Perceptual processing of audiovisual Lombard speech

Chris Davis[1], Amanda Sironic[2] and Jeesun Kim[1]
[1] MARCS Auditory Laboratories, UWS, Sydney; [2]Department of Psychology, The University of Melbourne.
chris.davis@uws.edu.au, a.sironic@pgrad.unimelb.edu.au, j.kim@uws.edu.au

Abstract

Seeing the talker improves the intelligibility of speech degraded by noise (a visual speech enhancement effect). This experiment examined whether this enhancement is greater when the speech signals were recorded in noise compared to when they were recorded in quiet. Ten sentences were spoken by four people either in-quiet or whilst they were listening to cocktail party noise (in-noise). The visual speech of these talkers was measured using 24 facial markers and associated video clips filmed. Sixty participants were tested in a speech-in-noise identification experiment under four different conditions. Visual speech enhancement was measured by mean percent words correctly identified in the audiovisual minus auditory-only condition scores. After ceiling effects were curtailed the results showed that the audiovisual enhancement for speech signals recorded in noise (43%) was significantly greater than that for speech recorded in quiet (30%).

1. Introduction

It has long been known that the provision of visual speech information (speech associated movements of the lips, mouth, jaw, face and head) enhances the intelligibility of degraded acoustic speech signals (Sumby & Pollack, 1954). What is interesting is that in many studies investigating this audiovisual (AV) enhancement have used speech signals presented in noise that were recorded under quiet conditions (e.g., Grant & Seitz, 1998). Given this, the audiovisual speech signals presented to perceivers in noise would not correspond with signals actually produced in noise.

It is also known that both auditory and visual speech signals are different (exaggerated) when produced in noise (Kim, Davis, Vignali & Hill, 2005). Given these differences, AV speech signals might be more intelligible in noise when also produced in noise. This is particularly because for such signals (compared to ones produced in quiet) there appears to be a greater degree of correlation between the speech-related aspects of the auditory and visual signals (e.g., RMS energy in the second formant band and mouth movement, Davis, Kim, Grauwinkel & Mixdorff, 2006). This greater bimodal linkage might afford better recovery of the auditory signal from the accompanying visual speech.

The current study investigated whether there will be greater AV enhancement in a speech-identification-in-noise (SPIN) task when the speech signals presented in noise were generated in noise (compared to those generated in quiet). To test this, we had four talkers utter sentences in quiet and noise. The change in the speech signal from in-quiet to in noise was quantified by measuring both the auditory and visual speech signals. Following this, a SPIN task was conducted to determine if these different production conditions produced different sized AV enhancement effects.

2. Methods

2.1 Participants

Movement data: Four talkers participated in the data capture session (3 males, 1 female). All were native speakers of English (one British, two Australian and one American). Ages ranged from 32 to 54 years.

Perception experiment: Sixty undergraduate students from the University of Melbourne participated in the for course credit. None of the participants reported any hearing loss and all had normal or corrected-to-normal vision. All were native speakers of English.

2.2 Materials and apparatus

The materials were 10 sentences selected from the 1969-revised Harvard list of phonetically balanced sentences (IEEE, 1969). The background noise consisted of a commercial babble track (www.auditec.com). Two linked Northern Digital Optotrak machines were used to record the visual speech movement data. The

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

configuration of the markers on the face and headrig is shown in Figure 1.

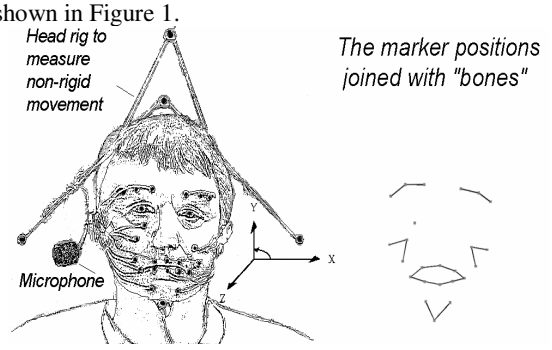

Figure 1. Position of the 24 facial markers and the four rigid body markers (on the extremities of the head-rig). Also shown are the directions of the XYZ axes. The marker positions are shown in the right side of the figure (joined with lines to show the mouth, chin, cheeks and eyebrows) for the neutral position.

