The effect of seeing the interlocutor on speech production in different noise types

Michael Fitzpatrick 1, Jeesun Kim 1, Chris Davis 1

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

Abstract

Talkers modify their speech production in noisy environments partly as a reflex but also as an intentional communicative strategy to facilitate the transmission of the speech signal to the interlocutor. Previous studies have shown that the characteristics of such modifications vary depending on the type of noise. The current study examined whether speech production (and its interaction with noise type) would be affected by being able to see their interlocutor or not. Participants completed an interactive communication game in various noise conditions with and without being able to see their interlocutor. The results show that speech modifications differed with noise condition and that the speech amplitude was significantly lower when interlocutors could see each other. These results suggest that talkers actively monitor their environment and adopt appropriate speech production for efficient communication.

Index Terms: Lombard Speech, Speech Production, Communication

1. Introduction

To communicate in everyday conditions, the listener/speaker often has to cope with noise in the auditory environment. One way we achieve this is by modifying the way that speech is produced in noise – such modifications are known as Lombard Speech (following [1]). The main characteristics of Lombard speech compared to speech produced in quiet include increases in loudness and vowel duration, an increase in f0, as well as a shifting of energy from low frequency to high frequency bands (e.g. [2]).

Although Lombard speech is (in part) an automatic reflex triggered by the reduced audibility of one’s own produced speech ([3]), recent evidence suggests that Lombard speech is also an intentional communicative strategy, such as ‘clear-speech’, that talkers adopt in noise to facilitate the perception of the speech signal at the ears of the listener ([4, 5]). For example, several studies have found more pronounced Lombard type modifications in noise for conditions where there is interaction and feedback from an interlocutor (e.g. as in interactive communicative tasks) compared to reading lists or repeating words out loud ([4, 5]).

Further evidence for the role of communication in Lombard speech comes from a recent study examining speech production across different types of background noise. [5] measured participants’ speech production in quiet and in three different types of background noise: competing speech (CS), speech-shaped-noise (SSN) and speech-modulated-noise (SMN). In addition to finding greater speech production modifications for communicative relative to non-communicative tasks, they found that the specific talker modifications (of increased speech output level, spectral tilt and fundamental frequency, as well as increased F1 F2 vowel space expansion) varied across the different background noise conditions (with the greatest modifications being made for the SSN condition). Furthermore, it was reported that listeners were able to monitor the content of the background noise and ‘time’ their speech output to exploit amplitude modulations in the background maskers. To account for the different speech production strategies across different noise types, Cooke and Lu [5] suggested that talkers use a “listening-while-speaking” strategy when producing speech in noise. That is, talkers actively monitor the background noise and adopt the most appropriate and efficient strategy of speech production to offset the effects of masking, and facilitate the perception of the speech signal at the ears of the interlocutor.

Given the above proposal that talkers actively adopt different speech production strategies to maintain efficient communication in different environments, it can be predicted that being able to see the conversational partner may also influence speech production in noise. This prediction is based on two lines of research evidence: (1) the provision of visual speech (e.g. seeing the talker articulating) is known to substantially improve speech intelligibility, especially in noisy conditions [e.g. 6]; (2) people produce more pronounced visual speech signals (e.g. lip/mouth movement [4]) when they can see each other. In face-to-face conditions therefore, as talkers are heard more effectively due to the availability of their visual speech, it is possible that they make less pronounced auditory modifications (such as speaking louder which is energy consuming), and instead increase the saliency of their visual speech signals (which will not be affected by noise). In other words, the availability of visual speech may lessen the need to actively modify auditory signal properties in response to the background noise.

The motivation for the current study therefore was to follow up the proposal of “listening-while-speaking” [5] and examine the role of visual speech cues on speech production in noise. Using a similar experimental paradigm to that of [5], the aim of the study was to examine speech production in noise across different noise types and for conditions where the participants could or could not see their interlocutor. It was hypothesised, that with the availability of visual speech cues, talkers would make less pronounced modifications to their auditory speech than in conditions where they could not see their interlocutor. Furthermore, with access to visual speech talkers would have less need to adapt their production to the background noise type.

2. Method

2.1. Participants

Six Males (Mean Age = 26 years) participated in the experiment. All were native speakers of Australian English and university graduates. All participants had self reported normal levels of hearing and vision.

