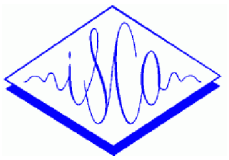

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

The Intelligibility of Lombard Speech: Communicative setting matters

Michael Fitzpatrick, Jeesun Kim, Chris Davis

MARCS Institute, University of Western Sydney, Australia
michael.fitzpatrick@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au

Abstract

Recently we reported that talkers modified their speech production strategies in noise as a function of whether their interlocutor could or could not be seen, i.e. face-to-face (FTF) or non-visual conditions (NV). Participants made greater auditory speech modifications (e.g. in terms of amplitude and F0) in NV condition, and greater visual speech modifications (in terms of inter-lip area) in FTF condition [1]. The current study examined whether such modifications led to corresponding differences in speech intelligibility in the different settings. In the current experiment, participants were presented with a set of consonant-vowel-consonant (CVC) phonemes in noise at a fixed SNR in auditory-only, visual-only and auditory-visual conditions. The CVC stimuli were drawn from speech recordings in quiet and in noise conditions, and also during NV and FTF conditions from [1]. The results showed that the speech in noise tokens produced in the FTF conditions had a greater AV benefit than for tokens produced in the NV conditions. Also, the AV benefit was greater for speech tokens produced in noise than for speech produced in quiet. The results were discussed in terms of efficient talker and listener strategies .

Index Terms: Lombard speech, AV speech, speech production.

10.21437/Interspeech.2012-470

1. Introduction

Talkers modify their speech production in noisy environments – the change in production style in noise is known as Lombard Speech (following [2]). The main acoustic characteristics of Lombard compared to speech produced in quiet include increases in loudness, vowel duration, f0, as well as a flattening of spectral tilt [3]. In addition to acoustic modifications, talkers also modify visual parameters of their speech, i.e., increases in both rigid (e.g. head movement – e.g. [4]) and non-rigid motion (e.g. inter-lip area, Lip protrusion, mouth and jaw opening – e.g. see [1, 4, 5, 6]) have been reported for speech produced in noise.

Although Lombard speech production is in part a reflexive response to talking in background noise [7], it also has features that are shaped by the communication setting [8, 9]. That is, both the environmental constraints (i.e. type and level of background noise [10]) as well as the needs of the interlocutor [11] influence the production modifications that talkers make. Supporting this idea, several studies have reported that talkers make increased, production modifications in noise during communicative tasks, compared to those modifications brought about where communicative intent is absent or minimal [5, 12],.

In a recent study [1] we examined talkers' in noise speech production in a communicative task where the interlocutors could or could not see each other (a Face-To-Face (FTF) vs. a Non-Visual (NV) condition) and found that in the NV conditions talkers made relatively greater *auditory* modifications to their speech production (in terms of amplitude, f0 and spectral tilt), whereas in the FTF conditions they made greater *visual* modifications (in terms of inter-lip area). These results support the proposal that the manner that speech is produced in noise is a flexible and adaptive process shaped by a talker's communicative intent [13]. Specifically, it appeared that talkers selectively modified aspects of the auditory-visual speech signal to suit the needs of their interlocutor with respect to constraints imposed by the communicative setting.

The above interpretation rests on the assumption that the speech signals produced in the different settings should lead to corresponding differences in intelligibility. This is what we tested this in the current study. That is, we examined whether there would be a difference in the auditory and AV *intelligibility* of the speech produced across the NV and FTF conditions from [1]. Given the difference in production style across the conditions, we hypothesised that there would be: A) greater *auditory* intelligibility for the perception of Lombard speech produced in the NV condition (relative to the FTF speech stimuli), and B) a greater *AV benefit* for Lombard speech produced in FTF condition (relative to the NV speech stimuli).

2. Method

2.1. Participants

Fifteen Australian English volunteers (three males; twelve females) participated in the study. All of the participants reported normal or corrected to normal vision and hearing. All were native speakers of English.

