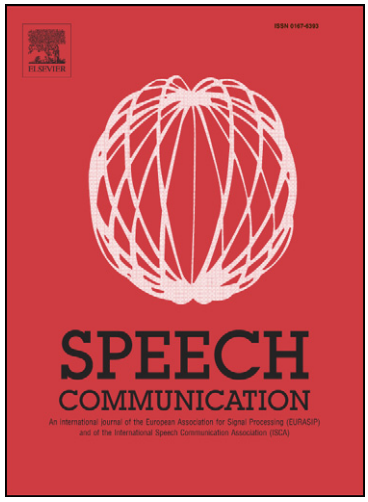
The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise

Youyi Lu, Martin Cooke



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

Youyi Lu, Martin Cooke

PII: S0167-6393(09)00125-3
DOI: 10.1016/j.specom.2009.07.002
Reference: SPECOM 1822

To appear in: *Speech Communication*

Received Date: 11 March 2009
Revised Date: 24 July 2009
Accepted Date: 24 July 2009

Please cite this article as: Lu, Y., Cooke, M., The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.07.002

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

Youyi Lu [a,*]

Martin Cooke [b,c]

[a] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK

[b] Ikerbasque: Basque Science Foundation

[c] Language and Speech Laboratory, Facultad de Letras, Universidad del País Vasco, Vitoria, Spain

[*] Corresponding author. Electronic mail: y.lu@dcs.shef.ac.uk; Tel.: +44 114 2221822; Fax: +44 114 2221810

Abstract

Talkers modify the way they speak in the presence of noise. As well as increases in voice level and fundamental frequency (F0), a flattening of spectral tilt is observed. The resulting "Lombard speech" is typically more intelligible than speech produced in quiet, even when level differences are removed. What is the cause of the enhanced intelligibility of Lombard speech? The current study explored the relative contributions to intelligibility of changes in mean F0 and spectral tilt. 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 simulate those of natural Lombard speech. In the presence of speech-shaped noise, flattening of spectral tilt contributed greatly to the intelligibility gain of noise-induced speech over speech produced in quiet while an increase in F0 did not have a significant influence. The perceptual effects of spectrum flattening was attributed to its ability of increasing the amount of speech time-frequency plane "glimpsed" in the presence of noise. However, spectral tilt changes alone could not fully account for the intelligibility of Lombard speech. Other changes observed in Lombard speech such as durational modifications may well contribute to intelligibility.

Keywords: intelligibility; noise; speech production, spectral tilt

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 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 on speech produced in noise, so called "Lombard speech". Lombard speech 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; Lu and Cooke, 2008). Apart from an increase in speech level, these and many other studies (e.g. Tartter et al., 1993; Hansen, 1996; Steeneken and Hansen, 1999; Garnier et al., 2006) have converged on a set of primary acoustic changes seen in Lombard speech relative to speech produced in quiet. Specifically, Lombard speech demonstrates an overall increase in duration (although vowels and consonants are differentially affected), and increase in F0 and a flattening of spectral tilt (more energy at higher frequencies). 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; Lu and Cooke, 2008). In addition to these primary changes, acoustic-phonetic modifications in consonant-vowel energy ratio and formant frequencies have been reported. 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. The first and second formant frequency (F1 and F2) also shift, with the consensus that F1 tends to increase (Summers et al., 1988; Lu and Cooke, 2008) while F2 has been reported to increase (Junqua, 1993) or decrease (Pisoni et al., 1985).

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) 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. 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 1 using a simple 6-word sentence, drawn from Lu and Cooke (2008), produced in quiet and in 3 Lombard conditions in which the utterance was produced in noise backgrounds of 82, 89 and 96 dB SPL.