2.2. Experimental Conditions

There were 8 (4 auditory by 2 visual) communication conditions. The auditory conditions consisted of a quiet no-noise condition (Q) and three types of background noise (as used by [5]): a single competing talker (CS), Speech-Shaped
Noise (SSN), and Speech Modulated Noise (SMN). These noise conditions allow the effects of informational masking (IM) and energetic masking (EM) to be examined by comparing the effects of each. The visual conditions consisted of a face-to-face (FTF) condition (where the speech partners could see and hear each other) and an auditory-only (AO) condition where the speech partners could only hear each other.

The creation of the three noise types followed the procedure outlined in [5]. The CS maskers were created by recording talkers’ speech when completing the task in quiet. From these recordings, all speech, non-speech (e.g. laughing, coughing, and paper-rustling) and silent pauses (>100ms) were transcribed using Praat. All non-speech segments were replaced with silence, and the resultant auditory stimuli were used as the CS masker. No talker heard their own voice as noise in the testing sessions. SSN was then generated from the CS stimuli by filtering white noise to match the long term spectrum of the CS masker using Akustyk. The SMN was created by shaping the resulting SSN signal to the amplitude envelope of the relevant CS masker. In this way, the SMN masker has similar energetic masking potential as the CS signal, but without the meaningful content.

2.3. Recording/Equipment

Speech data collection sessions took place in a sound attenuated booth. Two talkers sat across from each other at a table. Across the middle of the table was a removable paper screen. The screen was used to prevent talkers seeing each other, yet to minimize any differences in the audibility of the speech signal between the AO and FTF conditions.

All auditory recordings were made using Sennheiser HSP4 Headset microphone which were connected to an EDIROL UA-25 external sound card and routed to a PC. Noise maskers were presented through Sennheiser MX560B in-ear headphones at 82dB. The SNR level was chosen as it represents an adequate level to produce EM effects, as well as allow for IM effects for the CS maskers ([5]). The talker’s own speech was looped back through their headphones (at a level determined by the participant at the beginning of the experiment) to reduce any loss of audibility of his/her own voice due to wearing the headphones. By using the small microphone and in-ear headphones set-up, any potential loss of visibility of the talker due to equipment during FTF conditions was minimised.

2.4. Task:

In each communication condition, participants were asked to solve puzzles (similar to 9x9 Sudoku puzzles) in conjunction with a speech partner for a 10 minute period. Sudoku puzzles were used by [5] for the speech production task as such encourage natural and spontaneous conversation amongst the partners, and induce of repetition of particular words (usually the numbers “one” to “nine”) that can be used for reliable phonetic analyses. However, a problem with using numbers as the object of phonetic analyses is that such provide a limited representation of English vowels. To overcome this, for the task used in the current experiment, the numbers (1-9) were replaced with the following 9 hVd tokens: “hAd, hARd, hEd, hED, hOd, hOOd, hld, hUd, hU Ud” (these syllables were printed at the top of the Sudoku grids instead of the numbers of 1-9). In this way, a large set of vowels were produced within the same phoneme context that allowed for a more complete analysis of the vowel space.

Talkers were instructed to solve the puzzles collaboratively with their speech partner and to complete as many puzzles as possible within the time period. No instructions were given as to how to talk to each other, or whether to pay particular attention to each others’ visual speech (e.g. lip movement or facial expression) in conditions where they could see each other. Note that the same speech partner (a confederate) was employed for each of the 6 participants and as such, any differences in speech production attributed to the characteristics of the conversation partner were minimised.

2.5. Procedure

Participants always completed the quiet sessions first (in order to create the background noise material described above) before completing the three noise conditions. The completion order of the noise conditions, as well as the order with which the participants completed the AO or FTF conditions within each noise type, was balanced across the participants.

All participants completed the recordings in two sessions. Recording for each condition lasted for 10 minutes and participants were permitted a break between conditions.

3. Results

3.1. Data Processing

Recorded speech files were segmented using Praat along three tiers: the first containing the boundaries of the specific phoneme, the second for the boundaries of the full hVd token, and the third segmenting whether the material contained speech/non-speech (e.g. non meaningful utterances such as “ummm” or “mmm”, laughing, coughing, paper rustling, as well as any recorded talk from the speech partner, or a pause >100ms in length).