2.2. Stimuli

The stimuli were drawn from the auditory and visual recordings made from our previous study [1]. In [1] pairs of talkers played a game similar to a Sudoku task where the repetition of several consonant-vowel-consonant (hVd) tokens was required. The current stimuli were taken from renditions in the FTF and NV settings for both quiet and SSN noise conditions i.e., FTF Quiet, FTF Lombard, NV Quiet and NV Lombard. From the auditory and visual recordings made in each of the four conditions, 6 hVd tokens were selected: "hAd, hARd, hEEd, hId, hOd, hUd". Note that since the materials produced in the conversation task were spontaneous speech, not all tokens were suitable to be used as stimuli in the current experiment. Thus, the 6 tokens were selected such that:

a) The tokens could be segmented clearly with minimal auditory or visual co-articulation before or after the consonants (i.e. the tokens were not from running speech).
b) There were no idiosyncratic artefacts that could be used to identify the speech token (such as licking the lips or smiling before or after the token).

c) Tokens were representative of the median values in the production condition (i.e. we did not select outliers as stimuli from each condition).
d) Adequate repetitions of the items were available in each vowel category that met the above conditions.

Two exemplars of each hVd token were selected from the auditory-visual recordings from two talkers in the experiment, across the NV and FTF conditions, in quiet and noise. The summary of the acoustic and visual parameters of the chosen stimuli are presented in table 1 (cf. [1]).

Three types of stimuli were rendered from the auditory and visual recordings: Auditory-Only (AO), Visual-Only (VO) and Auditory-Visual (AV). The auditory stimuli were normalized with PRAAT [14] software such that their peak amplitude of the vowel portion of the items was 60dB. They were then mixed with speech-shaped-noise (SSN) at a SNR of -10dB (a value selected based on a pilot study to avoid floor and ceiling effects). To avoid the onset of the noise and speech signal coinciding, the masking noise was mixed with the stimuli such that the noise always preceded and followed the speech content by some variable amount (range: 500 to 1000 ms).

	Quiet		Lombard	
Measure	NV	FTF	NV	FTF
Duration (ms)	153	140	282	219
F0 (Hz)	138.38	124.13	229.88	198.43
Spectral-Tilt (dB / Octave)	-4.74	-5.29	-2.12	-2.79
% Inter-Lip Area	35.62	32.91	55.69	59.67

Table 1. *Mean Acoustic and Visual Parameters for the Non-Visual (NV) and Face-To-Face (FTF) stimuli across the Quiet and Lombard speech stimuli.*

In [1], colour video recordings were made using a miniature "spycamera" (at 25 fps) which was directed at the talkers' lips and mouth. In preparing the VO and AV stimuli, we tried to minimise any differences across the video stimuli (for example, the talkers eye-gaze varied from the interlocutor to the game grid) so the videos were cropped to contain only the lip, jaw and mouth information using VirtualDub. The AV items were created by realigning the normalised noise-mixed auditory items with the visual items.

In sum, there were three presentation types of stimuli (AO, VO, AV) drawn from the 4 speech production conditions (NV Quiet, NV Lombard, FTF Quiet and FTF Lombard). Each of these 12 conditions consisted of 2 tokens of 6 CVCs (hAd, hARd, hEEd, hId, hOd, hUd) by 2 talkers (N = 288). In the experiment, these were repeated three times so that there were 864 items in total.

2.3. Procedure

The test was run for each participant separately on a laptop PC in a sound attenuated booth. Test presentation and response collection was controlled by the DMDX software program [15]. Audio stimuli were presented through Sennheiser (HD 650) headphones at a comfortable listening level. The visual stimuli were presented in the centre of the screen. Participants were asked to identify which of the 6 tokens they were presented by mouse-clicking one of 6 available response options on the screen (which appeared after each stimulus presentation). Once a response was made, the response options disappeared from the screen and the next stimulus was presented after a 500ms interval.

The AO, AV and VO stimuli were presented in blocks. To avoid learning effects, the VO stimuli were always presented first, followed by either the AV or AO blocks (the presentation order of the AO and AV conditions was counterbalanced). Within each of the presentation conditions, the presented order of the 4 types of recorded stimuli (i.e. the Quiet/Lombard, NV and FTF) was randomized. Before the experiment, the participants completed a practice session that consisted of the 6 HVD tokens presented in an AO condition, 6 without noise, and 6 mixed with SSN set at an easy SNR (0dB SNR). These tokens were produced by a different talker to the ones used in the main experiment.

The entire experiment took around 60 minutes to complete which included breaks between the different presentation conditions.