2.3 Data processing

Non-rigid facial and rigid head movement was extracted from the raw marker positions. The data were recorded at a sampling rate of 60Hz. Each frame was represented in terms of its displacement from the first frame and Principle Component (PC) analysis was used to reduce the dimensionality of the data. Discrete data were fitted as continuous functions using B-spline basis functions (so called functional data, see Ramsay & Silverman, 1997). This process of converting discrete data to trajectories required that all instances were of the same time and this was achieved by a reversible time alignment procedure (with time differences recorded). The characteristic shape of the data was maintained over the time warping procedure by using manually placed landmarks that were then aligned in time (the beginning and end of each curve were also taken as landmarks).

2.4 Motion capture procedure

Recording: Each session began with the placement of the movement sensors (see Figure 1) during which time the talkers were asked to memorize the 10 sentences to be spoken. Each talker was recorded individually; they were seated in an adjustable dentist's chair and asked to say each sentence to a person directly facing them at a distance of approximately 2.5 m. In the Quiet condition the talker spoke each sentence in quiet; in the Noise condition, the talker wore a set of earplugs and heard the babble track as they spoke each sentence. In addition to motion capture, the audiovisual speech of four talkers was recorded using a high-quality video camera (Sony HDCAM HKDW-702). These video clips were used in the perception study.

2.5 Perception test procedure

The video clips (compressed to MPEG 2) were played back using the DMDX software (Forster & Forster, 2003) on a ViewSonic G810 21 inch monitor with the audio component presented through a pair of Sennheiser HD580 precision headphones.

There were four different conditions in the experiment: Auditory-Only Quiet; Auditory-Only Noise; Audiovisual Quiet and Audiovisual Noise. These conditions were constructed in the following fashion. First the audio tracks from the speech in Quiet and in Noise videos were stripped off normalized to 60 dB and mixed with babble speech at various Signal-to-Noise Ratios (SNRs). The SNRs for each Audio-Only conditions were determined in a pilot experiment so that the percentage of correct responses reported in the Auditory-Only conditions (Quiet and Noise) would be similar and well away from ceiling. As it turned out, speech produced in noise was very difficult to mask by babble. This was due to the Lombard effect (Lombard, 1911) in which speech produced in noise has a higher fundamental frequency (F0) than that produced in quiet (e.g., Davis, et al, 2005). In order to counter this effect and appropriately mask the auditory signal, the F0 of the babble track was raised to approximately match that of the speech produced in noise. Once the appropriate in-Quiet and in-Noise Auditory-Only tracks were created, copies of these were then dubbed back to the corresponding Videos to create the and in Noise Audiovisual conditions.

Conditions within the experiment were presented blocked (see Table 1). The initial block consisted of Auditory-Only presentation in which five sentences for one talker (recorded in quiet) and five sentences from another talker (recorded in noise) were presented. The second block consisted of Auditory-Only presentation of the remaining five sentences spoken by the talkers in Block one. The third block consisted of Audiovisual presentation of five items from a different talker (recorded in quiet) and five items from another talker (recorded in noise). The fourth block consisted of 12 Audio-Only filler items. The fifth block consisted of 12 Audiovisual filler items. The last block consisted of Audiovisual presentation of the remaining five sentences of the two talkers in Block three. The presentation of items within any block was randomized for each participant. Four versions of the experiment were prepared such that a specific item (as spoken by a particular talker) would not be repeated in any version. Each version consisted of a set of practice items that familiarized participants with the conditions to be presented in the experiment.

Participants were tested individually in a sound-attenuated booth and were asked to identify as many

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

words as they could and type these on the keyboard. In scoring these data, credit was given only if the typed word exactly matched the spoken word (except where the response was an obvious typo).

		Talker1	Talker2	Talker3	Talker4
Auditory Only	1st Blocks	Snt* 1-5 in quiet	Snt 6-10 in noise		
	2nd Blocks	Snt 6-10 in quiet	Sent 1-5 in noise		
Audio-visual	3rd Blocks			Snt* 1-5 in quiet	Snt 6-10 in noise
Fillers	4th Blocks	Audio Only fillers			
	5th Blocks	Audiovisual fillers			
Audio-visual	6th Blocks			Snt 6-10 in quiet	Sent 1-5 in noise

* Snt = sentences

Table 1. An example of the setup of one version of the speech perception experiment. Note that there were four versions of the experiment such that a specific item (as spoken by a particular talker) would not be repeated in any version.