Prior research (e.g., [2]) has shown that duration, amplitude and F0 typically increase for speech produced in noise compared to quiet conditions. Spectral tilt (a measure for the distribution of frequency energy between the lower and higher formants) also has been reported to change [e.g. 5]. Thus, in order to examine ‘Lombard effects’, measures of vowel duration, amplitude, fundamental frequency and spectral tilt were extracted from the vowels for the hVd tokens using a custom script in Praat. In the current study, spectral tilt was defined as the difference in intensity between the first and third formants for the vowels (e.g. as per [7]), and was expressed as dB/Octave.

In addition to these measures, the degree of ‘between’ and ‘within’ f1/f2 vowel space dispersion across the conditions was also examined [8]. Vowel space dispersion is a known characteristic of clear speech (e.g. [9]) and has been claimed to potentially underlie the intelligibility benefit of Lombard speech [e.g. 5]. That is, talkers may adopt a strategy where they produce more distinct vowels (as measured by a larger ‘between’ category dispersion) as well as adopting a more consistent production style (represented by tighter clustering ‘within’ vowel categories) in noisy compared to clear conditions. As such, both vowel-space measures were included in the current study.

The results were first analysed with a repeated measures ANOVA comparing Noise (Q vs CS vs SMN vs SSN) by Presentation Condition (AO vs FTF) for each of the speech production values separately (i.e. duration, amplitude, F0, spectral tilt, between dispersion and within dispersion). Any follow-up analyses were conducted using paired-t tests with Bonferroni adjusted alphas.
3.2. Lombard Measures

Figure 1 represents the duration, amplitude, F0 and spectral tilt values as a function of noise type and presentation condition (i.e. AO vs FTF).

Vowel duration, amplitude, and F0 all significantly increased for productions in noise compared to in quiet conditions [F(3,9)=49.96, p<0.05, η²=0.943 for amplitude; F(3,9)=7.23, p<0.05, η²=0.707 for F0; F(3,9)=3.93, p<0.05, η²=0.567 for duration]. These effects were predominantly driven by differences in production for the SSN condition compared to the other noise conditions (p<0.05). Follow-up post hoc analyses (collapsing across presentation condition) revealed that speech produced in the CS and SMN conditions did not differ from each other on any of the above Lombard measures. These results were consistent with those of [5] who have found larger production modifications as the energetic masking (EM) of the background noise increased.

In contrast to previous studies, there were no significant effects for spectral tilt across the noise conditions. That is, the talkers did not appear to shift the distribution of energy from lower to higher frequencies from quiet to noise conditions.

The main effects of presentation condition were significant (i.e. AO vs FTF) for both amplitude and F0 indicating that the talkers were significantly quieter, and spoke with a lower fundamental frequency in conditions where they could see each other compared to when they couldn’t [F(1,3)=14.29, p<0.05, η²=0.827 for amplitude; F(1,3)=223.52, p<0.05, η²=0.987 for F0]. Post-hoc analyses comparing the presentation condition for each or the noise types showed that the primary point of difference between FTF and AO conditions was for the SSN conditions (p<0.05 for both F0 and amplitude). This result provides support for our hypothesis that seeing the interlocutor leads to less pronounced modifications in the production of auditory speech.

3.3. Between and within category examination

Figure 2 shows between- and within-category vowel dispersion as a function of noise type and presentation condition. As represented in the top panel of figure 2, there were no significant differences across the four noise conditions for between-category dispersion. Similarly, presentation condition did not interact with noise type indicating that talkers did not increase the distinctiveness with which they produced the vowels from quiet to noise conditions. Further, whether or not the participants could see their interlocutor had no effect on category dispersion.

In contrast, there was a significant main effect of noise for the within-category dispersion measure (represented in the bottom panel of figure 2) [F(3,9)=4.225, p<0.05, η²=0.585]. Follow-up analyses revealed this effect was primarily driven by differences between the SSN and the CS conditions (p<0.05) – with talkers producing more consistent vowels in the SSN compared to the CS condition. Similar to between-category dispersion, seeing the interlocutor did not interact with the noise type, indicating that differences in the distribution of talkers’ vowel category production did not vary in terms of whether they could or could not see their interlocutor.

4. Discussion

The motivation for the current study was to examine whether the availability of visual speech cues from the interlocutor affected auditory speech production in noise. We hypothesised, that when visual speech cues were available, talkers would make less pronounced modifications to their auditory output compared to conditions where they could not
see their interlocutor. It was also predicted that there would be less variation in speech production as a function of the different noise conditions for FTF compared to AO conditions.

