Audiovisual Lombard speech: Reconciling production and perception

Eric Vatikiotis-Bateson, Adrian V Barbosa, Cheuk Yi Chow, Martin Oberg, Johanna Tan, and Hani C Yehia

1Department of Linguistics, University of British Columbia, Vancouver, Canada
2Clinical Audiology, Melbourne University, Melbourne, Australia
3Department of Electronics, Federal University of Minas Gerais, Belo Horizonte, Brazil
evb@interchange.ubc.ca, adriano.vilela@gmail.com, cycho@interchange.ubc.ca,
bythebreeze@gmail.com, jsient@gmail.com, hani@cefala.org

Abstract

An earlier study compared audiovisual perception of speech ‘produced in environmental noise’ (Lombard speech) and speech ‘produced in quiet’ with the same environmental noise added. The results showed that listeners make differential use of the visual information depending on the recording condition, but gave no indication of how or why this might be so. A possible confound in that study was that high audio presentation levels might account for the small visual enhancements observed for Lombard speech. This paper reports results for a second perception study using much lower acoustic presentation levels, compares them with the results of the previous study, and integrates the perception results with analyses of the audiovisual production data: face and head motion, audio amplitude (RMS), and parameters of the spectral acoustics (the spectrum pair).

Index Terms : audiovisual speech, Lombard speech, production and perception links.

1. Introduction

The research described here addresses a mismatch between our investigative methods and how they influence our subsequent understanding of the production and perception of multimodal speech, on the one hand, and the nature of the signals that speakers and hearers actually produce and process, on the other hand.

Although Sumby and Pollack [1] used speech presented in noise (SPIN) to demonstrate the role of visual information in enhancing the intelligibility of degraded acoustic signals, since then the vast majority of stimuli for SPIN tasks have been constructed by adding masking noise to speech signals that were recorded under quiet conditions. Thus, the speech signal presented to perceivers has no correlation with the noise accompanying it because that noise was not present during production. This may, therefore, introduce yet one more distortion in the already strained realism of laboratory test conditions.

This study examines to what extent having a correspondence between the speech and noise signals the so-called Lombard Effect [2] is different from the more commonly used, but uncorrelated, combination of a masking noise signal and speech recorded in quiet conditions. In what follows, the production of speech produced in noise (SPIN) and in quiet (SPIQ), and the effects these two production styles have on auditory (A) and auditory-visual perception (AV) when co-presented with different conditions of noise.

In an earlier presentation of this research [3], it was shown that listeners have greater difficulty recovering words in sentences for speech produced in quiet with party noise added than for two conditions of Lombard speech (see Fig. 2). However, as can be seen in the figure, listeners were able to recover 60% of the words even in the most difficult listening condition where speech recorded in quiet was presented in noise. This is a much higher (easy) listening level than we have used in the past—cf. 30-35% auditory word recovery in studies conducted with Kevin Munhall [4, 5]. Therefore, more stringent (hard) listening levels were set and the perception study was run again.

In what follows, we first compare the results for the easy and hard listening levels. Then we examine the production differences between speech produced in noise and in quiet via the correspondences between visible and audible components of the speech—e.g., between head and/or face motion and RMS amplitude, between face motion and LSP parameters. Finally, we attempt to integrate the production and perception results and say something coherent about auditory-visual speech processing.

2. Methods

2.1. Data Recording

Figure 1 shows the scheme for videotaping sentence recitations (Camera #1) and interactive discourse (Cameras #1-2). For the sentence phase the set of 100 CID Everyday Sentences was recited four times: in quiet and
masking with party noise both with and without blue paper dots stuck on the face and forehead. The female talker faced a 23 inch (15:9 aspect ratio), LCD monitor on which the sentences were displayed one at a time. During noise conditions, the speaker heard pre-recorded party noise (for description, see [6]) through headphones while she spoke.

![Diagram](image.png)

**Figure 1: Generic scheme for experimentation: 1 camera used for sentence recitation; 2 cameras used for conversation, with or without blue dots for motion tracking.**

Close-up high definition recordings using Camera 1 (Ikegami HDK-79E) and a TRAM (TR-50) lapel microphone were made to a SONY HDCAM recorder (Sony M2000). The party noise was recorded to a separate audio channel. During conversational interaction, simultaneous recordings were made (Camera 2) with a Canon XLS-1 DV camera with separate audio feeds from an audio mixer (Mackie XXX). Camera 1 video and audio were transferred in serial digital (HD-SDI) format (1920x1080 pixels) to a digital disk recorder (DOREMI UHD1) without compression. Camera 2 data were transferred in DV (NTSC–DVCpro) format to computer disk.