3. Results

Auditory Data. Analysis of the auditory data indicated that there was a Lombard effect. For instance, speech produced in the noise condition was louder than that produced in quiet. On average the size of this effect was 11 dB. The length of the renditions also increased for the Noise condition compared to Quiet with an average increased production time of 290 ms.

Movement data. In order to reduce the amount of data resulting from motion capture a principal components (PC) analysis was used. In order to visualize each PC, its trajectory over time was calculated in reference to a neutral face condition (see the right side of Figure 1). This resulted in a curve of the change in the contribution of the PC over time.

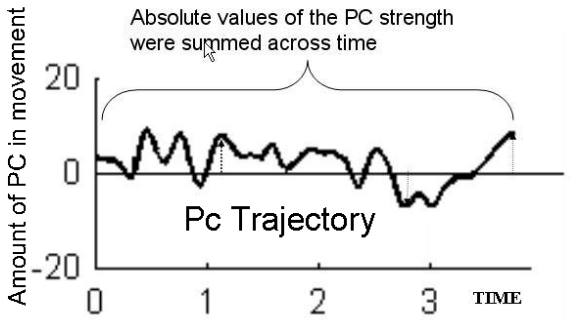

Figure 2. The absolute value of each PC over time was summed to create an index of that PC's contribution (used in Figure 4).

The total power of the PC's contribution was calculated by taking the sum of absolute values over time (Figure 2 above).

These data were then visualized in terms of each marker position using a TclTk graphic interface (see Figure 3). The movement of PC1 can be glossed as jaw motion (with mouth movement); PC2 as mouth/lip rounding/opening and eyebrow raising (without jaw motion); PC3 as head translation towards the listener; PC4 as lip protrusion; PC5 as mouth opening and eyebrow closure.

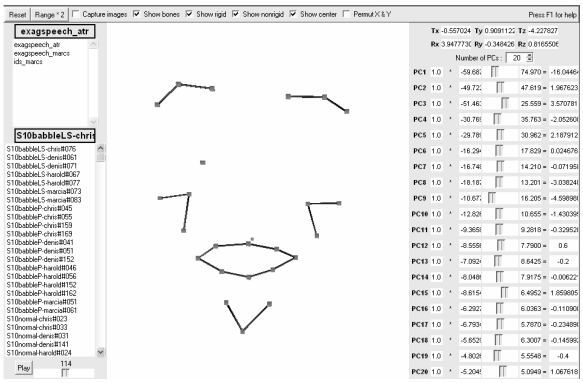

Figure 3. The TclTk Viewer that enables each PC to be visualized.

The summed absolute values of amount of change (speech produced in quiet compared to in noise) for these first 5 PCs for the four talkers are shown in Figure 4. As can be seen there was considerable variation across talkers in how the visual speech PCs changed from talking in quiet to in noise (Talker 1 showed the greatest change for PC1; Talker 2 for PC2 and so on).

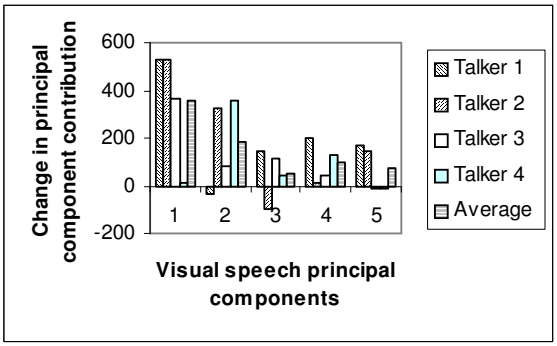

Figure 4. The change in the size of PCs 1-5 (relative to the quiet condition) for the four talkers.

