy. Second, the task for talkers did not involve communication of information, so the notion that talkers were motivated to make things easier for a listener is suspected. Further studies involving communicative tasks and continuous noise backgrounds may lead to different results. Finally, the observed change in speech level produced by noise may act to mask the effect of noise on other parameters. Experiments designed to inhibit the change in vocal effort (e.g. Pick et al., 1989) may provide a more sensitive measure of differential response to the spectral content of the background.

4.5 Conclusion

An effective speaking strategy for the maintenance of intelligibility in noise would be to place information in those spectral regions least affected by the noise. However, the study of this chapter found little evidence that speakers were able to modify their speech productions in this way to take advantage of noise-free regions. In the presence of high-pass noise, speech parameters such as F0 and F1 frequencies, and spectral centre of gravity did not shift downwards but instead increased relative to speaking in quiet conditions. One explanation for this result is that the increase in vocal effort caused by noise limited the scope for variability of other speech parameters such as fundamental frequency. However, there remains the possibility that under more realistic communicative conditions, speakers may adopt active strategies to reduce the effect of noise for listeners.

Chapter 5

The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise [1]

This chapter reports on a behavioral study of the relative contribution of noise-induced speech modifications in F0 and spectral tilt to the enhanced intelligibility of Lombard speech. This chapter also quantifies the perceptual effect of these two acoustic changes using a computational model based on the availability of glimpses.

5.1 Introduction

Speech intelligibility degrades in the presence of moderate and intense noise. Many studies have attempted to determine acoustic and acoustic-phonetic correlates of speech intelligibility, the discovery of which has important implications for the development of speech enhancement algorithms, particularly for listeners with hearing impairment. While factors such as an increase in speech output level can, to some extent, boost intelligibility by raising signal-to-noise ratio (SNR), level increases alone are undesirable due to their unpleasant and fatiguing effect on the

[1] A version of the work reported in this chapter appeared in Lu and Cooke (2009c).

listener. Fortunately, other acoustic and acoustic-phonetic properties have been shown to affect how well speech is understood in noise.

Hazan and Markham (2004) and Barker and Cooke (2007) reported higher intelligibility of female talkers compared to males, which might have been due to the differences in acoustic consequences resulting from the differing gender-based vocal tract characteristics. Laures and Bunton (2003) and Watson and Schlauch (2008) found a flattened fundamental frequency (F0) contour within individual utterance negatively influences sentence recognition accuracy in noise. Vowel formant space expansion (i.e. greater discrimination between vowel categories) has also been shown to benefit speech intelligibility (Bond and Moore, 1994; Ferguson and Kewley-Port, 2002). In the presence of noise, Gordon-Salant (1986) and Hazan and Simpson (1998) found that enhancement of consonant-to-vowel (C/V) amplitude ratio by 10 dB increased intelligibility by up to 10 percentage points. It has also been reported that the fine-grained acoustic-phonetic consequences of precision of articulation are able to affect speech intelligibility in noise (Bond and Moore, 1994; Hazan and Simpson, 1998). In addition, the intelligibility advantage of clear speech over normal conversational speech in the presence of noise is found to be associated with dynamic formant movement (Ferguson and Kewley-Port, 2002) and higher temporal amplitude modulation (Krause and Braida, 2004).

Further insights into the acoustic-phonetic correlates of intelligibility come from studies of Lombard speech, which has been found to be more intelligible than speech produced in quiet when both are mixed with noise at the same SNR (Dreher and O'Neill, 1957; Summers et al., 1988; Junqua, 1993; Pittman and Wiley, 2001). As reviewed in section 2.4 of chapter 2, amongst the most consistent features of Lombard speech are an overall increase in duration (although vowels and consonants are differentially affected), an increase in F0 and a flattening of spectral tilt. The scale of these changes varies with background noise level (Dreher and O'Neill, 1957; Summers et al., 1988; Tartter et al., 1993; Steeneken and Hansen, 1999).

The issue of how noise-induced speech production changes might contribute to the intelligibility advantage of Lombard speech in the presence of noise has also been addressed (Pittman and Wiley, 2001; Lu and Cooke, 2008). Pittman and Wiley (2001) suggested that the intelligibility gain of Lombard speech is likely to result from complex interactions between vocal level, spectral composition and other acoustic characteristics, rather than a simple relation between each of these parameters and intelligibility. Lu and Cooke (2008) (the study of which is also presented in chapter 3) found that Lombard speech was more intelligible than speech produced in quiet when both were mixed with stationary speech-shaped noise at -9 dB SNR. Using a model of energetic masking (Cooke, 2006), they found a strong positive correlation between speech intelligibility and the availability of spectro-temporal glimpses of the speech in the presence of noise. The intelligibility gain of Lombard speech over speech produced in quiet was thus attributed to durational increases (i.e. slow speaking rate) and more spectral energy in higher frequencies: an increase in duration provides more opportunities to glimpse acoustic information useful for phonetic distinctions and more spectral energy in higher frequencies leads to more glimpses in the presence of a speech-shaped masker (see figure 3.10).

Although an increase in the F0 of speech produced in noise has been widely reported, it is still not clear whether F0 is an attribute that affects Lombard speech intelligibility. In addition, while the study of chapter 3 suggested that the intelligibility advantage of Lombard speech over speech produced in quiet results from the increase in duration and the flattening of spectral tilt, the individual contribution of a flattened spectral tilt to the intelligibility gain of Lombard speech is unresolved. The primary purpose of the current study was to investigate the absolute and relative contributions, if any, of F0 increase and spectral tilt flattening to speech intelligibility in the presence of noise. Further, the quantitative effect of these parameters on intelligibility was studied using changes observed in Lombard speech induced by different levels of noise. The mean F0 and spectrum of speech produced in quiet were artificially manipulated either separately or together to simulate those of "natural" Lombard speech. Thus, speech intelligibility was measured as a function of parameter type and degree of manipulation. Intelligibility was also compared to that of "natural" Lombard speech to investigate the role of any secondary acoustic modifications in addition to those in F0 and spectrum (such as change in duration). Finally, in order to explore the origin of any difference in intelligibility resulting from different acoustic modifications, the current study used the glimpsing model (Cooke, 2006) as employed in chapter 3 to determine whether the resulting intelligibility difference of artificial and natural Lombard speech relative to normal speech can be explained by a change in the quantity of speech "glimpses" available in the noise.

5.2 Intelligibility of manipulated speech

5.2.1 Speech stimuli and masker

Speech stimuli produced in quiet and in the presence of noise at a number of levels were drawn from the corpus collected in the study of Chapter 3. Recall that in that study, 8 talkers were asked to read out 400 sentences in each of quiet and 3 speech-shaped noise conditions (presentation levels of 82, 89 and 96 dB SPL). Sentence structure was defined by the Grid multi-talker speech corpus (Cooke et al., 2006). Four identical sets of 100 Grid sentences, one set from each of the quiet and 3 noise conditions and balanced across the 8 talkers, were used to create the stimuli of the present study. All sentences were endpointed (i.e., leading and trailing silent intervals removed). These 4 conditions are denoted “Quiet”, “Lomb_82”, “Lomb_89” and “Lomb_96” respectively. An effect of the rise of noise level on the increase in F0 and flattening of spectral tilt is clearly demonstrated by the computations of mean F0 and spectral tilt of long-term average spectrum over the sentences in each of the 4 conditions (F0=148Hz, tilt=-1.62dB/octave for “Quiet”; F0=162Hz, tilt=-1.35dB/octave for “Lomb_82”; F0=166Hz, tilt=-1.29dB/octave for “Lomb_89”; F0=171Hz, tilt=-1.1dB/octave for “Lomb_96”). Mean F0 was obtained by averaging all the valid F0 estimates provided at 10 msec intervals using an autocorrelation-based method (Boersma, 1993). Spectral tilt was computed via a linear regression of energies at each 1/3-octave frequency.

To investigate the role of changes in mean F0 and spectral tilt on the intelligibility of Lombard speech, utterances collected in quiet were subjected to 3 types of manipulation on a sentence-by-sentence basis. To evaluate the contribution of increases in F0, each quiet sentence was artificially manipulated using a high-quality source-filter vocoder (STRAIGHT v40 [2]) to add a constant amount to the F0 across the utterance to obtain a signal having the same mean F0 as that of the corresponding Lombard sentence. Thus, corresponding to the 3 Lombard speech conditions, there were 3 sets of F0-manipulated sentences, denoted “F0_82”, “F0_89” and “F0_96”. Similarly, to examine the effect of spectral tilt flattening, each quiet sentence was

[2] STRAIGHT uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, and an excitation source design based on phase manipulation. It preserves the bilinear surface in the time-frequency region and allows for over 600% manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, without introducing the artificial timbre specific to synthetic speech signals while maintaining a high reproductive quality (Kawahara, 1997; Kawahara et al., 1999).

passed through an infinite impulse response filter of order 100 whose magnitude response was designed in such a way that the overall spectrum of the filtered signal was the same as that of the corresponding Lombard sentence, resulting in 3 sets of spectrum-manipulated sentences derived from the quiet speech (denoted “Spec_82”, “Spec_89” and “Spec_96” respectively). The filter was implemented using the MATLAB filter function “filter(*a*, *b*, *signal*)”, where *a* and *b* are vectors of LPC coefficients derived from the input quiet sentence and the corresponding Lombard sentence respectively. The output signal from the filter function then has the same overall spectrum as the Lombard sentence. Finally, to obtain stimuli having the same mean F0 *and* spectral tilt of the Lombard sentences, both F0 and spectrum manipulation were applied to each quiet sentence. F0 shift was applied before spectral manipulation. These 3 conditions are denoted “F0_Spec_82”, “F0_Spec_89” and “F0_Spec_96”.

To illustrate the processing of F0 and spectral tilt, the mean F0 and spectral tilt of a processed quiet sentence (from the condition of “F0_Spec_89”) and the corresponding Lombard sentence (from the condition of “Lomb_89”) together with the original unprocessed quiet signal (from the condition of “Quiet”) were measured as shown in figure 5.1.

In addition to the 9 manipulated speech conditions, the main experiment included 4 natural speech conditions: speech produced with no noise (“Quiet”), and speech produced in the presence of noise (“Lomb_82”, “Lomb_89” and “Lomb_96”). The quiet condition provides a baseline against which the contribution to intelligibility of the various speech manipulations can be measured, while the natural Lombard speech presumably represents a performance ceiling since it contains not only the manipulations represented in the artificial conditions but other changes, such as alterations to formant frequencies and bandwidths, some of which might conceivably contribute to intelligibility.

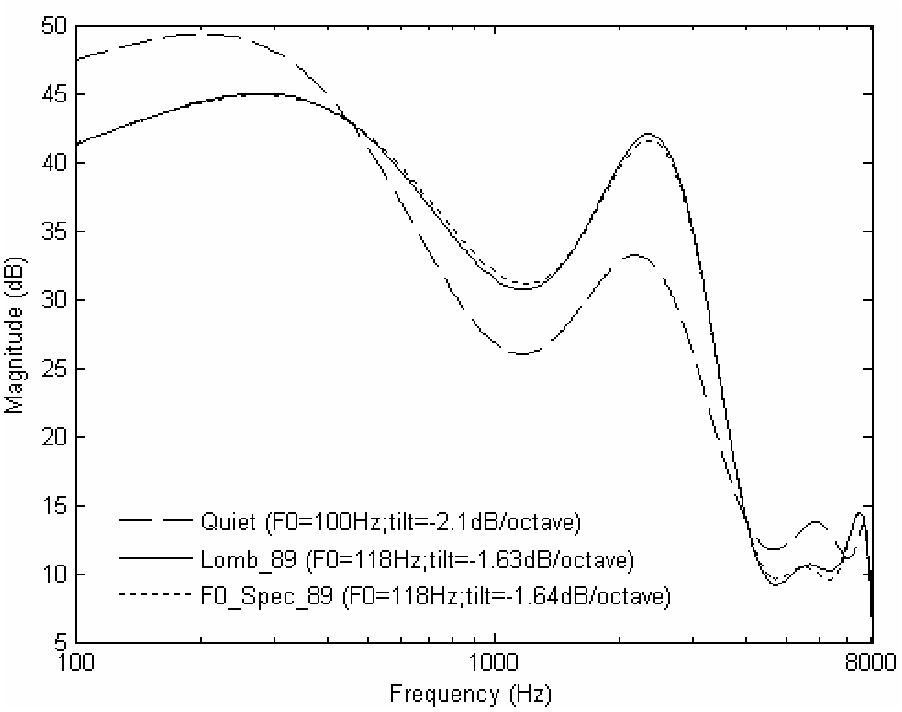


Figure 5.1: *The spectrum of a Grid sentence in the conditions of "Quiet", "Lomb_89" and "F0_Spec_89" with the values of mean F0 and spectral tilt. "Lomb_89" represents one of the Lombard conditions and "F0_Spec_89" is the condition that contains processed signals having the same mean F0 and spectral tilt of those in the condition of "Lomb_89". Signals were normalized to have equal rms energy.*

Since in the current study F0 manipulation was implemented via the tool STRAIGHT, any effect of F0 manipulation on speech intelligibility might also be accompanied by artefacts introduced by the resynthesis algorithm. To check for any such effects, an additional 3 conditions were tested in which the original stimuli from the "Quiet", "Spec_89" and "Lomb_89" conditions were re-synthesized by STRAIGHT without parameter manipulations.

In summary, the experiment contained 16 test conditions: 4 of natural speech, 9 with manipulated speech, and 3 to check any effects of the resynthesis algorithm. The same set of 100 Grid sentences was used for the 16 conditions. In all 16 conditions, each sentence was mixed with a speech-shaped noise masker at an overall SNR of -9 dB, a value chosen to avoid ceiling and floor effects as reported in the intelligibility experiment of chapter 3. The spectrum of the masker equaled the long-term average speech spectrum of the Grid corpus (see figure 3.1). A masker with a speech-shaped spectrum was chosen because it was found in chapter 3 to elicit both a strong overall Lombard effect and a flattening of the speech spectrum, which was suggested as a possible basis for intelligibility gains based on release from energetic masking in the presence of a speech-shaped noise masker. Maskers were gated on and off with the stimuli and the mixed signals were scaled to a presentation level of approximately 68 dB SPL.

5.2.2 Listeners

Ten native speakers of British English (7 males and 3 females) took part in the intelligibility experiment. All received a hearing test as described in section 3.2 of chapter 3. All had normal hearing level. Ages ranged from 20 to 31 years (mean: 26.2).

5.2.3 Procedure

Listening sessions took place in an IAC single-walled acoustically-isolated booth. Stimulus presentation and results collection was controlled by a computer program. Stimuli were presented diotically over Sennheiser HD 250 Linear II headphones. Listeners were asked to identify in each noisy utterance the letter and digit keywords by entering their results using a conventional computer keyboard. Those keys representing letters were activated immediately following the onset of each utterance. As soon as a letter key was pressed, the 10 digit keys were enabled. This approach allowed for rapid and accurate data entry. Since the structure of the speech materials provided no contextual information with which to predict the target keywords, the listeners were required to rely on the acoustic information rather than the semantic content of the sentence to identify the target words. Each participant completed the 16 conditions over 2 sessions. Each condition consisted of 100 sentences, and required 4-5 minutes to complete. For each condition, keyword identification rate was computed as the percentage of correctly identified keywords. Condition orders were randomized across listeners. There were 10 additional unscored tokens (5 in quiet and 5 in noise) for practice in the beginning of the first session for each listener.

5.2.4 Results

A. Effect of resynthesis procedure

Figure 5.2 compares speech intelligibility in the re-synthesized and original conditions of “Quiet”, “Spec_89” and “Lomb_89” to determine the effect of any artefacts which might have been introduced by STRAIGHT processing. A two-way repeated-measures ANOVA with factors of type of speech signal (re-synthesized, original) and type of manipulation (“Quiet”, “Spec_89”, “Lomb_89”) demonstrated that the effect of type of speech signal collapsed over the three conditions was not significant (F(1,9)=1.76, *p*=0.22) and none of the differences in any of the 3 manipulation conditions reached significance (*p*>0.20).

This finding supports that of Assmann and Katz (2005), who reported that when no parametric modifications were introduced, vowels synthesized with STRAIGHT were identified as accurately as the natural version. Kawahara (1998) also found the re-synthesized speech using STRAIGHT provided equivalent “naturalness” compared to the original speech, bearing out the claim (Kawahara et al., 1999) that STRAIGHT is capable of high-fidelity speech manipulation. Both subjective impressions and the results of the present listening test suggest that STRAIGHT processing in the current study was unlikely to introduce important artificial timbre or other deleterious effects when manipulating F0.