### 2.2. Data Processing for Perceptual Evaluation

For the perceptual evaluation, only the two sets of sentence recordings without dots were used from Camera 1. The start and end frames of each sentence were identified and stored in playlist format (EDL). The recording was then transferred to DVD format (720x486 pixels) with MPEG 2 compression via a Pioneer PRV-XL1 recorder.

Play lists were constructed such that differences due to presentation order and sentence intelligibility could be tested as well as our primary interest in the effects of noise level and modality (audio, audiovisual). The four noise conditions were

- Speech recorded in noise and presented at the same noise level – SPIN1,
- Speech recorded in noise and presented with the noise boosted – SPIN2,
- Speech recorded in quiet and presented at the same noise level as SPIN1 – SPIQ+N,
- Speech recorded and presented in quiet – SPIQ.

The four noise conditions contained four audio and four audiovisual tokens each and were presented in noise condition blocks. The noise conditions were presented in four presentation orders constructed by rotating through the order once e.g., Spin1, Spin2, SpiQN1, SPIQ =⇒ Spin2, SpiQN1, SPIQ, Spin1 =⇒ and so on. The sentence content of each block was varied for each subject so that each subject got a unique assignment of sentences to the four presentation orders, the four noise conditions, and the two modality conditions – 32 subjects saw 32 sentences for (4x4x2 = 32) conditions.

DVD playlists were made for each of the 32 presentation configurations and played to subjects on a Pioneer V7400 DVD player. Each experimental session was preceded by a block of four practice trials in which subjects could become accustomed to the task, but for which no feedback was given. Subjects viewed the stimuli on a 20" studio monitor (Sony XXX) and heard the stimulus sentences and noise through high quality headphones (Sennheiser 427). After each trial, subjects were recorded repeating what they had heard. This audio file was used for scoring the data.

The experiment was carried out twice at two listening levels using different subjects for each (64 subjects total). For the first, easy listening condition, playback levels of noise and speech were adjusted so that pilot subjects could recover about 50% of the words in the SPIQ+N audio-only (A) condition, which was deemed to be the most difficult. For the second, hard listening condition, the target recovery was 25%. Reliable signal-noise ratios (SNR) for the three conditions with noise (SPIN1, SPIN2, SPIQ+N) were computed for the hard listening level using a sound pressure level (SPL) meter. Due to a calibration error, the measures for the easy listening level could not be used.

### 2.3. Data Processing for Production Analysis

The recordings using dots were subjected to a video-based dot-tracking algorithm [7]. Extracted 2D measures of the 23 dots on the face and forehead were used to identify head and head-corrected face motion for 98 pairs of sentences recorded with and without party noise (2 sentences had to be discarded due to tracking problems).

### 3. Results and Discussion

#### 3.1. Perceptual Evaluation

The data were scored for correct identification of function (verbs, articles, prepositions) and content words (nouns, adjectives, adverbs). The results for the four tokens of each modality-specific condition were summed and the
result subjected to ANOVA in a repeated measures design where order (4x8) was the between factor, and noise (4) and modality (2) were the two within measures. The results for function and content words were effectively identical, therefore statistics are presented for both sets combined in what follows.

3.1.1. Easy Listening Level

The results for Easy Listening are summarized in Figure 2. There was no main effect of presentation order (F[3,28] = 0.71, p > 0.55). There were main effects of Noise (F[3,84] = 24.29, p < 0.0001) and Modality (F[1,28] = 22.63, P < 0.0001). While there were no interactions, Means Comparison tests showed that subjects retrieved more information from the noise-matched signal (SPIN1) than from the noise-unmatched signal (SPIQ+N) in A mode, but these differences were neutralized in the audiovisual condition by the larger visual contribution in the AV mode for SPIQ+N. Note that the probability of correct word identification for each AV condition with party noise was between .75 and .80. Also note that subjects scored about 60% correct in the A mode for SPIQ+N, about 10% higher than intended.