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 Lu and Cooke (2008) 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 a glimpsing model 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. The glimpsing model is based on the idea that listeners utilise spectro-temporal regions of favourable local signal-to-noise ratio to identify speech. Glimpsing has been used to predict the intelligibility of intervocalic consonants across a range of masker types (Cooke, 2006), the relative intelligibility of sentence material from different speakers (Barker and Cooke, 2007) as well as the intelligibility of Lombard speech (Lu and Cooke, 2008).

2. Intelligibility of manipulated speech

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 by Lu and Cooke (2008). In their 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 multitalker speech corpus (Cooke et al., 2006), which specifies simple 6-word sentence-like materials such as "bin green at K 4 now" or "place red by E 7 please". 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 current 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 ms 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[i]) 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 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). 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 2.

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.

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 Lu and Cooke (2008). The spectrum of the masker equalled the long-term average speech spectrum of the Grid corpus (figure 3). A masker with a speech-shaped spectrum was chosen because it was found in Lu and Cooke (2008) 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.

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 using a calibrated software audiometer which was used to test each ear at the 6 frequencies: 250, 500, 1000, 2000, 4000 and 8000 Hz. All had normal hearing level. Ages ranged from 20 to 31 years (mean: 26.2). Ethics permission was obtained following the University of Sheffield Ethics Procedure.

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.

2.4. Results

2.4.1 Effect of resynthesis procedure

Speech intelligibility in the re-synthesized and original conditions of "Quiet", "Spec_89" and "Lomb_89" was compared to determine the effect of any artefacts which might have been introduced by STRAIGHT processing. A 2-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.

2.4.2 Effects of manipulated speech on intelligibility

Figure 4 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 Lu and Cooke (2008), and consequently, the average increase of Lombard speech intelligibility over the quiet was somewhat lower in the current study compared to that reported in Lu and Cooke (16 versus 24 percentage points). This difference may be due to the fact that 7 of the 10 listeners recruited for the current study had prior experience of Grid sentences in other speech perception and production experiments.

Paired-samples *t*-tests 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 2-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 ($F(2.30,20.74)=127.96$, $p<0.001$, $\eta_p^2=0.93$) and manipulation level ($F(1.19,8.32)=8.67$, $p<0.05$, $\eta_p^2=0.55$).

Between the 4 types of manipulation collapsed across manipulation level, post-hoc pairwise comparisons (here and elsewhere 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=0.29, p=0.59).

2.5. Discussion

The behavioural 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 (Lu and Cooke, 2008) 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-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.

3. Does manipulated speech offer more glimpsing opportunities?

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). 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, Lu and Cooke (2008) 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.

3.2. Glimpse measures

Two glimpsing statistics were measured for the signal mixtures used in the intelligibility experiment described in the previous section. One, glimpse "area", is the number of spectro-temporal points where the glimpse criteria described above hold. In addition to glimpse area, which is dependent on signal duration, a second measure, the proportion of spectro-temporal points meeting the glimpse criteria was also computed, namely glimpse "proportion". This latter measure is independent of duration, and helps to distinguish those speech production processes which improve glimpsing opportunities by slowing speech rate from those which reallocate energy in the frequency domain.

Computation of glimpse measures was based on a spectro-temporal excitation pattern representation formed for the target and masker independently. This representation is produced by first passing the time-domain signal through a 64

channel gammatone filterbank, smoothing the Hilbert envelopes at the output of the filters, integrating the energy into 10 ms frames, followed by log compression. Following Cooke (2006), a minimum "area" of 5 "pixels" and a local SNR of -5 dB were used here.

3.3. Results

Figure 5 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 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 ($p<0.001$).

For glimpse area, a 2-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 ($F(2.10,14.71)=103.23$, $p<0.001$, $\eta_p^2=0.94$) 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 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 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 ($F(1.19,8.32)=8.67$, $p<0.05$, $\eta_p^2=0.53$). 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 listener intelligibility gain (r=0.988, p<0.001) as shown in figure 6 which plots relative increases in intelligibility for listeners against relative increase in glimpse area.