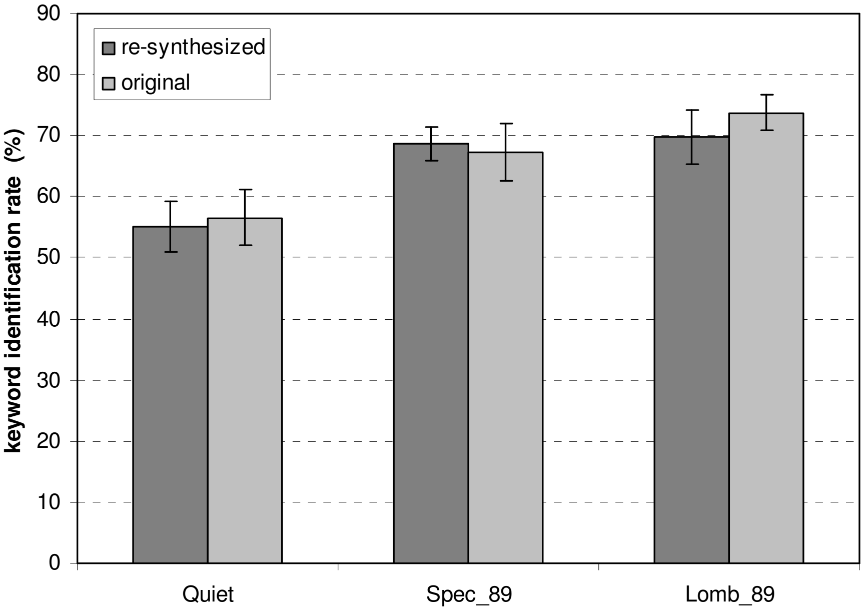


Figure 5.2: *Keyword identification rates for the re-synthesized and original speech, when added to speech-shaped noise. Values shown are means over listeners.*

B. Effects of manipulated speech on intelligibility

Figure 5.3 summarizes relative improvements in keyword identification rates in all 12 speech manipulation conditions over quiet, shown as the proportional increase in scores. The baseline performance in quiet was 56%, while intelligibility for both the manipulated and natural Lombard conditions exceeded this score, with up to 30% relative improvement. Using the same SNR and type of noisy stimuli, the baseline score for utterances produced in quiet was somewhat higher than the 42% reported in the study of chapter 3, and consequently, the average increase of Lombard speech intelligibility over the quiet was somewhat lower in the present study compared to that reported in chapter 3 (16 versus 24 percentage points). This difference may be due to the fact that 7 of the 10 listeners recruited for the current experiment had prior experience of Grid sentences in other speech perception and production experiments.

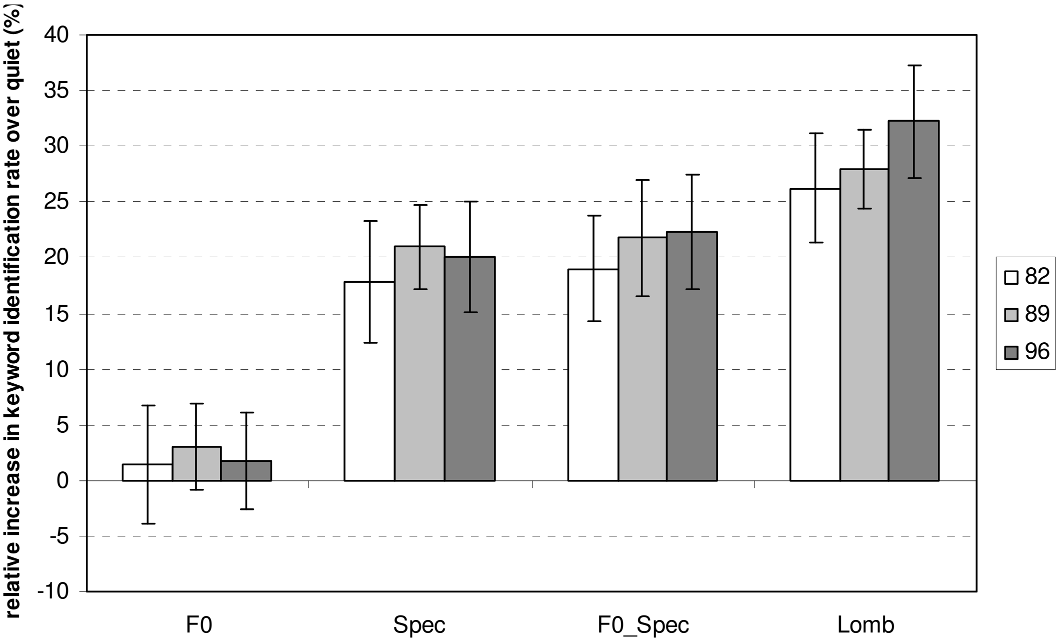


Figure 5.3: *Relative improvements in keyword identification rates for speech with acoustic manipulations over speech produced in quiet, in both cases presented in speech-shaped noise. Improvements are shown as proportional increases in scores. The baseline identification score for utterances produced in quiet was 56%.*

Paired-samples *t*-tests (with Bonferroni-adjustment) were computed between the quiet condition and each of the 12 speech manipulation conditions. Compared to quiet, the three F0-shifted speech conditions did not increase intelligibility (p>0.05) while all the other conditions did (p<0.001). For the 12 conditions, a two-way repeated-measures ANOVA with factors of manipulation type = {F0, Spec, F0_Spec, Lomb} and manipulation level = {82, 89, 96} was also computed. The analysis showed there was no significant interaction between these two factors

(F(3.50,31.53)=0.60, *p*=0.73) while demonstrating a significant main effect of manipulation type (F(2.30,20.74)=127.96, *p*<0.001, η^2=0.93) and manipulation level (F(1.19,8.32)=8.67, *p*<0.05, η^2=0.55).

Between the 4 types of manipulation collapsed across manipulation level, post-hoc pairwise comparisons (here and elsewhere in this chapter with Bonferroni-adjustment for multiple comparisons) indicated that the intelligibility of speech with a manipulated spectrum increased significantly compared to that with F0 shifted separately (*p*<0.001). There was no additional benefit of modifying F0 and spectrum together over changing the spectrum alone (*p*>0.05). Natural Lombard speech was more intelligible (*p*<0.01) than all other types of manipulation. Since manipulation type did not interact with manipulation level, a similar overall pattern was further confirmed at each of the 3 manipulation levels using pairwise comparisons between the 4 manipulation types (*p*<0.05), except that at the smallest manipulation scale (82 dB Lombard speech), the intelligibility gain of natural Lombard speech over spectrum-manipulated speech and speech with spectrum manipulated jointly with F0 failed to reach significance (*p*>0.11).

In addition, post-hoc pairwise comparisons between the 3 manipulation levels collapsed across manipulation type confirmed that there was a significant difference between the largest and smallest manipulation levels (*p*<0.05) although the 89 dB case did not differ significantly from the other two (*p*>0.27). This tendency was also observed in each of the 3 manipulation types ("Spec", "F0_Spec" and "Lomb") although none of these reached significance (*p*>0.08).

Since listeners were exposed to the same set of 100 Grid sentences across conditions, a check was made for learning effects using a repeated-measures ANOVA with factors of background condition and presentation order. This analysis suggested that condition order was not a significant factor for keyword identification score (F(1,9)=0.29, *p*=0.59).

5.2.5 Discussion

The behavioral experiment explored the extent to which an increase in F0 and a flattening of spectral tilt influence speech intelligibility in the presence of speech-shaped noise. The two findings that F0-shifted speech was no more intelligible than the baseline "quiet" speech and shifting F0 of spectrum-manipulated speech did not further improve intelligibility suggest that increases in F0 make little contribution. However, it was found that there were significant intelligibility gains of spectrum-manipulated speech over quiet speech and the gain tended to increase with manipulation scale. These findings support the claim in chapter 3 that a flattening of spectral tilt helps to improve intelligibility in the presence of speech-shaped noise.

Spectral modifications alone cannot account for the entire intelligibility increase of Lombard speech, since natural Lombard speech was significantly more intelligible than synthetic Lombard speech. Thus, part of the benefit must derive from factors other than a flattening of spectral tilt. Lombard speech has a number of other acoustic and acoustic-phonetic consequences, such as changes in consonant-to-vowel energy ratio and formant frequencies. A further difference between the natural and synthetic conditions is the durational lengthening in the former. In essence, the same amount of information is spread out over a longer interval in the natural Lombard case, leading to the possibility of a greater resistance to energetic masking. To investigate a role for durational differences, and to examine whether energetic masking can explain the superior intelligibility of spectrally-manipulated speech, the glimpsing model of speech perception in noise (Cooke, 2006) was employed.

5.3 Does manipulated speech offer more glimpsing opportunities?

5.3.1 Motivation

Cooke (2006) demonstrated that recognition of intervocalic consonants solely from those spectro-temporal regions ("glimpses") of clean speech least affected by background noise predicts listener scores across a range of conditions, ranging from competing speech, through *N*-talker babble to stationary speech-shaped noise. The glimpsing model has since been shown to make good detailed predictions of the intelligibility of individual spoken letters and different talkers in adverse conditions (Barker and Cooke, 2007). To recap, in Cooke (2006), glimpses of a signal are defined as those connected regions in an auditory-inspired spectro-temporal representation greater than a certain minimum "area" calculated from the number of spectro-temporal "pixels" and where each spectro-temporal "pixel" has a local SNR larger than a threshold. Using the same computational model, chapter 3 also reported a very high correlation between relative intelligibility gains for listeners against relative increases in the amount of information available through glimpsing (r=0.98, p<0.001). The current study tested the hypothesis that the intelligibility of speech with acoustic modifications is likewise dominated by the availability of glimpses of the speech in the presence of noise. Such glimpses could result from factors such as changes in F0, spectral tilt and duration. Same as section 3.5 of chapter 3, two glimpsing statistics, glimpse area and proportion, were measured for the signal mixtures used in the intelligibility experiment conducted in the previous section.

5.3.2 Results

Figure 5.4 depicts relative changes in the glimpse measures for each of the 12 speech manipulation conditions over speech produced in quiet, shown as percentage increases. For each of the utterances produced in quiet, there were on average 1390 spectro-temporal points meeting the glimpse criteria, leading to a glimpse proportion value of 11.4%. Only one measure is plotted for the manipulation types of "F0", "Spec" and "F0_Spec" because the speech in these conditions was derived from the quiet speech and thus had the same duration, which made the duration-dependent (glimpse area) and duration-independent (glimpse proportion) measures identical Paired-samples *t*-tests (with Bonferroni-adjustment) showed that compared to speech produced in quiet, there was no significant increase in glimpse area/proportion of F0-shifted speech (p>0.05). For all the other conditions, significant increases of glimpse area and proportion over quiet were reported (p<0.001).

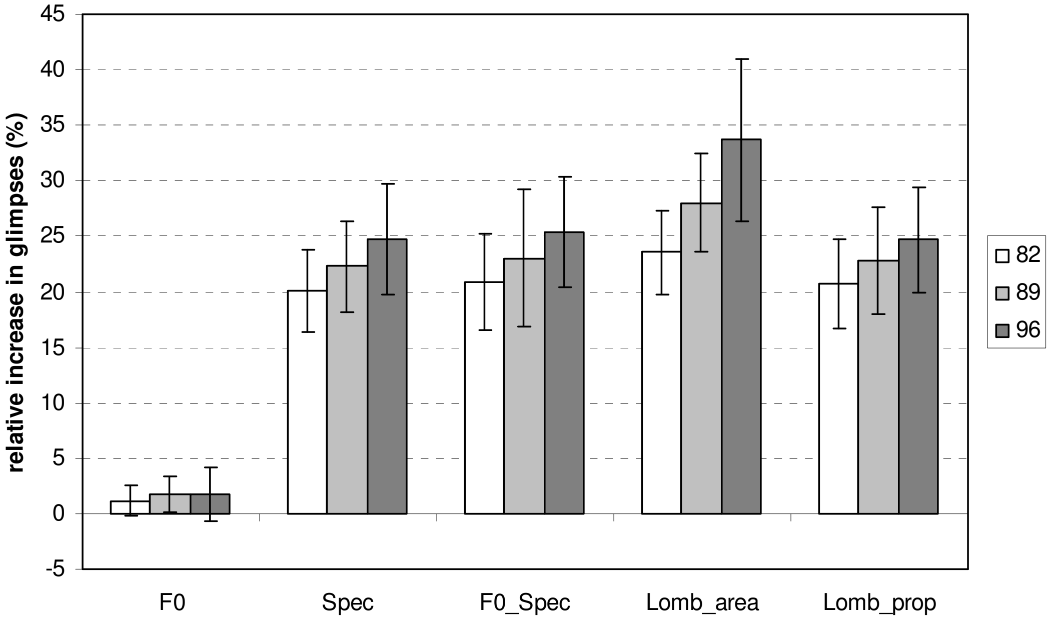


Figure 5.4: *Glimpse area and proportion for the stimuli used in the intelligibility experiment, expressed as percentage increase in area/proportion over speech produced in quiet. The baseline values in quiet for the two measures were 1390 and 11.4%. "Lomb_area" and "Lomb_prop" represent the area and proportion of glimpses measured for the Lombard speech conditions.*

For glimpse area, a two-way repeated-measures ANOVA with factors of manipulation type = {F0, Spec, F0_Spec, Lomb} and level = {82, 89, 96} demonstrated a significant main effect of manipulation type (F(2.10,14.71)=103.23, p<0.001, η^2=0.94) and the absence of an interaction with level (F(1,7)=4.01, p=0.08). To test the differences between the 4 types of manipulation collapsed across level, post-hoc pairwise comparisons showed that speech with its spectrum manipulated separately produced a larger increase than that with F0 shifted separately (p<0.001) while there was no significant change (p>0.05) between speech with spectrum manipulated separately and jointly with F0. Lombard speech produced more glimpses than all the other 3 types of manipulation (p<0.05). This pattern is echoed at each of the 3 manipulation levels as shown in figure 5.4 although for the 82 dB conditions the glimpse area for Lombard speech did not differ significantly from those for spectrum-manipulated (p=0.83) and both-parameter manipulated speech (p=0.68).

Figure 5.4 also shows that glimpse area tended to increase with manipulation level in all 4 types of manipulation apart from the conditions of F0-shift alone in which it changed little with level. This was confirmed by the significant main effect of level (F(1.19,8.32)=8.67, p<0.05, η^2=0.53). When collapsed over manipulation type, the difference between the largest and smallest level was significant (p<0.05). The tendency of glimpse area to increase with manipulation level was also observed in each of the 3 manipulation types.

Glimpse area was highly-correlated with intelligibility gain (r=0.988, p<0.001) as shown in figure 5.5 which plots relative increases in intelligibility for listeners against relative increase in glimpse area.

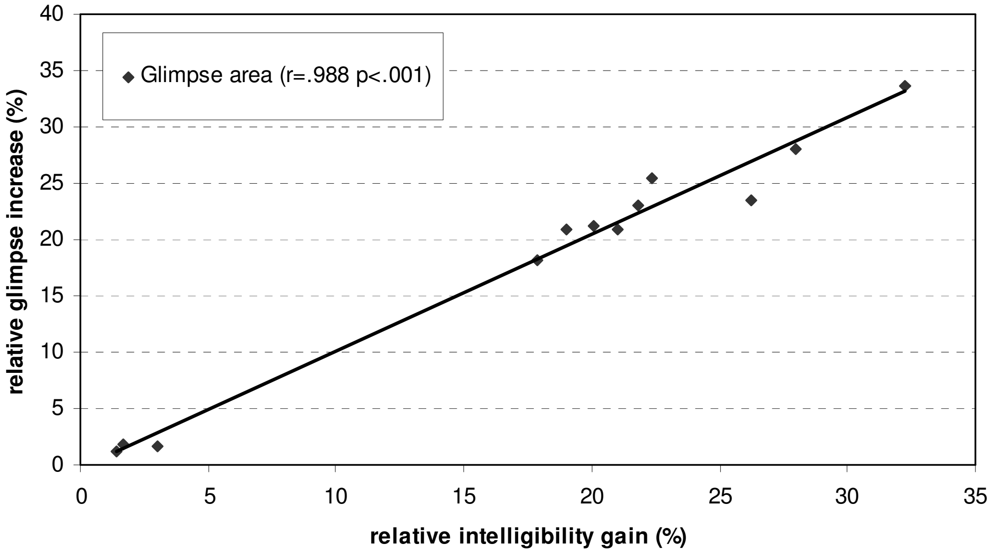


Figure 5.5: *Relation between increase in glimpse area and intelligibility, together with least-squares fit.*

For glimpse proportion, a two-way repeated-measures ANOVA with factors of manipulation type = {F0, Spec, F0_Spec, Lomb} and manipulation level = {82, 89, 96} was computed. At each level, there were significant differences (p<0.001) between manipulation type "F0" and each of the other 3 types ("Spec", "F0_Spec" and "Lomb") while the differences between "Spec", "F0_Spec" and "Lomb" were not significant (p>0.05). The increase in glimpse proportion with level for natural Lombard speech failed to reach significance (p>0.50).