3. Results

The percentage correct data for the 4 conditions and for the AO, VO and AV conditions are presented in the panels of figure 1. The percentage correct data were analysed with a 3-way repeated measures ANOVA, comparing: Modality (AO, VO, AV), Production Condition (NV, FTF) and Speaking Style (Quiet, Lombard). Significant effects were observed for all three factors: Modality, $F(2, 26) = 74.49$, $p < .001$, $\eta_p^2 = .85$; Production Condition, $F(1, 13) = 18.54$, $p < .001$, $\eta_p^2 = .59$; Speaking Style, $F(1, 13) = 412.92$, $p < .001$, $\eta_p^2 = .97$, as well as the two way and three way interactions amongst the conditions, ($p < .05$)

To explore the interactions, first, the results for each of the AO, VO and AV modalities were analysed separately, i.e., each with a 2-way repeated measures ANOVA comparing Production Condition and Speaking Style. Second, the relative benefit of visual speech on the auditory speech perception was examined. In all analyses, Bonferoni adjustments were made where appropriate.

3.1. Auditory Only (AO) conditions

The results of a 2 way ANOVA on the AO data revealed significant main effects of Production Condition, $F(1, 13) = 19.82$, $p = .001$, $\eta_p^2 = .604$, and Speaking Style, $F(1, 13) = 252$, $p < .001$, $\eta_p^2 = .95$. The interaction, however, was not significant ($p > 0.05$).

3.2. Visual Only (VO) conditions

A 2 way ANOVA on the VO data revealed a significant Main effect of Speaking Style, $F(1, 13) = 15.89$, $p = .002$, $\eta_p^2 = .55$, as well as the interaction between Production Condition and Speaking Style, $F(1, 13) = 13.21$, $p = .003$, $\eta_p^2 = .504$. Follow up simple effects analyses revealed that, for the Lombard speech, talkers were significantly more accurate at perceiving the VO stimuli drawn from the FTF condition than from the NV condition, $F(1, 13) = 11.65$, $p = .005$, $\eta_p^2 = .47$; for the Quiet speech, there was no significant difference ($p > 0.05$).

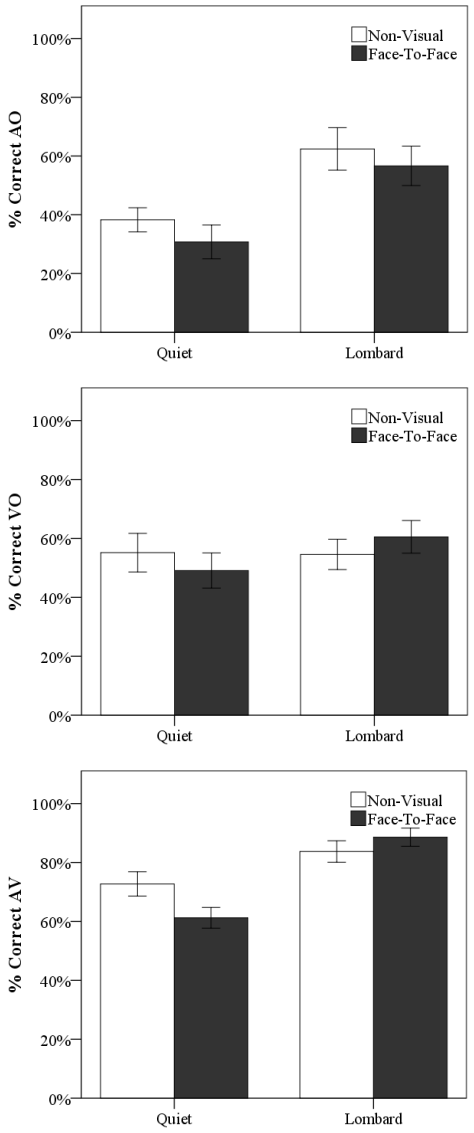

Figure 1. *Mean % Correct results for the Non-Visual (NV) and Face-To-Face (FTF) stimuli across the Quiet and Lombard speech stimuli. The three graphs represent the AO, VO and AV results. Error bars indicate SE.*

3.3. Auditory Visual (AV) conditions

A 2 way ANOVA on the AV data revealed significant main effect of Speaking Style, $F(1, 13) = 267.86$, $p < .001$, $\eta_p^2 = .96$, as well as a significant interaction between Production Condition and Speaking Style, $F(1, 13) = 53.53$, $p < .001$, $\eta_p^2 = .81$. Follow up analyses between the NV and FTF scores for Quiet and Lombard speech were conducted. For the Quiet speech, participants perceived the NV items significantly more accurately than those produced in the FTF setting, $F(1, 13) = 21.97$, $p < .001$, $\eta_p^2 = .63$. Conversely, consistent with our hypothesis, Lombard speech AV tokens produced in the FTF setting were significantly more intelligible than those of the NV conditions, $F(1, 13) = 14.34$, $p = .002$, $\eta_p^2 = .52$.