For glimpse proportion, a 2-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 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 2-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 ($F(1,7)=15.14$, $p<0.01$, $\eta_p^2=0.68$), 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.

4. General discussion

The current study 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 upwards migration of speech energy to regions less likely to be masked by speech-shaped noise (figure 3). The increase in F0 led to a rather small amount of energy

migration to higher frequencies compared to the speech in quiet (figure 7), 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 Lu and Cooke (2008), 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.

While a spectral flattening strategy is beneficial for noises with a falling spectrum typical of many natural noise types (e.g. multitalker babble) used to induce Lombard speech, it is not necessarily helpful for noises with a greater energy concentration in higher frequencies. However, Lu and Cooke (2009) 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 normalised 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 affects 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 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 current study that speech enhancement can be realised 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. Conclusions

The current study 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 Lu and Cooke (2008), the current study also found a high correlation between speech intelligibility and the amount of the time-frequency plane glimpsed. 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.

Acknowledgment

The second author acknowledges support from the EU Marie Curie Network “Sound to Sense”. The authors would like to thank Hideki Kawahara for providing the Matlab implementation of STRAIGHT v40.

[i] 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).

References

Assmann, P.F., Nearey, T.M., 2008. Identification of frequency-shifted vowles. J. Acoust. Soc. Am. 124, 3203-3212.

Assmann, P.F., Nearey, T.M., Scott, J.M., 2002. Modelling the perception of frequency-shifted vowels. Int. Conf. Spoken Lang. Proc., 425-428.

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

Barker, J., Cooke, M.P., 2007. Modelling speaker intelligibility in noise. Speech Communication 49, 402-417.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a samples sound. Proceedings of the Institute of Phonetic Sciences 17, 97-110.

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

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., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120, 2421-2424.

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

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

Ferguson, S.H., 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.

Garnier, M., Bailly L., Dohen, M., Welby, P., Loevenbruck H., 2006. An acoustic and articulatory study of Lombard speech: Global effects on the utterance. Int. Conf. Spoken Lang. Proc., 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.

Hansen, J.H.L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Communication, Special Issue on Speech Under Stress, 20(2), 151-170.

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

Hazan, V., 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.

Hillenbrand, J.M., Clark, M.J., 2000. Some effects of duration on vowel recognition. J. Acoust. Soc. Am. 108, 3013-3022.

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

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

Kawahara, H., 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. Int. Conf. Acoustics Speech and Sig. Proc., 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, vol.103, no.5, p.2776.

Kawahara, H., Masuda-Katsuse, I., 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.

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

Laures, J.S., 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, S.H., 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.

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

Lu, Y., Cooke, M.P., 2009. Speech production modifications produced in the presence of low-pass and high-pass filtered noise. To appear in J. Acoust. Soc. Am. 126.

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

Niederjohn, R.J., Grotelueschen, J.H., 1976. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. IEEE Trans. ASSP-24, 277.

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

Ryalls, J.H., Lieberman, P., 1982. Fundamental frequency and vowel perception. J. Acoust. Soc. Am. 72, 1631-1634.

Sawusch, J., 1996. Effects of duration and formant movement on vowel perception. Int. Conf. Spoken Lang. Proc., 2482-2485.

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

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.

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

Steeneken, H.J.M., 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. Proc., 2079-2082.

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

Tartter, V.C., Gomes, H., 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. 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.

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

Watson, P.J., 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.

Figure 1. 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.

Figure 2. 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”.

Figure 3. Long-term average spectrum of the Grid corpus.

Figure 4. Relative improvements in keyword identification scores for speech with acoustic modifications 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%.

Figure 5. 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% respectively. “Lomb_area” and “Lomb_prop” represent the area and proportion of glimpses measured for the Lombard speech conditions.

Figure 6. Relation between increase in glimpse area and intelligibility, together with least-squares fit.

Figure 7. 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 normalised to have equal root-mean-square energy.