Figure 5.4 also demonstrates that the relative increases of glimpse area ("Lomb_area") were larger than those of glimpse proportion ("Lomb_prop") for the natural Lombard speech conditions, a difference confirmed by a two-way repeated-measures ANOVA with factors of glimpse measure = {area, proportion} and level = {82, 89, 96}. Significantly larger increases in glimpse area over glimpse proportion were obtained (F(1,7)=15.14, p<0.01, η^2=0.68), which was further confirmed in the 89 and 96 dB (p<0.01) conditions but not at 82 dB (p=0.14). The difference in relative increase in glimpse area over proportion also tended to increase with manipulation level although the interaction between glimpse measure and level failed to reach significance (F(1,7)=5.50, *p*=0.052). The larger increases in glimpse area over proportion in the Lombard speech conditions are presumably due to the tendency of noise-induced sentences to increase in duration, since glimpse area increases in proportion to duration, while proportion is independent of duration.

5.4 General discussion

The study reported in this chapter estimated the relative contribution of F0 increase and spectral flattening to the improvement of speech intelligibility in the presence of speech-shaped noise. Compared to speech collected in quiet, an upward shift in F0 did not lead to an increase in intelligibility, while spectral flattening led to a large gain in intelligibility. However, the gain fell short of that obtained by natural Lombard speech. Such a pattern was found to be highly-correlated with a measure based on the amount of the time-frequency plane glimpsed, suggesting that the main effect of the speech manipulations examined was to create a release from energetic masking. Spectral flattening in the presence of speech-shaped noise is beneficial since it results in an upward migration of speech energy to regions less likely to be masked by speech-shaped noise (see figure 3.1). The increase in F0 led to a rather small amount of energy migration to higher frequencies compared to the speech in quiet (figure 5.6), which resulted in a small increase in glimpses and a non-significant improvement in intelligibility over the quiet speech. The presence of such an energy migration in F0-increased speech may be due to the wider spacing of harmonics. Since the Lombard speech materials used in the current study were collected in speech-shaped noise conditions drawn from the study of chapter 3, the F0 increase in Lombard speech could be a by-product of other speech changes such as an increase in vocal intensity, rather than a strategy that helps to improve intelligibility in speech-shaped noise. For other maskers (such as a competing voice) it remains possible that F0 changes could help to distinguish a speaker's output from the background.

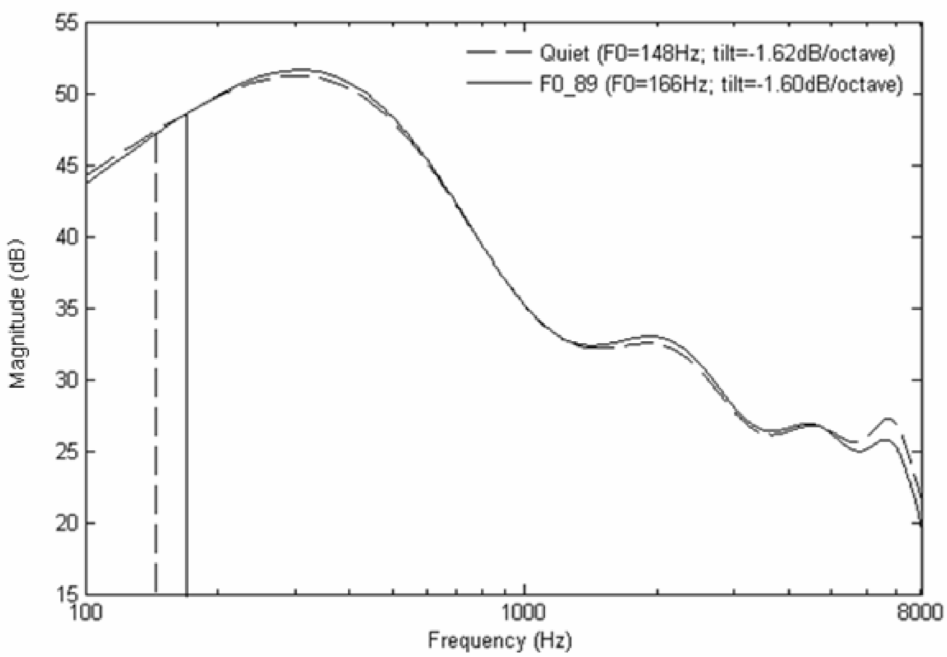


Figure 5.6: *Long-term average spectrum over sentences in "Quiet" and "F0_89" conditions. The location the mean F0 in each is represented by a vertical line. Signals were normalized to have equal rms energy.*

While a spectral flattening strategy is beneficial for noises with a falling spectrum typical of many natural noise types (e.g. multi-talker babble) used to induce Lombard speech, it is not necessarily helpful for noises with a greater energy concentration in higher frequencies. However, the study of chapter 4 demonstrated that speech produced in response to high-pass filtered noise also has a spectral centre of gravity which is shifted upwards into the frequency regions containing the noise. Talkers appear unable to adopt what might be considered the optimal strategy in such situations i.e. to shift spectral energy downwards in frequency to noise-free regions.

Evidence for the perceptual contribution of flattening spectral tilt has been mentioned in other studies. For instance, Krause and Braida (2004) found that a migration of spectral energy to high frequencies contributes to increased intelligibility of clear speech relative to conversational speech in the presence of speech-shaped noise. A significant effect on intelligibility in white noise was reported by Niederjohn and Grotelueschen (1976) who attempted to suppress the first formant by high-pass filtering to emphasize the energy in high frequencies. The intelligibility gain obtained was considered to be due to enhancement of F2 energy relative to that of F1. F2 is claimed to make a larger contribution to overall speech reception than F1 (Thomas, 1967, 1968). In addition, by moving formants upward in frequency via alteration of line spectral pairs derived from linear prediction parameters, McLoughlin and Chance (1997) reported an enhancement of vowel intelligibility in the presence of noise, which they attributed to the SNR improvement afforded by the low-frequency bias of the noise. However, Assmann et al. (2002) and Assmann and Nearey (2008) reported that an upward shift as well as a downward movement of formants due to a linear scaling of the frequency axis did not yield an improvement on the intelligibility of vowels in quiet, a finding which they attributed to the deterioration of learned relationships between formant frequencies.

The finding that Lombard speech resulted in more potential glimpses overall (as indicated by the glimpse area metric) compared to the spectrum-manipulated speech in the presence of a masker could be due to the increased duration of Lombard utterances, since the spectrum manipulation conditions were applied to utterances produced in quiet, which were shorter. When the effect of duration was normalized by measuring the proportion of the time-frequency plane glimpsed, Lombard speech led to an equivalent glimpsing density as the spectrum-manipulated speech. Given that there is a high correlation between the availability of overall glimpses and the speech intelligibility, it appears that the greater intelligibility of Lombard speech compared to spectrum-manipulated speech could result from the increase in glimpsing opportunities afforded by a slower speaking rate. This is compatible with the finding that the intelligibility gain was larger for the more intense Lombard speech, which was itself of longer duration than the less intense Lombard speech.

A number of studies have investigated the perceptual effect of duration lengthening (which is equivalent to a reduction in speaking rate if utterances of homogeneous length are used). However, evidence for the effect of durational change on speech intelligibility in noise is mixed. While several researchers (e.g. Cox et al., 1987; Jones et al., 2007) have demonstrated that slower speaking rates lead to increased speech intelligibility in noise, Sommers (1997) failed to find a perceptual correlate of speaking rate for young listeners with normal hearing. In addition, Bond and Moore (1994) and Hazan and Markham (2004) observed that words with longer duration led to an increased intelligibility in the presence of noise while no such effect of word duration was found in Uchanski et al. (2002). These findings suggest that while it is clear that duration lengthening can increase the amount of acoustic information available, the extent to which it can improve intelligibility in the presence of noise may depend on the characteristics of the listeners and speech materials employed.

The current study did not find a significant effect of increasing F0 on intelligibility, which echoes studies such as Bond and Moore (1994) and Hazan and Markham (2004), who reported that the intelligibility of speech in noise did not correlate with F0 mean. Barker and Cooke (2007) found that speech intelligibility was correlated with fundamental frequency (F0) only for female talkers at relative low SNRs. Ryalls and Lieberman (1982) and Assmann and Nearey (2008) even found a negative influence of large synthetic F0 increase on vowel intelligibility in quiet, which was attributed to the poorly resolved formant peaks that resulted from a sparsely sampled harmonic spectrum. This suggests that there could be other mechanisms apart from glimpsing that are involved in the way F0 increases affect speech intelligibility.

For the current study, the spectral energy reassignment due to spectral flattening contributed approximately 70% of the intelligibility gain, with the residual possibly due to temporal reassignment (slower speaking rate). In both cases, simple measures based on energetic masking (EM) provide a good quantitative explanation for the gains. In addition to a durational account, other non-EM factors also have a potential role for the observed residual. For instance, changes in vowel formant frequencies of Lombard speech that lead to a change in vowel space dispersion is likely to contribute since the perceptual confusion between different vowels could be reduced in an expanded vowel formant space. The improved Lombard speech intelligibility could also result from the enhancement of speech regions which contain acoustic cues to phonemic contrasts.

Various studies have attempted to improve speech intelligibility by enhancing perceptually-relevant acoustic cues, typically by identifying information-bearing regions of the signal, including those which contain important acoustic cues to phonetic contrasts, and increasing their relative intensity. Using a consonant identification task in a set of nonsense CV/VCV syllables, consonant intelligibility has been found to increase in a background of noise when their intensity relative to that of vowels was enhanced (Gordon-Salant, 1986; Hazan and Simpson, 1998; Skowronski and Harris, 2006). Hazan and Simpson (1998) reported significant improvement by applying amplitude enhancement to the formant transition regions at vowel onset and offset as well as the perceptually-important spectral regions of consonants. Tallal et al. (1996) also observed a benefit of amplifying regions of rapid spectral change to auditory training. However, these speech enhancement approaches are difficult to apply in a robust manner in real-time. The finding from the present work that speech enhancement can be realized by spectrum flattening is encouraging since it is certainly feasible to implement spectrum modifications online. Indeed, the successful application of real-time processing approach to speech enhancement in noise has been shown by Lee and Jeong (2007), for instance. By increasing the speech energy relative to noise in the frequency bands where the SNR is low, they were able to enhance speech intelligibility in noise in communication situations requiring real-time processing, such as in mobile phone applications.

5.5 Conclusion

The current chapter investigated the effects of an upward shift in F0 and a flattening of spectral tilt on speech intelligibility in noise with a speech-shaped spectrum. The results showed a significant contribution to Lombard speech intelligibility of spectrum flattening and failed to find a perceptual influence of an increase in F0. The possibility that a lengthened duration helps to improve the intelligibility of Lombard speech in noise was also suggested. Echoing one outcome of chapter 3, a high correlation between speech intelligibility and the amount of the time-frequency plane glimpsed was found. These findings suggest that speech modifications which reassign speech energy in time and frequency to introduce more glimpses in the presence of noise can be used in an attempt to improve speech intelligibility in everyday conditions.

Chapter 6

The effect of task on noise-induced speech production [1]

This chapter evaluates the influence of a communication factor on the Lombard effect induced by noise with differing degrees of energetic and informational masking potential. The possibility that speakers are able to avoid temporal overlap with a fluctuating noise masker to ameliorate any adverse effects for an interlocutor is also explored.

6.1 Introduction

The Lombard effect has been widely explored using a task of reading sentences alone (e.g. Dreher and O'Neill, 1957; Summers et al., 1988; Junqua, 1993; Pittman and Wiley, 2001) or a task with two speech partners talking to each other (e.g. Korn, 1954; Webster and Klumpp, 1962; Mixdorff et al., 2007; Patel and Schell, 2008; Bořil, 2008). Of particular interest are those studies who evaluated the effect of the presence or absence of a communicative task on noise-induced speech production modifications. In the study of Garnier (2007), individual talkers were asked to complete a non-interactive task alone and an interactive task with a speech partner. It

[1] A shorter version of the work reported in this chapter appeared in Lu and Cooke (2009b).

was also found that, compared to quiet, noise led to significant speech production changes (increases in F1, F0, speech level and duration, and more spectral energy in higher frequencies) in both tasks with and without a communication factor. The size of these Lombard effects tended to be larger in the task involving communication.

Among these studies employing tasks with a communicative element, the commonly used maskers were multi-talker babble noise and wideband stationary noise (Korn, 1954; Webster and Klumpp, 1962; Garnier, 2007; Mixdorff et al., 2007; Patel and Schell, 2008). In addition, vehicle noise and whisper were employed in Mixdorff et al. (2007) while Bořil (2008) also used bands of noise (62–125, 75–300, 220–1120, 840–2500 Hz) in spectral regions corresponding to typical energy concentrations for F0 and the lower formants. However, the effect of a single background competing talker on speech production compared to noise has not been examined when a communicative element is involved. The study reported in chapter 3 used a background talker but the task involved reading sentence prompts and was not communicative. That study found that a competing talker led to smaller acoustic-phonetic changes in F0, speech level, duration and spectral centre of gravity than those produced by stationary noise, possibly due to the small EM capacity of a competing talker at any given SNR. However, few effects of a competing talker masker above and beyond energetic masking were found. For instances, apart from a slightly larger disrupting influence of the competing speech background on the speech production process as evidenced by a larger number of false starts, and increased number and duration of short pauses, no evidence was found of speaking strategies which exploited the temporal fluctuations of specific competing utterances. These were attributed to the lack of communicative intent in the task employed.

The focus of this chapter is the influence of a communication factor on noise-induced speech production changes affected by maskers with differing degrees of energetic and informational masking potential. The two 'extreme' ends of the *N*-speaker continuum as used in the study of chapter 3, viz. a competing talker and speech-shaped noise, were employed in the current study. In addition, since a competing speaker produces both EM and IM, speech-modulated noise (SMN) backgrounds were also used. SMN is produced by modulating speech-shaped noise with the short-term temporal envelope of speech, and has approximately the same EM potential as natural speech but lacks intelligibility and thus is devoid of IM (Feston and Plomp, 1990), and its use here allows the additional IM effect of natural speech to be distinguished.

Thus, the primary purpose of the current study was to investigate the effects on speech production of the 3 types of maskers with differing degree of EM and IM compared to quiet when a communication factor was present or not. In particular, we were interested in speech production changes in word-level properties such as duration, intensity, F0 and spectral energy distribution as well as fine-grained acoustic-phonetic characteristics such as the effect of noise and task on the vowel space. Expansion or compaction of the vowel space formed by the first and second formant (F1 and F2) frequencies was examined. Vowel space dispersion is known to affect speech intelligibility (Bradlow et al., 1996) and studies by Bond et al. (1989), Garnier (2007) and Bořil (2008) hinted at the presence of differences in vowel space dispersion of speech produced in noise relative to quiet, but were not examined statistically. Bond et al. (1989) and Garnier (2007) found a compactness of vowel space (i.e. vowels cluster more tightly between vowel categories) while the tendency varied across noise-induced speech corpora (Bořil, 2008). A further aim of the present study was to investigate whether talkers could avoid overlapping with a fluctuating noise masker such as competing talker masker and speech-modulated noise by exploiting the temporal gaps in the masker when interacting with a speech partner. Such temporal modifications were not observed using a non-communicative task in the study of chapter 3.

6.2 Speech corpus collection

6.2.1 Talkers

Eight native speakers of British English (4 males and 4 females) drawn from staff and students in the Department of Computer Science at the University of Sheffield participated in the corpus collection. All received a hearing test as described in section 3.2. All had normal hearing level. These 8 speakers were grouped into 4 pairs, each of which had two speakers of the same gender.

6.2.2 Tasks and maskers

To determine the effect of task on speech production, tasks with and without a communication factor were employed. In one task, individual speakers were asked to speak aloud while solving sudoku puzzles, while in another task pairs of speakers solved these puzzles cooperatively. Sudoku puzzles naturally provoke the occurrence of spoken digits which serve as a solid basis for comparisons across conditions and speakers. The puzzles were randomly selected from the website "http://www.dailysudoku.com/sudoku/" with medium difficulty level, chosen on the basis of pilot tests which suggested that easy sudokus could lead to less communicative demand, while more difficult sudokus produce a less fluid interaction.