3.4. AV Relative Benefit

AV Relative Benefit (RB) was calculated by defining the scores as (% Correct AV - % Correct AO)/(100 - % Correct AO) as in [15]. RB is a useful metric of AV intelligibility as it normalises for differences in the accuracy levels of AO data. Thus, the benefit attributable to having visual speech available (as separate to that attributable to the auditory Lombard speech modifications) can be analysed. Here the important result was to examine whether there was firstly, a greater RB for Lombard speech over Quiet speech, and secondly, whether any RB differences were mediated by whether the stimuli was drawn from NV or FTF conditions.

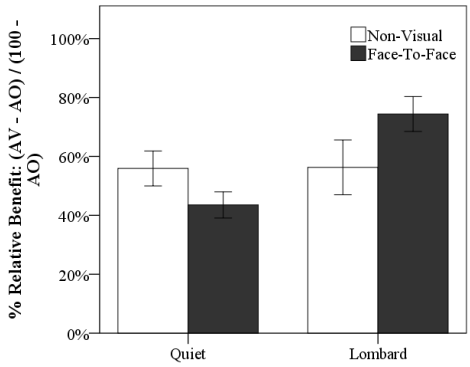

Figure 2. *Mean AV Relative Benefit results for the Non-Visual (NV) and Face-To-Face (FTF) conditions, across the Quiet and Lombard speech stimuli. Error bars indicate SE.*

The RB results are presented in figure 2. A 2 way repeated measures ANOVA revealed a significant main effect of Speaking Style, $F(1, 13) = 17.87$, $p = .001$, $\eta_p^2 = .58$, indicating that, averaged across the NV and FTF conditions, the benefit from having the visual speech available is indeed greater for Lombard speech than it is for quiet speech. There was also a significant interaction between Production Condition and Speaking Style, $F(1, 13) = 28.26$, $p < .001$, $\eta_p^2 = .69$. Follow up simple effects analyses examining the Quiet and Lombard speech separately revealed that although the RB was significantly greater for FTF ($M = 74.39\%$) relative to NV ($M = 52.17\%$) conditions for Lombard speech production, $F(1, 13) = 13.36$, $p < .05$, $\eta_p^2 = .51$, the opposite pattern was the case for the speech produced in quiet (NV $M = 55.91\%$, FTF $M = 43.52\%$), $F(1, 13) = 10.88$, $p < .05$, $\eta_p^2 = .46$.

4. Discussion

The current study used the materials from a Lombard speech production study (that compared FTF to NV conditions, [1]) to examine whether the produced speech translated into differences in auditory and auditory-visual intelligibility for perceivers.

For the auditory speech conditions, consistent with previous research, the results showed that there was a substantial perceptual benefit for Lombard speech over speech produced in Quiet even when amplitude differences have been normalised. Contrary to prior predictions, however, if only the Lombard

speech tokens are considered, there was no significant improvement in the perception of the NV compared to the FTF speech tokens in the AO condition. In contrast, across measures of AV perception (AV % correct, RB), there was a significant increase in the accuracy of perception for FTF relative to the NV Lombard items. This difference in intelligibility across NV and FTF Lombard speech conditions was also echoed in the VO results.

The finding of an improvement in the perception of AV Lombard speech tokens is consistent with previous studies (e.g. [4, 6]). Although the exact source of this improvement remains unclear, the VO results presented in the current study provide some data to support the proposal that the hyper-articulated lip-movements for Lombard speech serve to enhance the intelligibility of AV Lombard speech [5] – possibly by increasing the discriminability between visemes. Further, the fact there was no significant difference in the AO NV and FTF Lombard speech conditions, yet a significant difference in the AV performances (for both AV % correct and RB), supports this argument that significant improvements in AV perception of Lombard speech are strongly related to increased clarity of the visual signal. However, it should be cautioned that this was a limited data set, and not all vowels were equally confusable with each other. Further research with a wider set of stimuli will elaborate on the extent to which viseme categories are affected during Lombard speech.

It is unclear why there was a limited difference between the NV and FTF conditions for the auditory only Lombard speech conditions (compare the right set of white and grey bars in the top panel of figure 1). It may be the case that the difference in acoustic properties between the two sets of stimuli was not great enough to result in a significant perceptual accuracy. This may also have to do with the fact that auditory amplitudes were normalized in the current study – in [1], the largest point of difference between the NV and FTF Lombard speech conditions was for amplitude. While normalizing the amplitude was useful in the current study to examine the Quiet/Lombard speech, and visual speech benefit across the conditions, it may have minimised the acoustic differences between the two speech styles. A follow up study examining perceptual benefit for the original non-normalised stimuli is being conducted.