Perceptual data. The audiovisual enhancement for the Quiet and Noise conditions (expressed as mean percent words correctly identified in the Audiovisual minus

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Auditory Only condition scores) as a function of each talker is shown in Figure 5. The key result to be examined is whether there was more visual speech enhancement of auditory identification in the Noise versus the Quiet condition. As it turned out, the interaction between Visual enhancement and Noise versus Quiet was not significant [F2 (1,36) = 2.76, p > 0 .05]

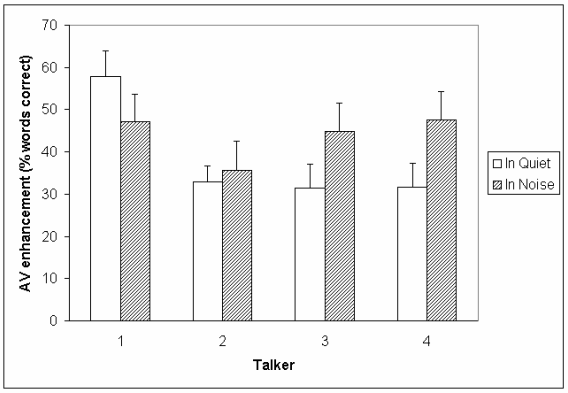

Figure 5. Audiovisual enhancement (expressed as mean percent words correctly identified in the Audiovisual minus Auditory Only condition scores) for the Quiet and Noise conditions across talkers.

However, as can be seen, the pattern of Talker 1 is different to that of the other talkers. Indeed, there was a significant interaction between the variables of Talker, Visual enhancement and Noise versus Quiet, F2 (3,36) = 3.62, p < 0.05. In order to understand why the AV enhancement scores for Talker 1 were so different from that of the others, it is necessary to examine how these scores were determined. As mentioned above, AV enhancement scores for the Quiet and Noise conditions were determined by subtracting the mean correct scores in the appropriate AO conditions (Quiet or Noise) from the corresponding AV scores. That is, the AO conditions act as the baselines against which the AV enhancement effects are measured. This means that the size of any potential enhancement score might be influenced by the magnitude of the relevant baseline score.

If possible, AO baselines across all talkers should be equivalent in the Quiet condition and in the Noise condition. If this were so, one could conclude that there were no effects due to baselines. However, as it turned out baselines differed across talkers both in Quiet and Noise. In the Quiet condition, Talkers 1, 2, 3 and 4 have baselines (AO scores) of 21.77%, 44.25%, 42.10% and 42.70%, respectively. One reason for why the pattern of Talker 1 might be different from those of the other talkers is that there was a ceiling effect in the percent correct identification for the Audiovisual condition for this talker. That is, Talker 1 has the lowest AO baseline of 21.77% and the highest percentage correct score of 79.67%. Thus, Talker 1 showed a very large AV enhancement effect of approximately 58% in the Quiet condition (all other talkers showed an AV enhancement effect of around 30%). The AO baseline for talker 1 in Noise was 38% correct; for a larger AV enhancement effect to be shown for this talker, participants would have to score greater than 100% accuracy (clearly an impossibility). If ceiling effects are curtailed (for all talkers) by analyzing only those items that had AV (in-quiet) scores of less than 80% correct, then there is a significant Visual enhancement and Noise versus Quiet interaction, [F2 (1,17) = 7.78, p < 0 .05].

3.1 Does the visual speech movement data relate to the perception data?

As it turned out, there were considerable cross talker variation in the size of visual speech enhancement and in the amount of visual speech movements (summed PC contribution). Given this, it might be the case that particular aspects of the visual movement data specifically relate to the perception data. To test this, first, AV enhancement in quiet was considered. There are two speech movement PCs that are mostly likely important indices of visual speech (PC1, jaw motion and PC2, mouth/lip rounding). The size of the contribution of these for each of the talkers is shown in Figure 6.

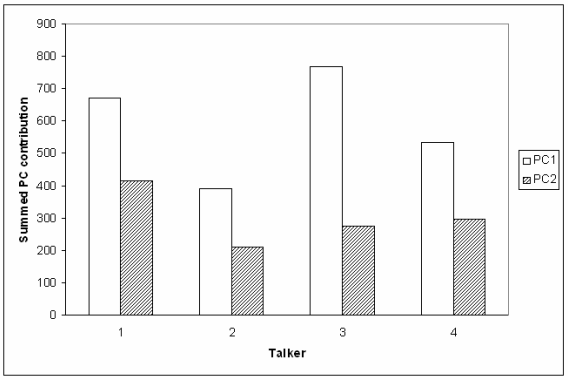

Figure 6. The power of the contribution of PC1 and PC2 for each talker in quiet.