To investigate how speech production changes are affected by different noises, the background was quiet (Q) or contained one of three types of noise: competing speech (CS), speech-modulated noise (SMN) and stationary speech-shaped noise (SSN). In summary, talkers produced speech in a total of 8 conditions consisting of 2 independent factors, type of task and background (2 tasks × 4 backgrounds).

Each of the 8 speakers attended three recording sessions. In the first session, without noise exposure, one speaker in each pair did a 10 minute recording while solving puzzles alone. Then, the speaker cooperated with his/her partner for 10 minutes, followed by another 10 minute recording when the partner was solving alone. From this material, speech from 2 males and 2 females was selected to be used as the basis for competing speech maskers in subsequent sessions. Ten minutes of speech from each of the 4 talkers was manually transcribed using WAVESURFER v1.8.4 to identify speech/nonspeech segments and silent pauses. The transcription was carried out using a combination of inspecting spectrogram, waveform and F0 plot, and listening, here and elsewhere in this chapter. Sound types such as *uh, um, ooh*, paper-rustle, breathing, laughing, coughing, and unintelligible utterances were labeled as *nonspeech*. Silent pauses longer than 100 msec were also identified. Each *nonspeech* segment was replaced with a silence of the same duration. The resulting four signals were used as the competing speech maskers. For each competing speech masker, the corresponding speech-shaped noise was generated by filtering white noise with a filter whose spectrum equaled the long-term spectrum of the speech segments of the competing masker, and the corresponding speech-modulated noise was formed by modulating the generated speech-shaped noise with the envelope of the competing speech masker using a procedure described in Brungart (2001).

Speakers participated in two further sessions on different days in which they were asked to solve puzzles alone (session 2) and with their partners (sessions 3) in the three noise conditions, whose orders were balanced. Each recording lasted 10 minutes for each noise condition. Speakers were permitted a short break between conditions. In each noise condition of the second session, for each individual talker, the masker was the same as that used in the corresponding condition of the third session. In the third session, for each pair of speakers, the masker used in the competing speech condition contained a talker with the same gender (not the same person as either of the speakers in the pair), which is known to lead to more IM for listeners (Cooke et al., 2008; Vestergaard et al., 2009). The maskers used in the other two noise conditions were those derived from the competing masker.

In each of the three sessions, individual speakers or pairs of speakers were given a few "sudoku" puzzles and asked to keep solving alone or together up to the 10 minute time limit. In practice, most individuals or pairs worked on a single puzzle in 10 minute interval.

6.2.3 Recording setup

Corpus collection sessions took place in an IAC single-walled acoustically-isolated booth, with a table placed inside. When working together, each pair of talkers sat at two sides of the table which had a screen barrier in the middle to prevent eye contact in order to provide some acoustic isolation to reduce crosstalk as well as require the talkers to rely only on acoustic cues to decode each other's speech. Two Bruel & Kjaer (B & K) type 4190 ½ inch microphones each coupled with a preamplifier (B&K type 2669) were fixed on the screen and directed towards each talker. When seated, the distance between the talker and the nearest microphone was set at approximately 20 cm. Once adjusted, they were asked to keep quite still during the recording session.

Each talker's signal collected by the microphone towards him/her was further processed by a conditioning amplifier (B & K Nexus model 2690) prior to digitisation at 25 kHz with a Tucker-Davis Technologies (TDT) System 3 RP2.1 through an individual channel. Simultaneously, maskers were presented diotically over Sennheiser HD 250 Linear II headphones using the same TDT system at 82 dB SPL, a level selected within the range known to provide sufficient EM (Summers et al., 1988 used 80, 90 and 100 dB; Junqua, 1993 and Garnier 2007 used 85 dB; Pittman and Wiley 2001 used 80 dB) but still relatively low in order to elicit IM effects, since a very intense competing speech background could cause a release from IM during the foreground talker's speech production if listeners were able to exploit level differences to separate their own speech and that of the background (Brungart, 2001; Cooke et al., 2008). Talkers wore the headphones throughout, including for the quiet condition. When solving puzzles alone, individual speakers simply sat at one side of the table with all the setups remaining the same. Signal collection and masker presentation was under computer control. Prior to saving, signals were scaled to produce a maximum absolute value of unity to make best use of the amplitude quantisation range. Scale factors were stored to allow the normalisation process to be reversed.

In order to compensate the sound attenuation introduced by the closed ear headphones, speakers' own voices were fed back via the TDT system and mixed with the noise signal prior to presentation over the headphones. At the beginning of each recording session, each speaker was asked to talk freely to the microphone while wearing the headphones. The level of voice feedback was manually and iteratively adjusted until the talker felt the overall loudness level matched that when not wearing the headphones. Voice feedback level was then held constant for the whole recording session and talkers were unable to adjust the level. While such a procedure does not compensate for frequency-specific headphone attenuation, the study of chapter 3 found no significant differences in Lombard speech with and without headphone attenuation inversion.

6.2.4 Transcription

Recordings were manually transcribed using WAVESURFER v1.8.4 at two different levels: (1) speech/nonspeech segments and silent pauses (>100msec); (2) individual digits “one” to “nine”. The reason to choose 100 msec as the minimum time restriction for consideration as a silent pause was that pauses with very short durations could coincide with other articulatory processes such as the articulatory closure of stop consonants as well as those associated with respiration, which occur when a speaker pauses in order to breathe. In addition, since every pair of talkers were sitting in the same room during the recording of corporative task, the crosstalk from the other speaker on each microphone could affect the transcription accuracy. A crosstalk level difference of approximately 12 dB was measured by comparing the recorded signals from a single talker on the two microphones. With such a level difference, it was found that the crosstalk was not an important factor in locating the speech segment boundaries. In practice, there were less than 5% occurrences of digit words in the crosstalk, leading to on average 12.3 “clean” instances (standard deviation: 4.2) of each digit available in each condition per talker.

Given that the segment boundaries of the collected signals were transcribed by only one person who is the author of this thesis, the reliability of the manual transcription was tested by re-transcribing the collected signals for one of the 8 talkers. Intra-transcriber reliability was measured by computing the Pearson’s correlation for the time boundaries between both sets of transcriptions. Significant correlation (r=0.99, p<0.001) was reported for speech/nonspeech and silent segments and for individual digits respectively, suggesting that intra-transcriber variability was not an important factor in the transcription of the segment boundaries.

6.3 Speech production analysis

6.3.1 Word-level analysis

Four acoustic properties were estimated for each digit word. Word duration, root-mean-square (rms) energy and mean fundamental frequency (F0) were computed via PRAAT v4.3.24 (Boersma and Weenink, 2005) as described in section 3.3.1 of chapter 3. Word spectral tilt was estimated using MATLAB via a linear regression of the long-term average spectrum (0-8 kHz), expressed in dB/octave. Since the number of instances of some digits can be as few as five in one condition per talker, in each of the 8 conditions the measurement for each talker was the average across the median of the instances of each digit. Here, the reason to use median rather than mean was because the median is not influenced by the extreme values of small data set. The measurement for each condition was the mean over those of the 8 talkers as shown in figure 6.1.

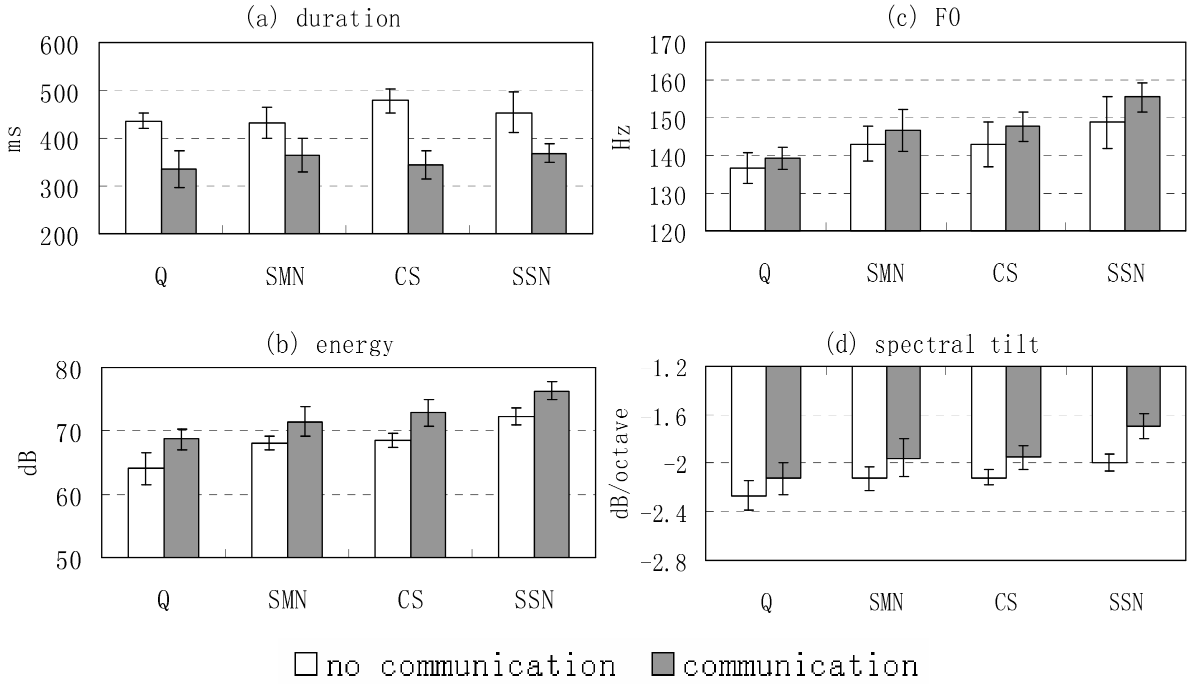


Figure 6.1: *Acoustic parameters of speech produced in the 4 types of background (quiet and 3 noisy backgrounds) with or without a communication factor. Values shown are means over talkers and error bars, here and elsewhere, indicate 95% confidence intervals. Background conditions are indicated as quiet (Q), speech-modulated noise (SMN), competing speech (CS) and stationary speech-shaped noise (SSN).*

To aid the interpretation of figure 6.1, a two-way repeated measure ANOVA with within-subjects factors of type of task and background (2 tasks × 4 backgrounds) was carried out for each of the 4 parameters. There was no significant interaction between the factors of task and background for any of the 4 parameters (F(1.47,10.27)=2.61, *p*=0.13 for duration; F(3,21)=0.24, *p*=0.87 for energy; F(3,21)=0.84, *p*=0.49 for F0; F(3,21)=1.27, *p*=0.31 for spectral tilt). In both tasks, compared to quiet, speech-shaped noise (SSN) produced largest increases and the other two maskers, SMN and CS, led to comparable but smaller changes in energy, mean F0 and spectral tilt. This was confirmed by the significant main effect of type of background (F(3,21)=17.98, p<0.001, η^2=0.72 for energy; F(3,21)=8.98, p<0.01, η^2=0.56 for F0;

$F(3,21)=7.70$, $p<0.01$, $\eta^2=0.52$ for spectral tilt). For these 3 parameters, post-hoc pairwise comparisons (here and elsewhere in this chapter by paired *t*-tests with Bonferroni-adjustment) between the 4 types of background collapsed across tasks showed that the differences between SMN and CS were not significant ($p>0.05$). Also, both of these two conditions significantly differed from quiet ($p<0.05$) and SSN ($p<0.05$). In addition, word duration was similar across the 4 types of background in both tasks. This was confirmed by the insignificant main effect of background ($F(3,21)=1.02$, $p=0.40$) as well as post-hoc pairwise comparisons between the 4 background conditions, which reported that compared to the quiet, none of the 3 masking noises led to significant durational changes in either type of task.

Figure 6.1 also shows clear differences in speech production between tasks with and without a communication factor. The main effect of task type was significant for duration ($F(1,7)=29.16$, $p<0.01$, $\eta^2=0.81$), energy ($F(1,7)=26.08$, $p<0.01$, $\eta^2=0.79$) and spectral tilt ($F(1,7)=28.57$, $p<0.01$, $\eta^2=0.80$). Such a task effect for these 3 acoustic parameters was further confirmed by post-hoc pairwise comparisons in each background condition ($p<0.05$), apart from spectral tilt in which the difference in competing speech condition was not significant ($p=0.12$). Mean F0 also tended to increase when the communication factor was present across all 4 background conditions (figure 6.1) although only the task effect in speech-shaped noise background was significant ($p<0.05$) and the main effect of task type on F0 also failed to reach significance ($F(1,7)=3.74$, $p=0.07$). However, further inspection on the task effect for individual talkers showed an increased F0 across tasks for most of the speakers and background conditions (figure 6.2). This suggested that the small number of speakers employed in the current study may have limited the chance of seeing a significant task effect for mean F0.

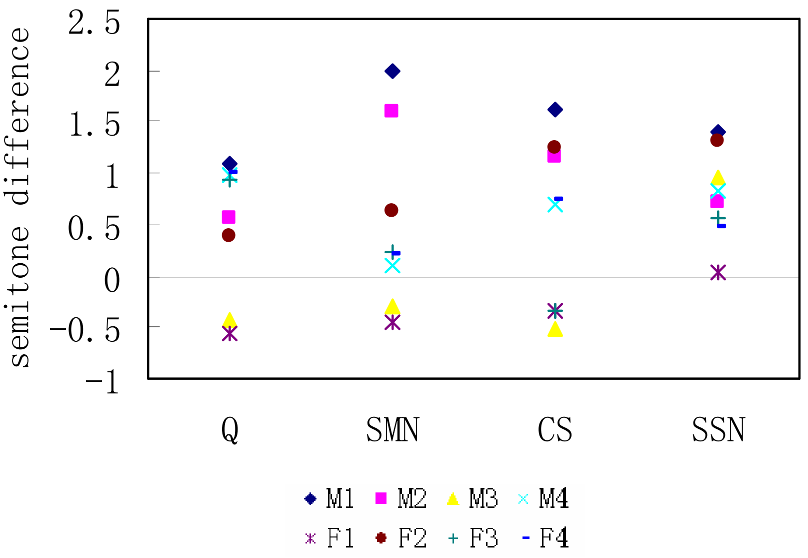


Figure 6.2: *Mean F0 differences across tasks in the 4 background conditions for individual talkers.*

6.3.2 Vowel space analysis

To examine any effect of task and the three noise types compared to quiet on expansion or compactness of the F1-F2 vowel space, F1 and F2 frequencies were estimated for the steady vowels /iː/, /ɪ/, /e/ and /uː/ in the words “three”, “six”, “seven” and “two” respectively in the 8 conditions (2 tasks × 4 backgrounds). Vowel instances were manually segmented using WAVESURFER v1.8.4. Frequencies were computed as the average of the central 3 frames in each vowel instance using the Burg algorithm (Burg, 1975) implemented in PRAAT v4.3.24 (Boersma and Weenink, 2005). F1 and F2 values were then converted into the perceptually motivated mel scale (Fant, 1973).

$$M = (1000 / \log_{10}2) \times \log_{10}((F / 1000) + 1)$$

where M and F are frequencies in mels and Hertz respectively.

To provide a single quantity indicative of vowel space expansion or compactness between vowel categories for each of the 8 conditions, a measure of each talker’s "between-category dispersion" was calculated as the mean of the Euclidean distances of each vowel token from the central point in the talker's F1-F2 space. A second measure of the compactness of individual vowel categories, "within-category dispersion", was also carried out. First, the mean of the Euclidean distances of each individual vowel token from the category mean was computed, as for the measure of between-category dispersion. Then a single measure for each talker was calculated as the mean within-category dispersion across all four vowel categories. These two measures follow the study of Bradlow et al. (1996).

Figure 6.3 shows that, compared to the task with no communication, the communicative task led to larger between-category dispersion in the conditions of quiet and speech-modulated noise, but produced similar values in the other two conditions, suggesting that the effect of communication factor on between-category dispersion differed with the type of background. This was confirmed by the significant interaction between task and background (F(1.51, 10.53)=6.71, $p<0.05$, $\eta^2=0.49$), reported by a two-way repeated measure ANOVA with within-subjects factors of type of task and background (2 tasks × 4 backgrounds). Although the main effect of task was significant (F(1,7)=18.13, $p<0.01$, $\eta^2=0.72$), post-hoc pairwise comparisons between tasks in each of the 4 background conditions showed that only the task effect in the conditions of quiet and SSN reached significance ($p<0.01$ and $p<0.001$ respectively). In addition, post-hoc pairwise comparisons between the 4 types of background reported that compared to quiet, none of the 3 masking noise led to significant change of between-category dispersion in the communicative task ($p=0.82$ for SMN; $p=0.09$ for CS; $p=0.70$ for SSN). However, for the task with no communication, compared to quiet, speech-shaped noise led to a significant increase in between-category dispersion ($p<0.05$), although the increases in the other two masker conditions failed to reach significance (p=0.73 for SMN; p=0.14 for CS).