3.1.2. Hard Listening Level

The results for Hard Listening are shown in Figure 3. Again, there was no main effect of presentation order (F[3,28] = 2.01, p > 0.13), and there were main effects of Noise (F[3,84] = 89.16, p < 0.001) and Modality (F[1,28] = 133.41, P < 0.0001). There was a Noise x Modality interaction (F[3,84] = 13.47, p < 0.001). As can be seen in the figure, the intended target of 25% recoverability in A mode was attained for SPIQ+N and SPIN2. Means Comparisons test showed that the AV modality always contributed to the perception of noisy speech. Contrary to the trend observed for Easy Listening, under Hard Listening conditions vision contributed more to word identification when the speech was recorded in noise (SPIN1 vs SPIQ+N, F[1,3] = 2.89, p = .093; SPIN2 vs SPIQ+N, F[1,3] = 9.52, P < 0.003). Means Comparison also showed the AV performance for SPIQ+N to be reliably worse than for either SPIN1 or SPIN2 (SPIN1 vs SPIN2, F = 5.33, p = 0.023; SPIN2 vs SPIQ+N, F = 32.29, P < 0.0001). Adding noise to speech recorded in quiet (SPIQ+N) made only a modest contribution to word recognition scores.

As can be seen in Figure 3, the SNR values computed for the three conditions where speech was presented in noise track the A mode word scores quite closely.

3.2. Mean P(Correct) and Standard Deviation

Figure 4: Relationship between Standard Deviation (σ) and mean probability of correct word identification (P(Correct)) shown for easy and hard listening conditions.
3.1.3. Interpreting the Results

Examining the results for Easy Listening alone would suggest 1) that there may be a ceiling on AV processing of sentences presented in noise at about 80% correct, regardless of their ecological validity. The compressed confidence intervals for AV conditions in Figure 2 support this interpretation. Adding the results for the Hard Listening level further supports a possible ceiling, because the visual modality is capable of more than doubling the word recognition performance so long as the speech presented in noise was also recorded in noise (SPIN1 and SPIN2). Further weight for the ceiling argument is given by the relation between mean word recognition scores and standard deviation (σ) depicted in Figure 4. Generally, variance scales with the mean, but when the results for the two listening levels are combined, the slightly positive relation between σ and Mean P(Correct), that we would expect, changes direction at about P(Correct) = 0.75 and heads for zero as word recognition scores improve.

Although the mean results are compelling, the intelligibility results for specific sentences, shown in Figure 5, suggests caution in at least two areas. First, eight of the 32 sentences were recovered nearly perfectly by all subjects, regardless of listening condition and regardless of how well they were recognized auditorily – compare D2 and D9 as extreme cases. Thus, the apparent ceiling at 80% might be nothing more than a distributional artifact for this set of sentences. Second, despite the overwhelming tendency for sentence intelligibility to be better in AV than in Audio conditions, there are five instances (A3, F3, H10 are the most notable) where the Audio mode was more intelligible than the AV mode.

With these caveats in place, the results of this study are clear on several important points:

- As we would expect, Lombard (SPIN) speech is more intelligible auditorily than quiet (SPIQ) speech that has had noise added to it (SPIQ+N).
- Similarly, but by no means expectedly, visual enhancement of Lombard speech is greater than that of SPIQ+N – compare SPIN2 with SPIQ+N.
- Results are substantially altered depending on what listening levels (SNRs) are used.

What we cannot know from perception testing alone is what it is about the audible and visible behavior under these different conditions that may account for this pattern of results. For this, we must examine the motion of the face and head and their correspondence with the acoustics.

3.2. Analysis of Production Data

Analysis of the face and head motion data should help us distinguish the components of visible motion crucial to recovery of the visual speech information from those that arise simply from generation of clearer acoustic signals under noisy conditions. Following our previous work in this area (e.g., [8]), we expect there to be strong correspondences between face motion and spectral properties of the acoustics and between head motion and fundamental frequency (F0) during production of speech in quiet (SPIQ). We also have seen correspondences of acoustic amplitude (RMS) to both face and head motion. However, to our knowledge no one has computed these correspondences for speech produced in noise (SPIN). In what follows, preliminary results for the analysis and comparison of 98 pairs of sentences produced in quiet (SPIQ) and in noise (SPIN) are presented.