It is important to note that in the current study we only considered the AV intelligibility benefit related to the perception of talkers lip-and mouth movements. This was a constraint of drawing stimuli from the materials collected in [1]. Informative visual speech information is conveyed in both rigid and non-rigid features [16]. Thus, the videos may have limited the visual information available to talkers in AV speech perception. Alternatively, it is also possible that presenting only the lip and mouth region may have artificially *focussed* talkers toward the most informative aspects of AV speech production in noise. As such, it would be of interest to extend the current findings to more complete videos of talkers in noise to examine which visual regions are the most informative in AV Lombard speech perception.

In summary, the results of current perception study mirror those shown in the acoustic and visual measures of speech produced in the NV and FTF conditions for Lombard and Quiet speech in [1]. In [1] it was argued that talkers are receptive both to the needs of their interlocutor as well as to the constraints placed on them by the environment and are able to modify both auditory and visual elements of speech production to effectively and efficiently communicate in noise [13]. The results of the current study provide perceptual evidence that the shift in production style across NV and FTF communication conditions in noise also corresponds with changes in the intelligibility of the speech signal for the perceiver.

5. Acknowledgements

We thank the participants and acknowledge support from the Australian Research Council (DP0666857).

6. References

[1] Fitzpatrick, M., Kim, J. and Davis, C. “The effect of seeing the interlocutor on auditory and visual speech production in noise”, 11th international conference on Auditory-Visual Speech Processing (AVSP2011), Volterra, Italy, 2011.

[2] Lombard, E., “Le signe de l'elevation de la voix”, Annales des maladies de l'oreille et du larynx, 37:101-119, 1911.

[3] Junqua, J., “The lombard reflex and it role on human listener and automatic speech recognizers”, Journal of the Acoustic Society of America, 93(1):510-524, 1993.

[4] Kim, J., Sironic A., & Davis C. “Hearing speech in noise: Seeing a loud talker is better”, Perception, 40:853-862, 2011.

[5] Garnier M., Henrich N., Dubois D., “Influence of sound immersion and communicative interaction on the Lombard effect”, Journal of Speech, Language, and Hearing Research, 53:588-608, 2010.

[6] Vatikiotis-Bateson, E., Barbosa, A. V., Chow, C. Y., Oberg, M., Tan, J. and Yehia, H. C., “Audiovisual lombard speech: Reconciling production and perception”, in Proceedings of the international Conference on Auditory-Visual Speech Processing (AVSP 2007), Hilvarenbeek, The Netherlands, 2007.

[7] Siegel, G. M., Schork, E. J., Pick, H. L., Garber, S. R., “Parameters of auditory feedback”, Journal of Speech and Hearing Research, 25 (3):473–5, 1982.

[8] Lane, H., and Tranel, B., “The Lombard sign and the role of hearing in speech”, Journal of Speech and Hearing Research, 14:677-709, 1971.

[9] Junqua, J. C., Fincke, S., and Field, K., “The Lombard effect: A reflex to better communicate with others in noise”, in International Conference on Acoustics, Speech and Signal Processing, 2083–2086, 1999.

[10] Lu, Y. and Cooke, M., “Speech production modifications produced by competing talkers, babble, and stationary noise,” Journal of the Acoustic Society of America, 124:3261–3275, 2008.

[11] Hazan, V. and Baker, R., “Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions”, Journal of the Acoustic Society of America, 130(4):2139-2152, 2011.

[12] Lu, Y. and Cooke, M., “Spectral and temporal changes to speech produced in the presence of energetic and informational maskers,” Journal of the Acoustic Society of America, 128(4):2059–2069, 2010.

[13] Lindblom, B., “Role of articulation in speech perception: Cues from production”, Journal of the Acoustic Society of America, 99:1683–1692, 1996.

[14] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer” [computer program]. version 5.3.11, retrieved 25 march 2012 from http://www.praat.org/,” 2012.

[15] Sumby, H. and Pollack, I. W., “Visual Contribution to Speech Intelligibility in Noise", Journal of the Acoustic Society of America, 26:212-215, 1954.

[16] Cvejic, E., Kim J., and Davis C., “Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion,” Speech Communication, 52:555-564, 2010.