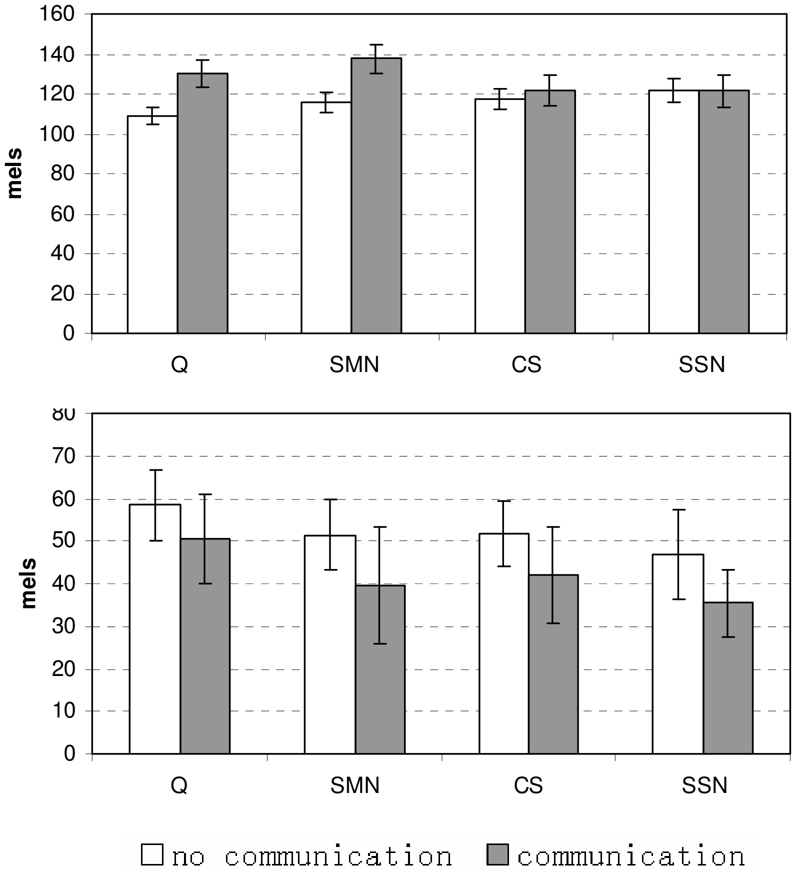


Figure 6.3: *Between-category dispersion (top) and within-category dispersion (bottom) for speech produced in the 4 types of background (quiet and 3 noisy backgrounds) with or without a communication factor. Values shown are averages across the 8 talkers.*

Figure 6.3 also shows the within-category dispersion measure. A two-way repeated measure ANOVA with within-subjects factors of type of task and background reported that there was no significant interaction between task and background (F(3,21)=0.05, p=0.91). For both tasks, there was a similar tendency that compared to quiet, all 3 types of masker led to a decrease in within-category dispersion (i.e. vowel tokens clustered more tightly within each vowel category), with the largest fall in the SSN condition. This was confirmed by the significant main effect of background (F(3,21)=9.13, p<0.05, η^2=0.62). In addition, within-category dispersion was significantly smaller in the communicative task as reported by the main effect of task type (F(1,7)=10.35, p<0.05, η^2=0.59) although post-hoc comparisons between tasks in each of the 4 background conditions showed that only the difference between tasks in speech-shaped noise condition reached significance (p<0.05).

6.3.3 Temporal effects

A. Foreground-background overlap

Here, the issue of whether talkers could avoid overlapping in time with a noise background was studied by measuring the length of temporal overlap between speech activity in the foreground talker and speech or speechlike (in the case of SMN) activity in the background masker. The overlap values were computed relative to the length of speech from the foreground, expressed as overlap percent, in order to normalize for differences in the amount of speech produced across conditions. For each talker, the overlap was computed between the *foreground* speech segments produced in the backgrounds with temporal fluctuations (i.e. competing speech "CS" or speech-modulated noise "SMN") and the *background* in which the speech was collected, shown as "CS" and "SMN" in figure 6.4. As a reference, for each talker, the overlap between speech segments produced in quiet and the background used in the fluctuating masker case was also computed, shown as "Q" in figure 6.4. If talkers were attempting to make use of the gaps in the fluctuating background, one would expect to see a smaller degree of overlap relative to the reference case.

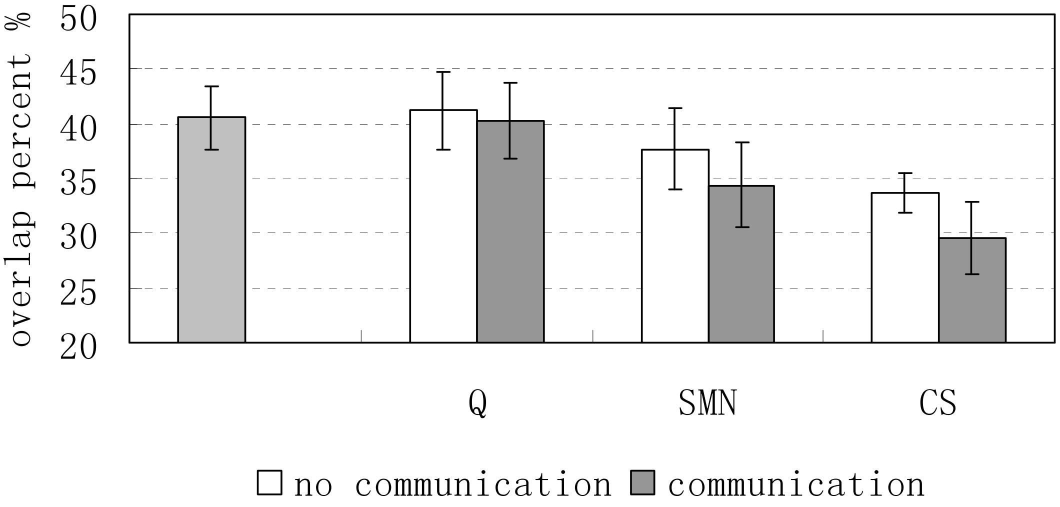


Figure 6.4: *Overlap percent as a function of task and background. Values shown are averages across the 8 talkers. The leftmost bar shows the degree of overlap for a simulated talker (see section 6.3.3.D).*

Compared to quiet, both tasks produced a reduction in overlap in the conditions of speech modulated noise and competing speech maskers, with a greater reduction for the latter. Such a difference between backgrounds was significant as confirmed by the main effect of background (F(2,14)=44.82, p<0.001, η^2=0.87) tested using a two-way repeated measure ANOVA with within-subjects factors of type of task and background (2 tasks × 3 backgrounds). Further, for both tasks, the differences between individual conditions of quiet and SMN, and between SMN and CS were also significant (p<0.01). In addition, the type of background did not interact significantly with task type (F(2,14)=2.66, p=0.11). However, post-hoc pairwise comparisons reported that compared to the task with no communication, the communicative task led to a significantly smaller overlap percent in the backgrounds of SMN (p<0.05) and CS (p<0.01) while produced statistically the same value in quiet (p=0.25). The task effects in SMN and CS conditions also resulted in a significant main effect of task type (F(1,7)=110.39, p<0.001, η^2=0.94).

The same value in the quiet condition was not surprising because the speaker did not need to make any responses to a background without noise. There are a number of ways in which speakers could reduce foreground-background overlap in the conditions of CS and SMN relative to quiet. It is possible that talking more rapidly or changing pause length distribution might result in overlap reduction *without any active attempt* to time contributions relative to the background. Subsequent analyses addressed these issues.

B. Speaking rate

The mean speaking rate in each condition and for each talker was estimated using the digits extracted during corpus transcription. To accommodate the different numbers of digit exemplars in each condition, a certain number n_i of each of the digits i = 1..9 (different for each digit but fixed across conditions) was chosen and speaking rate $rate_c$ for condition c was computed according to:

$$rate_c = \frac{\sum_{i=1}^{9} n_i}{\sum_{i=1}^{9} \sum_{k=1}^{n_i} d_{cik}}$$

where d_{cik} is the duration of the kth exemplar of digit i in condition c. Figure 6.5 shows across-talker means of speaking rate for the 6 conditions. A clear increase in speaking rate for the communicative task relative to the task without communication was observed (F(1,7)=28.44, p<0.01, η^2=0.80). Such a task effect was present in all 3 background conditions (p<0.05). While the difference in speaking rate across tasks as shown in figure 6.5 might at first sight be considered as a contributory factor given the task differences in SMN and CS conditions in figure 6.4, this is unlikely since in the quiet condition there was no task effect on overlap yet the task produced a significantly faster speaking rate. Further, the effect of noise background was not significant (F(2,14)=0.54, p=0.59) and none of the speaking rate differences between background conditions reached significance (p>0.05). These suggested that speaking rate changes can not account for the overlap reduction either as a function of task type or noise background.

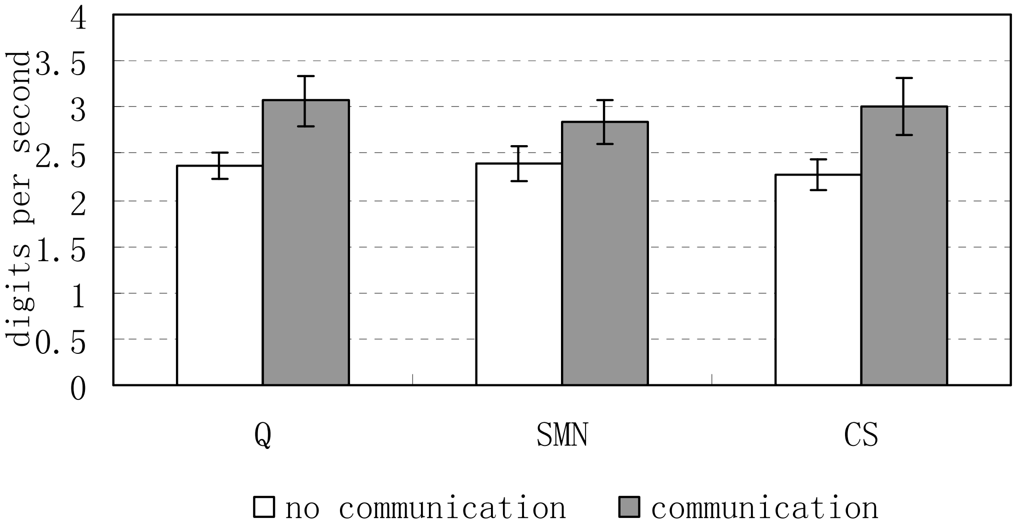


Figure 6.5: *Speaking rate as a function of task and background.*

C. Mean pause duration

Another factor which could lead to reduced overlap is a change in pause structure as a function of the background or task. Mean pause durations (figure 6.6) do indeed show both task and background effects. The communicative task resulted in longer pauses overall (F(1,7)=9.70, p<0.05, η^2=0.58), although not in quiet. Both tasks showed longer pauses in the modulated noise conditions. For the communicative task, this trend was statistically significant (F(1.98,13.88)=9.04, p<0.01, η^2=0.56). Comparison of figure 6.6 and figure 6.4 reveals a common pattern. Longer pause durations correlate strongly with decreasing amounts of overlap (r=-0.90, p<0.05). This finding is consistent with the idea that speakers wait until an appropriate point to make their contributions in the face of a modulated background. However, it is also possible that the mere presence of noise results in longer pauses. The rightmost bars of figure 6.6 suggest otherwise. The mean pause duration for stationary noise is barely different from quiet (p>0.05). Speech-shaped noise produces the largest Lombard effects (figure 6.1) but has little effect on pause duration.

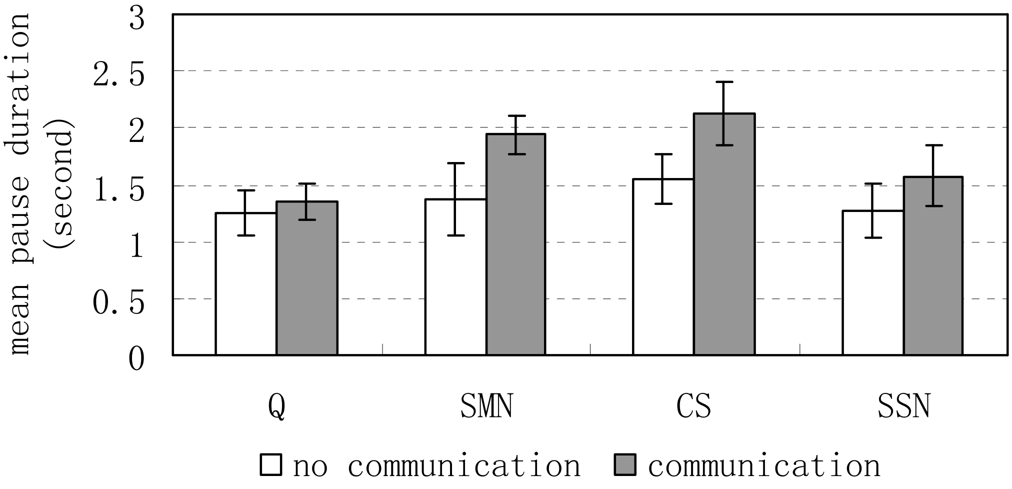


Figure 6.6: *Mean pause durations.*

D. Simulated talkers

There remains the possibility that the pause distribution varies as a function of the background (e.g. speakers matching their rhythm to that of a competing talker) without necessarily requiring *active* timing of contributions to avoid overlap. To test this idea, a simulated talker having the same distribution of pause and contribution lengths as the real talkers was constructed.

Example distributions of pause and contribution lengths for a single talker in quiet and competing speaker backgrounds are shown in figure 6.7. To accommodate the long one-sided tail, gamma distributions with density given by

$$f(x;\alpha;\beta)=\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

parameterized by α (“shape”) and β^{-1} (“rate”) were fitted to each pause and contribution distribution. A talker’s pause structure in each condition was then simulated by alternately sampling from the gamma distributions for pauses and contributions to produce a sequence of the same length as the real speaker data. One-hundred simulation sequences were produced for each condition.

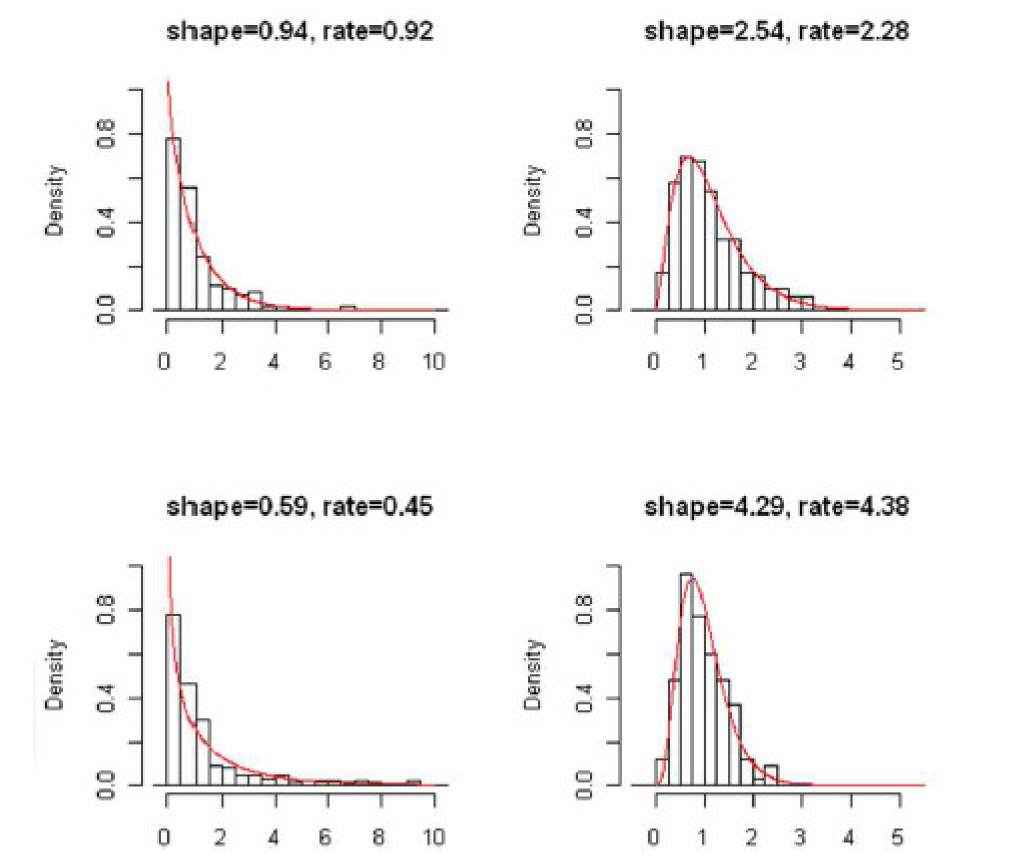


Figure 6.7: *Pause length (left) and contribution length (right) densities for a single talker in quiet (top) and a competing talker background (bottom). Horizontal axis is duration in seconds. Gamma fits are also plotted with shape and rate values shown.*

The overlap rates for these *simulated* talkers were statistically-identical across tasks and noise backgrounds. The degree of overlap for the simulated talkers is plotted in figure 6.4 (see the *leftmost* bar) and matches very closely the real talker data in the quiet condition. An additional simulation was performed by randomizing the order of consecutive pause-contribution pairs from the original data. Again, overlap scores (40%) similar to those in quiet were obtained. These simulations demonstrate that random sampling from the different pause and contribution duration distributions cannot account for the differences in overlap rate across the tasks and backgrounds.