In terms of the effect of seeing/not seeing the interlocutor, the results showed that talkers’ were quieter, and spoke with a lower fundamental frequency for conditions where they could see their interlocutor. Furthermore, this effect of seeing the interlocutor on amplitude and pitch was most pronounced for the SSN compared to the quiet, CS and SMN noise conditions. However, there were no significant differences between FTF and AO conditions for measures of duration, spectral tilt or within- and between-vowel category dispersion.

The finding of a reduction in amplitude for FTF conditions supports the current hypothesis. That is, the availability of visual speech enables interlocutors to perceive each others’ speech without the need to employ effortful strategies such as increasing the amplitude of their speech production. A similar explanation could also account for the finding of reduced auditory modifications for SSN compared to CS and SMN in FTF relative to AO conditions. In CS and SMN conditions, listeners may have little need to utilize the visual speech information from their interlocutor. That is, producing slightly louder speech, or also exploiting temporal glimpses of the partners’ speech signal (e.g. as per [5]) may be sufficient to communicate effectively in CS and SMN conditions. In contrast, for the SSN conditions where the EM of the talkers’ auditory signal is greatest, employing auditory modifications such as increasing the amplitude, reducing the spectral tilt (by adding speech energy to higher frequencies) or varying the temporal production of the speech signal may not be enough to offset the effects of the noise and effectively communicate with the interlocutor. As such, talkers may compensate by relying more heavily on the partners’ visual speech cues to decode the speech signal in energetic maskers such as SSN.

In terms of other indicators of speech production strategies, we examined vowel space dispersion and spectral tilt. For vowel space dispersion there was no evidence that talkers produced more distinctive vowels from quiet to noise conditions (i.e. as measured by between-category dispersion). However, talkers did produce more consistent vowel categories in the SSN compared to the other conditions. This result is similar to that found in [5] who reported a slight tendency toward between-category vowel space expansion for quiet and CS noise, and a strong tendency for more consistent production of vowel tokens (i.e. smaller within-category dispersion). In the current study, the degree of between- and within-vowel space dispersion did not interact with being able to see the interlocutor or not. This finding may indicate that vowel dispersion is useful in both auditory and visual speech modes. That is, research examining ‘clear’ speech (that shares some characteristics of Lombard speech) has demonstrated that the production of distinct and consistent vowel categories is beneficial both auditorily and visually [e.g. 10]. That talkers produced more consistent auditory speech in the SSN may indicate that they were also producing consistent visual speech tokens. In FTF conditions therefore, the availability of visual speech is unlikely to reduce the benefit afforded by producing distinct and consistent vowels. Rather, FTF communication is likely to encourage talkers to produce distinct and consistent speech in noise. Future testing specifically examining both auditory and visual speech modifications in relation to vowel space expansion will test this possibility.

In contrast to previous studies [e.g. 5], there were no main effects for spectral tilt across the noise conditions. That is, the talkers did not appear to shift the distribution of energy from lower to higher frequencies from quiet to noise conditions. However, the discrepancy between the results might be due to measurement differences. For instance, [5] used a linear regression to examine spectral properties whereas in the current study the difference in energy between F1 and F3 amplitudes was used. Furthermore, in the [5], spectral tilt was calculated over the long-term average spectrum (0–8 kHz) across all speech content, but here it was calculated over single vowels (which may have lead to more variable estimates).

The current results found little or no differences between the SMN and CS conditions for the various auditory measures used. This is consistent with the results of [5] who found little difference between the two conditions for the above auditory measures. Importantly, [5] also examined temporal speech production modifications across the noise conditions and found that talkers exploited amplitude modulations in the CS and SMN conditions. At this point we have yet to examine how talkers’ timed their utterances in the various conditions. Furthermore, although the current study only reported on measures of auditory speech, it is also seems reasonable that talkers may alter the saliency of their visual speech production when speaking in face-to-face conditions. For example, a recent study by [4] found a greater degree of lip pinching and pursing during face-to-face communication in conditions of background noise. In the current study, recordings were made of the talkers’ lip, mouth and rigid head movement using a novel head-set and camera set-up, and it will be of interest to examine these data across the different noise and visual conditions.

5. Acknowledgements

We thank the participants and acknowledge support from the Australian Research Council (TS0669874 and DP0666857). We also thank Erin Cvejic for assistance with the analyses.

6. References