Not surprisingly, Lombard speech is accompanied by larger motions of the head and face (even when corrected
for head motion). Figure 6 shows this on a marker-by-marker basis for one sentence by overlaying variance ellipses (three standard deviations – 3σ) for SPIQ and SPIN productions. That the Lombard speech effect on movement amplitude is not uniform in the face plane can be seen in the differences in elliptical shape of the variance pattern between the head, cheeks, and upper lip (horizontal) and the lower lip-jaw (vertical).

For the analyses currently underway (and presented below), the horizontal and vertical motion for 22 markers was extracted. Since the resulting dimensionality (44) is high and the individual markers co-vary to a large extent, principal component analysis (PCA) was used to reduce the dimensionality to three components of head motion (1 orientation and 2 x-y position) and seven components of face motion (with head motion removed). In addition, the time-varying RMS amplitude of the acoustics was computed, as were the 10 line spectrum pairs derived by autoregression of the output of a 10th-order LPC (linear predictive coding) analysis of the spectral acoustics (for details, see [8, 9]).

We have shown previously that multilinear prediction of face motion from acoustic LSP parameters is better than the reverse (estimation of LSP from face). Here, we estimate the relation between face motion and LSP parameters in both directions (see Table 1). On the other hand, predicting the position and orientation of the head from a single RMS parameter is an ill-posed, one-to-many mapping problem, so these relations are estimated only in one direction – visible face and head to audible RMS amplitude.

Table 1: Mean global correlation results for estimating acoustic parameters (LSP, RMS) from motion of the face and head for SPIQ and SPIN versions of 98 sentences

<table>
<thead>
<tr>
<th>Estimation</th>
<th>SPIQ</th>
<th>SPIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSP =&gt; Face</td>
<td>0.82</td>
<td>0.83</td>
</tr>
<tr>
<td>Face =&gt; LSP</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Face =&gt; RMS</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>Head =&gt; RMS</td>
<td>0.51</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 1 shows estimation of face motion from LSP parameters is better than the reverse as expected. Figure 7 shows the correlations for all 98 pairs of sentences for estimating face motion from LSP parameters. Estimation of RMS amplitude from the face motion PCAs is also good, accounting for 60% of the variance ($r^2$). For the RMS estimation from face and head, there is a small tradeoff when SPIQ and SPIN are compared; the face => RMS estimation is stronger for SPIN than SPIQ, while the head => RMS amplitude estimation is weaker.

More detailed analysis is needed to determine if these differences are meaningful. Such a finding might help rationalize the perception differences observed for SPIQ and SPIN. In noisy speech conditions, the physical correspondences between acoustic and kinematic amplitudes might shift from the head to the speech articulators, and aid multimodal processing. However, this account is limited because the physical correspondence between visible and audible events is largely the same for quiet speech (SPIQ) and Lombard speech (SPIN). Thus, we may have to look elsewhere to account for perceivers’ greater ability to recover visual information from Lombard than quiet speech. Indeed, we may have to analyze the relation between the environmental noise and the speaker’s face and head motion.

4. Conclusions

This study combines analyses of multimodal speech production and perception for two types of speech that have been used in speech-in-noise experiments. While it is clear that listeners can make better use of visual information in Lombard speech conditions, the production data analyses have revealed how this might occur. Lombard speech is louder and is accompanied by bigger motions of the face and head. While the relatively strong correspondence between face motion and the spectral acous-
Figure 7: Pairwise (n=98) comparison of SPIQ (solid) and SPIN (dashed) recordings of the CID Everyday Sentences. Shown is the average correlation coefficient ($r$), per sentence, for estimating RMS amplitude from head motion. Sentence #65, whose marker variances for SPIQ and SPIN productions are compared in Figure 6, is indicated (arrow).

tics does not change, amplitude does not map equally well to face and head for the two types of speech production. We suspect that in order to uncover the linkage between production and multisensory perception, we will need to examine the more complex three-way interaction between the noise, speech acoustics, and visible motions of the face and head.

5. Acknowledgements
Supported by grants from CFI (Canada), NSERC (Canada), and the Telecommunications Assistance Organization (TAO Japan). Thanks to Victor Chung, Kevin Lutz, Nicole Mirante, Kevin Munhall, Jolien Otten, and Rebecca Weiss who assisted in this work.

6. References