6.4 Discussion

The current findings of increases in speech level, F0 and a flattening of spectral tilt (i.e. more spectral energy in higher frequencies) for speech produced in stationary noise compared to quiet for both tasks with and without a communication factor confirm those found in previous studies. Here, these effects were also observed in competing talker and speech-modulated noise conditions, both of which tended to produce approximately the same size of effect, which in turn was smaller than that produced by stationary noise. Such a tendency is compatible with the hypothesis of chapter 3 that the size of speech production changes scales with the energetic masking (EM) potential of background noise. In response to noise, an increase in speech level can benefit speech intelligibility due to an increase in signal-to-noise ratio, as well as the flattening of spectral slope which enables more of the speech to escape masking, at least for the maskers used here which had a low-frequency bias. On the other hand, increases in F0 might be correlated with a change in speech level as a result of the raised subglottal pressure in order to produce a louder voice (Schulman, 1985; Gramming et al., 1988; Stevens, 2000), and have been found to contribute little to speech intelligibility in noise (Bond and Moore, 1994; Barker and Cooke, 2007) as well as in quiet (Bradlow et al., 1996; Assmann et al., 2002). These findings extend the results of chapter 3 using read sentences to a task involving communication.

A very significant effect of communication on speech was observed throughout the present study in quiet and noisy backgrounds. Specifically, as shown in figure 6.1, the communication factor led to increases in speech level and F0, a flattening of spectral tilt and shorter word duration. The increase of speech level and F0 and the shift of spectral energy towards higher frequencies produced by the communication factor in the presence of noise confirm the findings of Garnier (2007), although

Junqua (1998) observed a decrease in speech level when speakers were in communication with a speech recognition device compared to reading word lists, which Bořil (2008) attributed to speakers consciously lowering their voices to obtain accurate results from the recognition system. The shorter word duration due to a communicative intent was also found in Junqua et al., (1998). In the current study, even in the absence of background noise, it may be that speakers imagine that an increase in speech level can help create more intelligible speech at the ears of the speech partner, while the observed flattening of spectral tilt places more speech energy in regions above 1 kHz known to be important for speech perception (French and Steinberg, 1947; Black, 1959; Schum et al., 1991; Studebaker and Sherbecoe, 1991). The reduction in word length by up to approximately 25% due to a communicative intent could result from the speaker's effort to maintain a more fluid interaction with their partner. Intelligibility does not necessarily degrade with moderate increases in speech rate (Bradlow et al., 1996; Uchanski et al., 2002).

In the present study, there was no *additional* effect of communication on the size of the noise-induced speech production changes in speech level, F0 and spectral tilt. This is at odds with some previous studies which reported a larger speech production change from quiet to noisy condition due to the effect of communication. Garnier (2007) observed a larger shift of spectral energy towards higher frequencies as well as greater increases in speech level and F0 from quiet to noisy conditions when subjects were interacting with a speech partner compared to while talking alone. A greater noise-induced effect in speech level when a communication factor is present is also reflected in the steeper slope of linear regression of vocal intensity as a function of ambient noise level for communicational speech (0.5, Webster and Klump, 1962; 0.39, Gardner, 1964; 0.29-0.61, Gardner, 1966) compared to read speech (0.11, Dreher and

O'Neill, 1958; 0.12, Lane et al., 1970; 0.15, Egan, 1972). Nevertheless, other findings support the current results. For instances, comparing the size of Lombard effects between the tasks with and without a communication factor, Junqua et al. (1998) showed similar size of F0 increase, although no statistical tests were performed. By asking speakers to read word/sentence lists, Kryter (1946) and Pickett (1958) also reported slopes of the voice-noise level function at 0.33 and 0.40 respectively, which are similar as 0.39 (Gardner, 1964) or 0.29-0.61 (Gardner, 1966) involving communication. One possible explanation of the difference between the current pattern and that in Garnier (2007) could be the higher baseline in quiet for speech produced in the communicative task. Without noise exposure, Garnier (2007) found that the communication factor led to small speech production changes while in the current study, significant changes in speech level and spectral tilt were reported in the communicative task compared to the task with no communication. Since a very forceful vocal effort may degrade the speech intelligibility due to the distortion of the normal speech production (Pickett, 1956; Rostolland, 1985), another possibility of the current finding that the communication factor did not yield a larger size of the Lombard effect might be the presence of the ceiling effect especially in the speech-shaped noise condition, the most adverse one, when talking to a speech partner.

Vowel space expansion (i.e. greater discrimination between vowel categories) has been associated with an intelligibility advantage on the basis of intertalker differences in overall intelligibility within normal, conversational speech (Byrd, 1994; Bond and Moore, 1994; Bradlow et al., 1996; Hazan and Markham, 2004) as well as on the basis of clear versus conversational style comparisons (Picheny et al., 1986; Moon and Lindblom, 1994; Bradlow et al., 2003; Smiljanić and Bradlow, 2005). The current study found that communication led to an expansion of vowel space between categories in quiet and speech-modulated noise conditions. In the task with no communication, the between-category vowel space of noise-induced speech also tended to expand relative to the speech produced in quiet. These tendencies of vowel space expansion, which could lead to greater ease in the discrimination of vowels, might result from a speaker's articulatory attempt to reduce the potential of perceptual confusion between different vowels when a communicative intent as well as a masking noise was present. The same tendency of Lombard speech as the current study was reported in Mixdorff et al. (2006) and Bořil (2008). However, the expanded between-category vowel space of noise-induced speech found is at odds with the findings of Bond et al. (1989) and Garnier (2007), who reported that the vowel space of speech produced in noise tended to be more compact compared to that produced in quiet. In addition to the change in overall vowel space dispersion, the current study also showed that vowels tended to cluster more tightly within each category under a communicative load and in the presence of noise. This tendency of within-category vowel clustering due to the speaker's more precise articulation of each vowel, also found in studies of clear speech (Chen, 1980), could benefit vowel discrimination because the more tightly clustered categories are less likely to lead to inter-category confusion, although Bradlow et al. (1996) showed that tightness of within-category clustering may not be a good correlate of perceptual performance.

Another important finding of the current results, in addition to these acoustic-phonetic ones, was that speakers attempt to avoid overlapping with fluctuating noise backgrounds in tasks with and without a communication factor. The reduction in overlap could not be accounted for by "passive" factors such as speaking rate changes or simulated talkers with identical pause distributions as natural talkers.

The reduction was greater for competing speech than for speech modulated noise, and greater for the communicative task.

Avoidance of temporal overlapping of foreground speech with competing talker masker or speech-modulated noise leads to a release from EM of the background due to fewer foreground speech signal elements being obscured by the background noise, aiding segregation of foreground and background speech for the interlocutor. The additional overlap reduction produced by the competing talker background relative to the speech-modulated noise may also result in reduced IM due to improved foreground-background segregation (Kidd et al., 1994). The perceptual mechanisms which drive the reduction in overlap are unclear. One possibility is that intelligibility of the competing speech masker relative to the speech-modulated noise allows a better prediction of upcoming pauses. This strategy is supported by the data of figures 6.6 and 6.7: for the competing talker background, there is evidence that the increased mean pause duration is largely due to a greater number of long pauses, perhaps due to speakers' monitoring the background for a suitable place to interject.

Interestingly, the present study found a reduction in foreground-background speech overlap when a competing speech masker was present not only in the communicative task but also in the task with no communication, while using a non-communicative task, the study of chapter 3 did not find such a tendency. This could result from the use of different competing speech material in the two studies. While the current work used a long section of spontaneous speech material, chapter 3 employed utterances of 3 seconds duration with almost all short pauses less than 100 ms. In the latter case, talkers may have been less able to attend to and track competing speech material sufficiently rapidly to modify their own productions in response.

While the current results showed the possible presence of a temporal-domain strategy to yield a release from energetic and informational masking, there are other mechanisms open to speakers. For example, differences in speech level or F0 between foreground and background are known to reduce IM (Bird and Darwin, 1998; Brungart, 2001; Vestergaard et al., 2009). In the study of this chapter, observed changes in speech level and F0 in competing speech condition appeared to be governed primarily by EM factors since speech-modulated noise induced very similar speech level and F0 changes. It may be that temporal domain speech manipulation is an efficient form of talker behavior compared to manipulations of vocal level and F0: increasing speech level is energy consuming and the extent to which talkers can manipulate F0 is constrained by physiological and articulatory constraints.

6.5 Conclusion

The study reported in this chapter investigated the effect of the presence or absence of a communicative task on speech modifications produced by noise maskers whose degrees of EM and IM differ. Acoustic changes such as an increased speech level and F0, a flattened spectral tilt and a clustering of within-category vowel space were found in the noisy conditions as well as in the communicative task. For both tasks, the size of these changes scaled with the energetic masking potential of the background, extending the finding in chapter 3 to a communicative task. In addition, an active overlap avoidance strategy in the backgrounds with temporal fluctuations was found. Overall, these findings suggest that when exposed to noise, talkers adopt a "listening-while-speaking" strategy which helps to increase the probability of message reception at the ears of the interlocutor. Most of the benefit arises from a reduction in EM, by both spectral and temporal reallocation of speech energy to frequency regions and time intervals where it is least likely to be masked.

Chapter 7

Thesis review, implications and future work

7.1 Review of the thesis

This thesis reports on an investigation into the effect of noise on speech production, namely the Lombard effect, and its perceptual consequences, using both behavioral and computational modeling studies. While authors such as Lombard (1911) and Pick et al. (1989) have suggested an unconscious and reflexive interpretation of the Lombard effect, others are in favor of the idea that the Lombard effect involves a conscious component driven by the need to maintain intelligible communication. The primary goal of this thesis has been to study the origin of the noise-induced speech modifications and in turn, how these changes affect speech intelligibility in noise.

Chapter 3 examined the acoustic and phonetic consequences of *N*-talker noise on sentence production for a range of *N* values from 1 (competing talker) to “infinity” (speech-shaped noise). Results of the noise-induced speech changes confirm those found in previous studies using stationary noise and extend to a single talker background. More interestingly, the results demonstrated a dependency of the scale of acoustic modifications on the overall energetic masking effect produced by the background signal. In addition, by comparing the intelligibility of Lombard speech with that of speech produced in quiet, it was found that the largest intelligibility gain of Lombard speech in noise results from the speech production with greatest acoustic modifications. The increased intelligibility of noise-induced speech was further found to be related to the quantity of the spectro-temporal plane “glimpsed” (Cooke, 2006). Given these findings, one interpretation of the speech modifications in the presence of noise, beyond the pure Lombard “reflex”, was proposed in chapter 3. Specifically, it was suggested that speakers may make an effort to ameliorate the EM effect of noise at the ears of listeners, leading to speech production modifications such as a shift of spectral energy to high frequencies and a lengthening of duration, increasing the opportunities for speech to be glimpsed in the presence of noise and thus yield a release from masking. However, speaking strategies that utilize the temporal fluctuations of specific competing sentences were not observed, which might have resulted from the absence of a communicative element in the task employed.

One of the questions arising from the study reported in chapter 3 is to what extent the shift of spectral energy is due to a speaker’s attempt to place spectral information in those spectral regions least affected by the noise. The effect found in chapter 3 may be coincidentally in the right direction to be advantageous for the masker types used, which had a low-frequency energy bias. To address this issue, chapter 4 measured a selection of spectral properties such as F0, F1 frequency and spectral centre of gravity for read speech produced in conditions of full-band as well as low- and high-pass filtered stationary noise whose noise energy is concentrated in different spectral regions. Results showed little evidence that speakers were able to adopt production strategies in noise which optimize listeners’ information reception. This could be due to the speakers’ desire to increase vocal level in response to noise, limiting the scope for active control of spectral properties such as F0 and F1 frequency. In addition, since the task employed in that study involved only read speech, the lack of a communicative intent might have lessened speakers' motivation to reduce the effect of noise for (non-existent) listeners.

The cause of the enhanced intelligibility of Lombard speech collected in chapter 3 was further analyzed behaviorally and quantitatively in chapter 5 by measuring the relative contribution of acoustic changes in F0 and spectral tilt, the Lombard effects most reliably observed, to speech intelligibility in noise. The roles of F0 and spectral tilt were assessed by measuring the intelligibility gain of non-Lombard speech whose mean F0 and spectrum were manipulated, both independently and in concert, to match those of natural Lombard speech. In the presence of noise with a falling spectrum, typical of many natural noise types, the contribution to the increased intelligibility of Lombard speech was large for a flattening of spectrum while little for an increase in F0. Computational modeling based on glimpses echoed the findings of chapter 3, and found that those speech modifications which reallocate speech energy in time and frequency to introduce more glimpses in noise are able to contribute significantly to speech intelligibility.

To test the notion that the presence of a need to communicate with a speech partner might lead to a differential effect of noise on speech production, chapter 6 evaluated the effect of communication on noise-induced speech modifications. Changes in speech level, F0 and spectral tilt extended the hypotheses of chapter 3 that the size of speech production scales with the EM capacity of background signal to a communicative task. Although no *additional* effect of communication on the size of the Lombard effect was found, evidence that speech production was affected by a demand of communication were observed in quiet as well as noisy backgrounds. Chapter 6 found that speakers are able to adjust the timing of their utterances to take advantages of temporal fluctuations in the background, reducing the adverse effects of the masker for an interlocutor. The findings of chapter 6 collectively suggest that talkers adopt a "listening-while-speaking" strategy of speech production which helps to benefit effective communication.

In summary, the main novel contributions of the work presented in this thesis are as follows:

- The effect of a competing talker background on speech production was measured for a number of spectral and temporal speech properties (chapters 3 and 6).
- An explanation of the cause of the Lombard effect was proposed on the basis of masking release (chapter 3).
- The inability of speakers to shift speech energy downwards to a region devoid of masker energy was reported, suggesting that speakers do not adopt optimal speaking strategies in noise (chapter 4).
- The contribution of different types and scales of noise-induced acoustic modifications to the increased intelligibility of Lombard speech was quantified based on the availability of glimpses (chapter 5).
- The impact of a communication factor on the Lombard effect induced by noise with differing degrees of EM and IM was evaluated. Evidence of speaking behaviors that improve an interlocutor's information reception were found. In particular, the evidence of an active speaking strategy which retimes speech contributions to take advantage of noise-free temporal regions was demonstrated for the first time (chapter 6).

7.2 Implications

By employing a computational model, the studies in chapters 3 and 5 demonstrated that the intelligibility advantage of Lombard speech in noise results from an increased amount of glimpses as a consequence of speech production modifications. However, it is noticeable that the glimpsing model first proposed by Cooke (2006) and used in this thesis does not involve the idea that the amount of energetic masking effect yielded by noise on speech perception could also be affected by the location of the frequency regions glimpsed as observed by Li and Loizou (2007). For example, Li and Loizou (2007) found that the availability of glimpses in frequency regions containing the first two formants caused more masking release compared to in higher frequencies. Therefore, such a factor that the availability of glimpses in different frequency regions has relatively different perceptual importance needs to be taken into account in order to obtain a more realistic computational model of speech perception in noise.

The findings of this thesis serve not only to increase our understanding of the links between speech production and perception, but also have technological relevance. The results of chapters 3 and 6 suggest a need to incorporate more specific information about acoustic-phonetic speech changes provoked by factors that could be present in the talker’s physical environment such as an unintelligible noise, another talker’s voice and a communicative intent, in order to make automatic speech recognition systems robust to real-world conditions. Algorithms for recognizing noise-induced speech have been successfully implemented via model compensation approaches which adjust the parameters of neutral-trained acoustic models to accommodate Lombard speech (e.g. Womack and Hansen, 1999; Bořil, 2008). Other similar techniques involve speech enhancement to transform Lombard speech towards neutral speech (e.g. Lee and Rose, 1996; Bou-Ghazale and Hansen, 2000).