The amount of visual enhancement for the Quiet condition is shown in Figure 5 (unfilled bars). As can be seen, the largest enhancement was for Talker 1 with the other talkers having similar, smaller scores. For visual speech movements (indexed by PC1 and PC2, see Figure 6), the same relative pattern (large for Talker 1, similar sized smaller scores for the other talkers) was shown by PC2 (filled bars, mouth/lip rounding). This suggests that the absolute amount of mouth and lip

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

rounding was an important factor in determining the size of the visual enhancement effect for sentence produced in quiet. Indeed there were significant correlations between the size of the AV enhancement effect for the Quiet productions and PC1 (r = 0.33) and PC2 (r = 0.47). No other significant correlation between AV enhancement and movement PCs were found.

The size of the contribution of PC1 and PC2 for each of the talkers in noise is shown in Figure 7. The In-Noise visual enhancement (Figure 5, filled bars) does not appear to be accounted for by the relative size of cross talker PC2 movement. Of course, as mentioned above, there was a ceiling effect (at least for Talker 1) and this might have obscured the relationship. However, even if the data from Talker 1 is ignored, it is clear from Figure 7 (filled bars) that the contribution of PC2 is greatest for Talker 2, yet this talker actually showed the smallest audiovisual enhancement effect (see the filled bars in Figure 5).

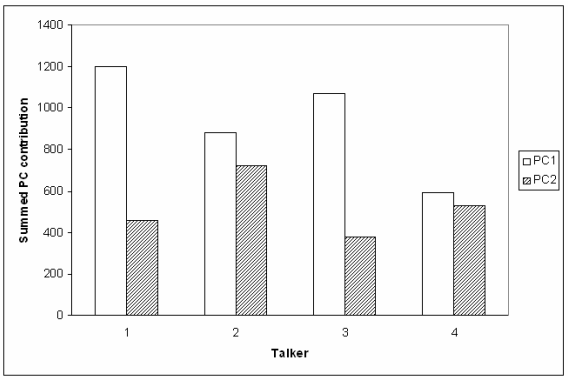

Figure 7. The power of the contribution of PC1 and PC2 for each talker in Noise.

4. Discussion

The results of perceptual experiment showed that the visual speech enhancement of a degraded auditory signal is greater when that visual speech was produced in noise compared to in-quiet. This finding suggests that hearers make use of speech signals that are relevant to listening conditions. That is, just as Lombard audio speech assists speech comprehension in noise, so too does Lombard visual speech.

It is not clear from the results which particular properties of visual speech might be responsible for this additional enhancement effect for speech produced in noise. Whereas the degree of visual enhancement of speech produced in quiet appeared to be related to the degree of lip and mouth rounding (PC2), this was not the case for speech produced in noise. That is, the talker who showed the largest contribution of PC2 (Talker 2) in their production showed the smallest visual enhancement effect. It might be that the visual facilitation from Lombard visual speech has a more complex causation which requires further investigation.

5. Acknowledgements

The authors wish to acknowledge that this work was supported by the Australian Research Council. We would like to thank Harold Hill, Guillaume Vignali, Denis Burnham, Kevin Munhall, Marcia Riley and Kuratate Takaaki for their help in conducting the Optotrak recordings.

6. References

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentence. The *Journal of the Acoustical Society of America*, *104*, 2438-2450.

Davis, C., Kim, J. Grauwinkel, K., & Mixdorff, H. (2006). Lombard speech: Auditory (A), Visual (V) and AV effects. In Hoffmann, Rüdiger and Mixdorff, Hansjörg (Eds). Proceedings of the 3rd International Conference on Speech Prosody, Dresden, Germany. TUDPress. Verlag der Wissenschaften GmbH: Dresden.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers*, *35*, 116-124.

Kim, J., Davis, C., Vignali, G., & Hill, H. (2005). A visual concomitant of the Lombard reflex. In E. Vatikiotis-Bateson, D. Burnham, & S. Fels (Eds.). Proceedings of Auditory Visual Speech Processing 2005 (pp 17-22), Vancouver, Canada, Casual Productions.

Lombard, E. (1911). Le signe de l'elevation de la voix. *ANN. MAL. OREIL. LARYNX*, *37*, 101-199.

Ramsay J. O., & Silverman, B. W. (1997) *Functional Data Analysis*. Springer

Sumby, W. H., & Pollack, I (1954). Visual contribution to speech intelligibility in noise" *Journal of the Acoustical Society of America*, 26, 212-215.