The current findings also demonstrate that people can adapt their speech production to compensate for the interference provoked by different noise sources. Such an adaptive capacity is necessary for reliable and effective speech communication technology, an idea also raised by Moore (2007). For instance, human-robot interactions have been explored in recent years. Among others, Martinson and Brock (2007) demonstrated a robotic system that interacts with humans by adaptively turning up the volume of the spoken output when a noise is present. However, this approach may not be desirable since over-amplification of sound could result in a loss of fidelity and is likely to cause damage to the hearing of the listener. This thesis shows the prospect of speech enhancement in noise by reassigning speech energy in time and frequency to produce more speech-dominated regions. Such an idea could facilitate the development of speech communication applications where an adaptive capacity of enhancing speech in the presence of noise is desired such as spoken output application e.g. human-robot interaction and talking GPS as well as real-time application e.g. telephone communication.

7.3 Future work

It was found in chapter 4 that compared to quiet, in the presence of high-pass filtered noise, speech parameters such as F0 and F1 frequency did not shift to low frequencies, as would have been predicted for an optimal strategy to avoid the noise-concentrated frequency region. Since the effect of noise on the spectral parameters could be obscured by the increase in vocal effort typically induced by noise, it is of interest to investigate whether optimal strategies occur when the increase in speech level is inhibited. One possible way of doing this is to ask subjects to suppress their speech intensity by monitoring visual feedback which shows their vocal level, as was done by Pick et al. (1989). However, the possibility that the visual feedback could distract the subject's attention from the background noise suggests the need to find alternative approaches.

The study in chapter 6 reported that speakers attempt to avoid temporal overlap with background signals. Such an effect was stronger in the condition of competing talker background compared to speech-modulated noise. This might result from that intelligibility of the competing speech background relative to the speech-modulated noise allows a better prediction of upcoming pauses because presumably, if listeners are capable of predicting an upcoming gap, they will be better able to retime their own speech to utilize the noise-free temporal regions. The issue of how the predictability of pauses contained in the background affects temporal overlap between foreground and background signals merits further study. The idea is that prediction of the end of an ongoing utterance may be affected by low level prosodic and intonational factors such as rhythm and pitch contour as well as high level semantic content. For instances, a falling pitch contour may indicate the utterance is coming to a stop, and the end of a sentence with simple semantic content might be easier to predict compared to that of a complicated one.

This thesis has presented evidence that the Lombard effect is dependent on the EM capacity of the background noise. Future work could explore how speech production is affected by noise with differing degrees of IM. Experiments which employ those background speech signals that have different degrees of similarity compared to the foreground spoken utterances in respect of linguistic content, language as well as talker characteristics are required to address this issue. One of the hypotheses of particular interest is that speakers might adopt a strategy of differentiating speech properties such as F0 or speaking rate of their own voice from those of the background competing talker to reduce the foreground-background similarity in an attempt to make a release from IM.

While the work described here was concerned with the effect of noise on native speech production, spoken communication in noise is particularly difficult for non-natives. It has been reported that perceiving speech in noise suffers non-natives more than natives (Florentine et al., 1984; Mayo et al., 1997; Takata and Nábělek, 1990). Garcia-Lecumberri and Cooke (2006) and Cooke et al. (2008) also found that non-native listeners are more adversely affected by both energetic and informational masking. The native advantage in speech perception could be attributed to more familiarity with linguistic patterning at all levels, from acoustic to pragmatic. Since the distinguished features between native and non-native can differentiate the effect of noise on speech perception, future work could be to investigate whether they are able to yield any differing effect of noise on speech production between native and non-native. In addition, although the Lombard effect has been studied in a number of spoken languages such as English (Junqua, 1993; Lu and Cooke, 2008), German (Mixdorff et al., 2006), French (Ramez, 1992; Garnier, 2007), Czech (Bořil, 2008) and Spanish (Castellanos et al., 1996), little effort has been made to compare the difference between languages. It is worthwhile to find out the cross-language difference of the Lombard effect since some elements of the effect might be language specific due to the large linguistic variability across languages.

7.4 Final summary

When talkers speak, they also listen. They listen not only to their own voice but are also affected by background noise and other people's speech. Some of the speaking strategies that talkers might use to ensure effective speech communication under these circumstances have been explored in this thesis. Understanding the full extent to which speakers are able to make things easier for listeners remains a challenging issue for further research.

Bibliography

Alku, P., Vintturi, J. and Vilkman, E. (2002). “Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation,” Speech Communication 38, 321-334.

Andruski, J. E. and Kuhl, P. K. (1996). “The acoustic structure of vowels in mothers’ speech to infants and adults,” Int. Conf. Spoken Lang. Process., 1545-1548.

Arbogast, T. L., Mason, C. R. and Kidd, Jr., G. (2002). “The effect of spatial separation on informational and energetic masking of speech,” J. Acoust. Soc. Am. 112, 2086-2098.

Arlinger, S. D. (1986). “Sound attenuation of TDH-39 earphones in a diffuse field of narrow-band noise,” J. Acoust. Soc. Am. 79, 189-191.

Assmann, P. F. and Katz, W. F. (2005). “Synthesis fidelity and vowel identification,” J. Acoust. Soc. Am. 117, 886-895.

Assmann, P. F. and Nearey, T. M. (2008). “Identification of frequency-shifted vowels,” J. Acoust. Soc. Am. 124, 3203-3212.

Assmann, P. F., Nearey, T. M. and Scott, J. M. (2002). “Modeling the perception of frequency-shifted vowels,” Int. Conf. Spoken Lang. Process., 425-428.

Barker, J. and Cooke, M. P. (2007). “Modeling speaker intelligibility in noise,” Speech Communication 49, 402-417.

Biersack, S., Kempe, V. and Knapton, L. (2005). “Fine-tuning speech registers: A comparison of the prosodic features of child-directed and foreigner-directed speech,” INTERSPEECH Proc., 2401-2404.

Bird, J. and Darwin, C. (1998). “Effects of a difference in fundamental frequency in separating two sentences,” In Palmer, A., Rees, A., Summerfield, Q., and Meddis, R. (Eds.), Psychophysical and Physiological Advances in Hearing, London: Whurr.

Black, J. W. (1959). “Equally contributing frequency bands in intelligibility testing,” J. Speech and Hear. Res. 2, 81-83.

Boersma, P. (1993). “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” Proceedings of the Institute of

Phonetic Sciences 17, 97–110.

Boersma, P. and Weenink, D. (2005). “Praat: doing phonetics by computer (version 4.3.14) (computer program),” (Last viewed May, 2005) from http://www.praat.org.

Boll, S. F. (1979). “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., 27, 113-120.

Bond, Z. S. and Moore, T. J. (1994). “A note on the acoustic-phonetic characteristics of inadvertently clear speech,” Speech Communication 14, 325-337.

Bond, Z. S., Moore, T. J. and Gable, B. (1989). “Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask,” J. Acoust. Soc. Am. 85, 907-912.

Bořil, H. (2008). “Robust speech recognition: analysis and equalization of Lombard effect in Czech corpora,” Ph.D. Thesis, Czech Technical University, Prague.

Bořil, H., Bořil, T. and Pollák, P. (2006). “Methodology of Lombard speech database acquisition: experiences with CLSD,” LREC 2006 - 5th Conference on Language Resources and Evaluation, 1644-1647.

Bou-Ghazale, S. E. and Hansen, J. H. L. (1994). “Duration and spectral based stress token generation for HMM speech recognition under stress,” Int. Conf. Acoustics Speech and Sig. Process., 413-416.

Bou-Ghazale, S. E. and Hansen, J. H. L. (2000). “A comparative study of traditional and newly proposed features for recognition of speech under stress,” IEEE Trans. Speech and Audio Process., 8, 429-442.

Bradlow, A. R., Krause, N. and Hayes, E. (2003). “Speaking clearly for children with learning disabilities: sentence perception in noise,” J. Speech Lang. Hear. Res. 46, 80-97.

Bradlow, A. R., Torretta, G. M. and Pisoni, D. B. (1996). “Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics,” Speech Communication 20, 255-272.

Bronkhorst, A. W. and Plomp, R. (1992). “Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing,” J. Acoust. Soc. Am. 92, 3132-3139.

Brungart, D. S. (2001). “Informational and energetic masking effects in the perception of two simultaneous talkers,” J. Acoust. Soc. Am. 109, 1101-1109.

Burg, J. P. (1975). “Maximum entropy spectrum analysis,” Ph.D. Thesis, Stanford University, USA.

Burnett, T. A., Freedland, M. B. and Larson, C. R. (1998). “Voice F0 responses to manipulation in pitch feedback,” J. Acoust. Soc. Am. 103, 3153-3161.

Burnham, D. K., Vollmer-Conna, U. and Kitamura, C. (2000). “Talking to infants, pets, and adults: What’s the difference?” Paper presented at the XIIth Biennial International Conference on Infant Studies, Brighton, UK.

Byrd, D. (1994). “Relations of sex and dialect to reduction,” Speech Communication 15, 39-54.

Carhart, R., Johnson, C. and Goodman, J. (1975). “Perceptual masking of spondees by combinations of talkers,” J. Acoust. Soc. Am. 58, S35.

Castellanos, A., Benedi, J-M. and Casacuberta, F. (1996). “An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect,” Speech Communication 20, 23-35.

Charlip, W. S. and Burk, K. W. (1969). “Effects of noise on selected speech parameters,” J. Commun. Disord. 2, 212-219.

Chen, F. R. (1980). “Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level,” Unpublished master’s thesis, Massachusetts Institute of Technology, Cambridge.

Chi, S. M. and Oh, Y. H. (1996). “Lombard effect compensation and noise suppression for noisy Lombard speech recognition,” Int. Conf. Spoken Lang. Process., 2013-2016.

Cooke, M. P. (2006). “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am. 119, 1562-1573.

Cooke, M. P., Barker, J., Cunningham, S. and Shao, X. (2006). “An audio-visual corpus for speech perception and automatic speech recognition,” J. Acoust. Soc. Am. 120, 2421-2424.

Cooke, M. P., Garcia-Lecumberri, M. L. and Barker, J. P. (2008). “The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception,” J. Acoust. Soc. Am. 123, 414-427.

Cox, R. M., Alexander, G. C. and Gilmore, C. (1987). “Intelligibility of average talkers in typical listening environments,” J. Acoust. Soc. Am. 81, 1598-1608.

Culling, J. F. and Darwin, C. J. (1993). “Perceptual separation of simultaneous vowels: within and across-formant grouping by F0,” J. Acoust. Soc. Am. 93, 3454-3467.

Denes, P. B. and Pinson, E. N. (1973). “The Speech Chain: The Physics and Biology of Spoken Language,” New York: Anchor Press.

Dreher, J. J. and O'Neill, J. (1957). "Effects of ambient noise on speaker intelligibility for words and phrases," J. Acoust. Soc. Am. 29, 1320-1323.

Drullman, R. (1995). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. 97, 585-592.

Durlach, N. I. (2006). "Auditory masking: Need for improved conceptual structure (L)," J. Acoust. Soc. Am. 120, 1787-1790.

Durlach, N. I., Mason, C. R., Kidd, Jr., G., Arbogast, T. L., Colburn, H. S. and Shinn-Cunningham, B. G. (2003a). "Note on informational masking (L)," J. Acoust. Soc. Am. 113, 2984-2987.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S. and Kidd, G. Jr. (2003b). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. 114, 368-379.

Egan, J. J. (1972). "Psychoacoustics of the Lombard voice response," Journal of Auditory Research 12, 318-324.

Egan, J. P. and Hake, H. W. (1950). "On the masking pattern of a simple auditory stimulus," J. Acoust. Soc. Am. 22, 622-630.

Ferguson, S. H. and Kewley-Port, D. (2002). "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 112, 259-271.

Festen, J. M. and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. 88, 1725-1736.

Fletcher, H., Raff, G. M. and Parmley, F. (1918). "Study of the effects of different sidetones in the telephone set," Western Electrical Company, Report no. 19412, Case no. 120622.

Florentine, M., Buus, S., Scharf, B. and Canevet, G. (1984). "Speech reception thresholds in noise for native and non-native listeners," J. Acoust. Soc. Am. 75, S84.

French, N. R. and Steinberg, J. C. (1947). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. 19, 90-119.

Freyman, R. L., Balakrishnan, U. and Helfer, K. S. (2001). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. 109, 2112-2122.

Freyman, R. L., Balakrishnan, U. and Helfer, K. S. (2004). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. 115, 2246-2256.

Gaillard, A. W. K. and Wientjes, C. J. E. (1994). "Mental load and work stress as two types of energy mobilization," Work Stress 8, 141-152.

Garcia-Lecumberri, M. L. and Cooke, M. P. (2006). "Effect of noise type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. 119, 2445-2454.

Gardner, M. B. (1964). "Effect of noise on listening levels in conference telephony," J. Acoust. Soc. Am. 36, 2354-2362.

Gardner, M. B. (1966). "Effect of noise, system gain, and assigned task on talking levels in loudspeaker communication," J. Acoust. Soc. Am. 40, 955-965.

Garnier, M. (2007). "Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal [Communication in noisy environments : from adaptation to vocal straining]," These de Doctorat de l'Universite Paris 6.

Garnier, M., Bailly L., Dohen, M., Welby, P. and Loevenbruck H. (2006). "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," Int. Conf. Spoken Lang. Process., 2246-2249.

Gordon-Salant, S. (1986). "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," J. Acoust. Soc. Am. 80, 1599-1607.

Gramming, P., Sundberg, J., Ternstöm, S., Leanderson, R. and Perkins, W. (1988). "Relationship between changes in voice pitch and loudness," J. Voice 2, 118-126.

Gu, Y. and Mason, J. S. (1989). "Speaker normalization via a linear transformation on a perceptual feature space and its benefits in ASR adaptation," Proc. European Conf. on Speech Communication and Technology, 258-261.

Hain, T. C., Burnett, T. A., Larson, C. R. and Kiran, S. (2001). "Effects of delayed auditory feedback (DAF) on the pitch-shift reflex," J. Acoust. Soc. Am. 109, 2146-2152.

Hain, T. C., Larson, C. R., Burnett, T. A., Kiran, S. and Singh, S. (2000). "Instructing participants to make a voluntary response reveals the presence of two vocal responses to pitch-shifted stimuli," Exp. Brain Res. 130, 133-141.

Hanley, T. D. and Steer, M. D. (1949). "Effect of level of distracting noise upon speaking rate, duration, and intensity," J. Speech Hear. Disord. 14, 363-368.

Hansen, J. H. L. (1988). "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. Thesis, Georgia Institute of Technology, USA.

Hansen, J. H. L. (1994). "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," IEEE Trans. Speech and Audio Process., 2, 598-614.

Hansen, J. H. L. (1996). "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," Speech Communication 20, 151-170.

Hansen, J. H. L. and Bria, O. N. (1990). "Lombard effect compensation for robust automatic speech recognition in noise," Int. Conf. Spoken Lang. Process., 1125-1128.

Hazan, V. and Markham, D. (2004). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. 116, 3108-3118.

Hazan, V. and Simpson, A. (1998). "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Communication 24, 211-226.

Howell, P. (1990). "Changes in voice level caused by several forms of altered auditory feedback in fluent speakers and stutterers," Lang. Speech 33, 325-338.

Howell, P. and Sackin, S. (2002). "Timing interference to speech in altered listening conditions," J. Acoust. Soc. Am. 111, 2842-2852.

Howell, P., Young, K. and Sackin, S. (1992). "Acoustical changes to speech in noisy and echoey environments," Proc. of ISCA Tutorial and Research Workshop (ETRW) on Speech Processing in Adverse Conditions, 223-226.

Jones, C., Berry, L. and Stevens, C. (2007). "Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners," Computer Speech and Language 21, 641-651.

Jones, J. A. and Munhall, K. G. (2000). "Perceptual calibration of *F0* production: Evidence from feedback perturbation," J. Acoust. Soc. Am. 108, 1246-1251.

Junqua, J. C. (1993). "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am. 93, 510-524.

Junqua, J. C. (1994). "A duration study of speech vowels produced in noise," Int. Conf. Spoken Lang. Process., 419-422.

Junqua, J. C. (1996). "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Communication 20, 13-22.

Junqua, J. C. and Anglade, Y. (1990). "Acoustic and perceptual studies of Lombard speech: Application to isolated-word automatic speech recognition," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2, 841-844.

Junqua, J. C., Fincke, S. and Field, K. (1998). "Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition," Int. Conf. Spoken Lang. Process., 467-470.

Junqua, J. C., Fincke, S. and Field, K. (1999). "The Lombard effect: a reflex to better communicate with others in noise," Int. Conf. Acoustics Speech and Sig. Process., 2083-2086.

Kawahara, H. (1997). "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," Int. Conf. Acoustics Speech and Sig. Process., 1303-1306.

Kawahara, H. (1998). "Perceptual effects of spectral envelope and F0 manipulations using the STRAIGHT method," Proceedings of 135th Meeting of the Acoustical Society of America, 103, 2776.

Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication 27, 187-207.

Kidd, Jr. G., Mason, C. R. and Arbogast, T. L. (2002). "Similarity, uncertainty and masking in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. 111, 1367-1376.

Kidd, Jr., G., Mason, C. R., Deliwala, P. S., Woods, W. S. and Colburn, H. S. (1994). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. 95, 3475-3480.

Kidd, Jr., G., Mason, C. R., Rohtla, T. L. and Deliwala, P. S. (1998). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. 104, 422-431.

Kitamura, C. and Burnham, D. (1998). "Acoustic and affective qualities of IDS in English," Int. Conf. Spoken Lang. Process., 441-444.

Kitamura, C. and Lorenzo, J. (2004). "Vowel Duration and Pitch Contour as Contenders for Infant Attention," Proceedings of the 10th Australian International Conference on Speech Science and Technology, Macquarie University, Sydney.

Knoll, M. and Uther, M. (2004). "Motherese and Chinese: Evidence of acoustic changes in speech directed at infants and foreigners," J. Acoust. Soc. Am. 116, 2522.

Korn, T. S. (1954). “Effect of psychological feedback on conversationnal noise reduction in rooms,” J. Acoust. Soc. Am. 26, 793-794.

Krause, J. C. and Braida, L. D. (2004). “Acoustic properties of naturally produced clear speech at normal speaking rates,” J. Acoust. Soc. Am. 115, 362-378.

Kryter, K. D. (1946). “Effects of ear protective devices on the intelligibility of speech in noise,” J. Acoust. Soc. Am. 18, 413-417.

Lane, H. L. and Tranel, B. (1971). “The Lombard sign and the role of hearing in speech,” J. Speech Lang. Hear. Res. 14, 677-709.

Lane, H. L., Tranel, B. and Sisson, C. (1970). “Regulation of voice communication by sensory dynamics,” J. Acoust. Soc. Am. 47, 618-624.

Larson, C. R., Burnett, T. A. and Kiran, S. (2000). “Effects of pitch-shift velocity on voice F0 responses,” J. Acoust. Soc. Am. 107, 559-564.

Laures, J. S., and Bunton, K. (2003). “Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions,” J. Commun. Disord. 36, 449-464.

Lee, B. S. (1950). “Effects of delayed speech feedback,” J. Acoust. Soc. Am. 22, 824-826.

Lee, L. and Rose, R. (1996). “Speaker normalization using efficient frequency warping procedure,” Int. Conf. Acoustics Speech and Sig. Process., 353-356.

Lee, S. H. and Jeong, H. (2007). “Real-time speech intelligibility enhancement based on the background noise analysis,” Proceedings of the 4th conference on IASTED international conference: Signal Processing, Pattern Recognition, and Applications, 287-292.

Letowski, T., Frank, T. and Caravella, J. (1993). “Acoustical properties of speech produced in noise presented through supra-aural earphones,” Ear Hear. 14, 332-338.

Levelt, W. J. M. (1983). “Monitoring and self-repair in speech,” Cognition 14, 41-104.

Levelt, W. J. M. (1989). “Speaking: From intention to articulation,” Cambridge, MA: Bradford Books.

Li, N. and Loizou, P. C. (2007). “Factors influencing glimpsing of speech in noise,” J. Acoust. Soc. Am. 122, 1165-1172.

Lindblom, B. (1990). “Explaining phonetic variation: A sketch of the H&H theory,” In Speech Production and Speech Modeling, edited by Hardcastle, W. J. and Marchal,

A. (Kluwer Academic, The Netherlands), 403-439.

Lindblom, B. and Sundberg, J. (1971). "Acoustical consequences of lip, tongue, jaw, and larynx movement," J. Acoust. Soc. Am. 50, 1166-1179.

Lippmann, R. P., Martin, E. A. and Paul, D. B. (1987). "Multi-style training for robust isolated-word speech recognition," Int. Conf. Acoustics Speech and Sig. Process., 705-708

Liu, S., Del Rio, E., Bradlow, A. R. and Zeng, F. G. (2004). "Clear speech perception in acoustic and electric hearing," J. Acoust. Soc. Am. 116, 2374-2383.

Lockwood, P. and Boudy, J. (1991). "Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," INTERSPEECH Proc., 79-82.

Lombard, E. (1911). "Le Signe de l'Elevation de la Voix (The sign of the rise in the voice)," Ann. Maladiers Oreille, Larynx, Nez, Pharynx (Annals of diseases of the ear, larynx, nose and pharynx), 37, 101-119.

Lu, Y. and Cooke, M. P. (2008). "Speech production modifications produced by competing talkers, babble, and stationary noise," J. Acoust. Soc. Am. 124, 3261-3275.

Lu, Y. and Cooke, M. P. (2009a) "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," J. Acoust. Soc. Am. 126, 1495-1499.

Lu, Y. and Cooke, M. P. (2009b) "Speaking in the presence of a competing talker," in Proc. INTERSPEECH, Brighton.

Lu, Y. and Cooke, M. P. (2009c). "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," Speech Communication 51, 1253-1262.

Martinson, E. and Brock, D. (2007). "Improving human-robot interaction through adaptation to the auditory scene," Proc. ACM/IEEE International Conference on Human-Robot Interaction, 113-120.

Mayo, L. H., Florentine, M. and Buus, S. (1997). "Age of second-language acquisition and perception of speech in noise," J. Speech Lang. Hear. Res. 40, 686-693.

McLoughlin, I. V. and Chance, R. J. (1997). "LSP-based speech modification for intelligibility enhancement," Proceedings of 13th International Conference on Digital Signal Processing, 2, 591-594.

Mixdorff, H., Grauwinkel, K., and Vainio, M. (2006). “Time-domain noise subtraction applied in the analysis of Lombard speech,” Proceedings of Speech Prosody.

Mixdorff, H., Pech, U., Davis, C., and Kim, J. (2007). “Map task dialogs in noise – a paradigm for examining Lombard speech,” Proc. Int. Congress of Phonetic Sciences, 1329-1332.

Mokbel, C. (1992). “Reconnaissance de la parole dans le bruit: Bruitage/debruitage [Voice recognition in noisy environments: Sound/denoising],” PhD thesis, Ecole Nationale Superieure des Telecommunications.

Mokbel, C. and Chollet, G. (1991). “Speech recognition in adverse environments: speech enhancement and spectral transformations,” Int. Conf. Acoustics Speech and Sig. Process., 925-928.

Moon, S-J. and Lindblom, B. (1994). “Interaction between duration, context and speaking style in English stressed vowels,” J. Acoust. Soc. Am. 96, 40-55.

Moore, R. K. (2007). “Spoken language processing: Piecing together the puzzle,” Speech Communication 49, 418-435.

Natke, U. and Kalveram, K. T. (2001). “Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables,” J. Speech Lang. Hear. Res. 44, 1-8.

Neff, D. L. and Callaghan, B. P. (1988). “Effective properties of multicomponent simultaneous maskers under conditions of uncertainty,” J. Acoust. Soc. Am. 83, 1833-1838.

Neff, D. L. and Green, D. M. (1987). “Masking produced by spectral uncertainty with multicomponent maskers,” Percept. Psychophys. 41, 409-415.

Niederjohn, R. J. and Grotelueschen, J. H. (1976). “The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression,” IEEE Trans. Acoust. Speech Signal Process., 24, 277.

Patel, R. and Schell, K. W. (2008). “The influence of linguistic content on the Lombard effect,” J. Speech Lang. Hear. Res. 51, 209-220.

Payton, K. L., Uchanski, R. M. and Braida, L. D. (1994). “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing,” J. Acoust. Soc. Am. 95, 1581-1592.

Picheny, M. A., Durlach, N. I. and Braida, L. D. (1986). “Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech,” J. Speech Lang. Hear. Res. 29, 434-446.

Pick, H. L., Jr., Siegel, G. M., Fox, P. W., Garber, S. R. and Kearney, J. K. (1989). "Inhibiting the Lombard effect," J. Acoust. Soc. Am. 85, 894-900.

Pickett, J. M. (1956). "Effects of vocal force on the intelligibility of speech sounds," J. Acoust. Soc. Am. 28, 902-905.

Pickett, J. M. (1958). "Limits of direct speech communication in noise," J. Acoust. Soc. Am. 30, 278-281.

Pile, E. J. S., Dajani, H. R., Purcell, D. W. and Munhall, K. G. (2007). "Talking under conditions of altered auditory feedback: Does adaptation of one vowel generalize to other vowels," Proceedings of 16th International Congress of Phonetic Sciences, 645-648.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C. and Yuchtman, M. (1985). "Some acoustic-phonetic correlates of speech produced in noise," Int. Conf. Acoustics Speech and Sig. Process., 1581-1584.

Pittman, A. L. and Wiley, T. L. (2001). "Recognition of speech produced in noise," J. Speech Lang. Hear. Res. 44, 487-496.

Pollack, I. (1975). "Auditory informational masking," J. Acoust. Soc. Am. 57, S5.

Rabiner, L. and Juang, B. H. (1993). "Fundamentals of speech recognition," Prentice Hall, New Jersy.

Rajasekaran, P., Doddington, G. and Picone, J. (1986). "Recognition of speech under stress and in noise," Int. Conf. Acoustics Speech and Sig. Process., 733-736.

Ramez, R. (1992). "Changes in speaker articulation due to ambient noise," Technical report, CRIN-CNRS/INRIA Lorraine, ESPRIT II (5516) ROARS Project, D25 Report.

Rivers, C. and Rastatter, M. P. (1985). "The effects of multi-talker and masker noise on fundamental frequency variability during spontaneous speech for children and adults," Journal of Auditory Research 25, 37-45.

Rostolland, D. (1985). "Intelligibility of shouted speech," Acustica, 57, 104-121.

Ryalls, J. H. and Lieberman, P. (1982). "Fundamental frequency and vowel perception," J. Acoust. Soc. Am. 72, 1631–1634.

Scarborough, R., Brenier, J., Zhao, Y., Hall-Lew, L. and Dmitrieva, O. (2007). "An acoustic study of real and imagined foreigner-directed speech," Proc. Int. Congress of Phonetic Sciences, 2165-2168.

Schulman, R. (1985). "Dynamic and perceptual constraints of loud speech," J. Acoust. Soc. Am. Suppl. 1 78, S37.

Schulman, R. (1989). “Articulatory dynamics of loud and normal speech,” J Acoust. Soc. Am. 85, 295-312.

Schum, D. J., Matthews, L. J. and Lee, F. (1991). “Actual and predicted word-recognition performance of elderly hearing-impaired listeners,” J. Speech and Hear. Res. 34, 636-642.

Simpson, S. A. and Cooke, M. P. (2005). “Consonant identification in *N*-talker babble is a nonmonotonic function of *N*,” J. Acoust. Soc. Am. 118, 2775-2778.

Skowronski, M. D. and Harris, J. G. (2006). “Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments,” Speech Communication 48, 549-558.

Smiljanić, R. and Bradlow, A. R. (2005). “Production and perception of clear speech in Croatian and English,” J. Acoust. Soc. Am. 118, 1677-1688.

Smith, C. L. (2007). “Prosodic accommodation by French speakers to a non-native interlocutor,” Proc. Int. Congress of Phonetic Sciences, 1081-1084.

Sommers, M. S. (1997). “Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment,” J. Acoust. Soc. Am. 101, 2278-2288.

Stanton, B., Jamieson, L. and Allen, G. (1988). “Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions,” Int. Conf. Acoustics Speech and Sig. Process., 331-334.

Steeneken, H. J. M. and Hansen, J. H. L. (1999). “Speech under stress conditions: overview of the effect on speech production and on system performance,” Int. Conf. Acoustics Speech and Sig. Process., 2079-2082.

Stevens, K. N. (2000). ”Acoustic Phonetics (Current Studies in Linguistics),” The MIT Press, New Ed edition.

Stuart, A., Kalinowski, J., Rastatter, M. P. and Lynch, K. (2002). “Effect of delayed auditory feedback on normal speakers at two speech rates,” J. Acoust. Soc. Am. 111, 2237-2241.

Studebaker, G. A. and Sherbecoe, R. L. (1991). “Frequency-importance and transfer functions for recorded CID W-22 word lists,” J. Speech and Hear. Res. 34, 427-438.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I. and Stokes, M. A. (1988). “Effects of noise on speech production: Acoustic and perceptual analysis,” J. Acoust. Soc. Am. 84, 917-928.

Suzuki, T., Nakajima, K. and Abe, Y. (1994). “Isolated word recognition using models for acoustic phonetic variability by Lombard effect,” Int. Conf. Spoken Lang.

Process., 999-1002.

Takata, Y. and Náběelek, A. K. (1990). “English consonant recognition in noise and in reverberation by Japanese and American listeners,” J. Acoust. Soc. Am. 88, 663-666.

Takizawa, Y. and Hamada, M. (1990). “Lombard speech recognition by formant-frequency-shifted LPC cepstrum,” Int. Conf. Spoken Lang. Process., 293-296.

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S., Schreiner, C., Jenkins, W. and Merzenich, M. (1996). “Language comprehension in language-learning impaired children improved with acoustically modified speech,” Science 271, 81-84.

Tartter, V. C., Gomes, H. and Litwin, E. (1993). ”Some acoustic effects of listening to noise on speech production,” J. Acoust. Soc. Am. 94, 2437-2440.

Thomas, I. B. (1967). “The second formant and speech intelligibility,” in Proc. Nut. Electronics Conf. 23, 544-548.

Thomas, I. B. (1968). “The influence of first and second formants on the intelligibility of clipped speech,” J. Audio Eng. Soc. 16, 182-185.

Treisman, A. (1964). “Verbal cues, language, and meaning in selective attention,” Am. J. Psychol., 77, 206-219.

Trainor, L. J. and Desjardins, R. N. (2002). “Pitch characteristics of infant-directed speech affect infants’ ability to discriminate vowels,” Psychonomic Bulletin and Review 9, 335-340.

Tufts, J. B. and Frank, T. (2003). “Speech production in noise with and without hearing protection,” J. Acoust. Soc. Am. 114, 1069-1080.

Uchanski, R. M., Geers, A. E. and Protopapas, A. (2002). “Intelligibility of modified speech for young listeners with normal and impaired hearing,’’ J. Speech Lang. Hear. Res. 45, 1027-1038.

Uther, M., Knoll, M. A. and Burnham, D. (2007). “Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech,” Speech Communication 49, 2-7.

Varadarajan, V. S. and Hansen, J. H. L. (2006). “Analysis of Lombard effect under different types and levels of noise with application to In-set Speaker ID system,” Int. Conf. Spoken Lang. Process., 937-940.

Vestergaard, M. D., Fyson, N. R. C. and Patterson, R. D. (2009). “The interaction of vocal characteristics and audibility in the recognition of concurrent syllables," J. Acoust. Soc. Am. 125, 1114-1124.

Watson, C. S. (1987). "Uncertainty, informational masking and the capacity of immediate auditory memory," in Auditory Processing of Complex Sounds, edited by Yost, W. A. and Watson, C. S. (Erlbaum, Hillsdale, NJ), 267-277.

Watson, C. S., Wroton, H. W., Kelly, W. J. and Benbassat, C. A. (1975). "Factors in the discrimination of tonal patterns. I: Component frequency, temporal position, and silent intervals," J. Acoust. Soc. Am. 60, 1175-1185.

Watson, P. J. and Schlauch, R. S. (2008). "The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours," Am. J. Speech Lang. Pathol. 17, 348-355.

Webster, J. C. and Klumpp, R. G. (1962). "Effects of ambient noise and nearby talkers on a face-to-face communication task," J. Acoust. Soc. Am. 34, 936-941.

Womack, B. and Hansen, J. (1996). "Classification of speech under stress using target driven features," Speech Communication 20, 131-150.

Womack, B. and Hansen, J. (1999). "N-channel hidden Markov models for combined stress speech classification and recognition," IEEE Trans. Speech and Audio Process., 7, 668-677.

Xu, Y., Larson, C. R., Bauer, J. J. and Hain, T. C. (2004). "Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences," J. Acoust. Soc. Am. 116, 1168-1178.

Young, K., Sackin, S., and Howell, P. (1993). "The effects of noise on connected speech: a consideration for automatic speech processing," In Visual Representation of Speech Signals, edited by Cooke, M. P., Beet, S. and Crawford, M. (John Wiley & Sons, Inc. New York, USA), 371-378.

Young, S., Kershaw, D., Odell J, Ollason, D., Valtchev, V. and Woodland, P. (1999). "The HTK Book 2.2," Entropic, Cambridge.