Lombard speech: Auditory (A), Visual (V) and AV effects

Chris Davis[1], *Jeesun Kim*[1, 2], *Katja Grauwinkel*[3], *Hansjörg Mixdorff*[3]

[1] Department of Psychology, The University of Melbourne, Australia
[2] Graduate School of Education, Sejong University, Korea
[3] TFH Berlin University of Applied Sciences, Germany

Abstract

This study examined Auditory (A) and Visual (V) speech (speech-related head and face movement) as a function of noise environment. Measures of AV speech were recorded for 3 males and 1 female for 10 sentences spoken in quiet as well as four styles of background noise (Lombard speech). Auditory speech was analyzed in terms of overall intensity, duration, spectral tilt and prosodic parameters employing Fujisaki model based parameterizations of F0 contours. Visual speech was analyzed in terms of Principal Components (PC) of head and face movement. Compared to speech in quiet, Lombard speech was louder, of longer duration, had more energy at higher frequencies (particularly with babble speech) and had greater amplitude mean accent and phrase commands. Visual Lombard speech showed greater influence of the PCs associated with jaw and mouth movement, face expansion and contraction and head rotation (pitch). Lombard speech showed increased AV speech correlations between RMS speech intensity and the PCs that involved jaw and mouth movement. A similar increased correlation occurred for intensity and head rotation (pitch). For Lombard speech, all talkers showed an increased correlation between F0 and head translation (raising and lowering). Increased F0 correlations for other head movements were more idiosyncratic. These findings suggest that the relationships underlying Audio-Visual speech perception differ depending on how that speech was produced

1. Introduction

The Lombard effect (or Lombard reflex/sign) refers to the tendency for a person to increase the loudness of his/her voice in the presence of noise [1]. Compared with speech in quiet, speech in noise is typically produced with increased volume, decreased speaking rate, and changes in articulation and pitch [2]. It has been suggested that such changes lead to improved speech communication in noisy environments. The Lombard effect occurs in children [3], macaques [4] and birds [5]. In most instances the Lombard effect occurs reflexively; it is not diminished by training or instruction [6; although see 7].

Studies on the Lombard effect have concentrated on describing the changes that occur in the auditory signal. In this paper, we follow-up the approach of [8] and consider changes that also occur in the visual correlates of speech articulation. This study is important for describing speech in a variety of natural conditions and particularly for characterizing changes in the relationship of AV speech signals.

2. Method

Participants. Four people participated in the experiment (3 males, 1 female). All were native speakers of English (one British, two Australian and one American); ages ranged from 32 to 54 years.

Materials. The materials were 10 sentences selected from the 1965 revised list of phonetically balanced sentences (Harvard Sentences, [9]).

Noise. Two types of background noise were employed, multi-talker babble and white noise. A commercial babble track (Auditec, St. Louis, MO) was used; the white noise was generated at the same average RMS as the babble track. The noise was presented to participants either through ear plugs or through two loud speakers (Yamaha MS 101-II). The conditions will be referred to as **babbleP**; **babbleLS** for the babble noise ear plug and loud speaker conditions and **whiteP** and **whiteLS** for the white noise plug and loud speaker conditions.

Movement capture. Two Northern Digital Optotrak machines were used to record the movement data.

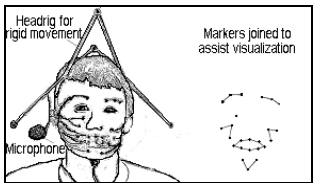

Figure 1. Position of the 24 facial markers and the four rigid body markers (on the extremities of the head-rig).

The configuration of the markers on the face and head rig is shown in Figure 1. Sound was captured from both a head-mounted (Share SM12A) and a floor microphone (Sennheiser MKH416P48U-3). Video was captured using a digital camera (Sony HDCAM HKDW-702).

Procedure Each session began with the placement of the movement sensors (see Fig 1) during which time participants were asked to memorize the ten sentences to be spoken. Each participant was recorded individually in a session that lasted approximately 90 minutes. Participants were seated in an adjustable dentist chair in a quite room and were asked to say

aloud ten sentences (one at a time) to a person who was directly facing them at a distance of approximately 2.5M. The participant then repeated the ten sentences. This basic procedure was repeated once for each speech condition. These conditions consisted of the participant speaking while hearing multi-talker babble through a set of ear plugs (at approximately 80 dB SPL); hearing the same babble through two Loud Speakers; hearing white noise through ear plugs (at the same intensity); hearing white noise through the loud speakers (participants also whispered the sentences at a level judged loud enough for the conversational partner to hear, we will not consider this condition here).

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 was fitted as continuous functions using B-spline basis functions (so called functional data, see [10]). This process of converting discrete data to trajectories required that all instances were of the same time, 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).

The acoustic analyses were carried out on the earplug conditions only. F0 contours were calculated at 10 ms intervals using the *Praat* [11] default pitch estimation. Contours were checked and corrected within the *Praat PitchEditor*. Fujisaki parameters were estimated automatically [12] and if necessary corrected using the interactive *FujiParaEditor* [13]. Time constants α and β were generally set to 2/s and 20/s, respectively. *Fb* was 95, 96, 76 and 144 Hz for the different subjects. Formant frequencies and energy were analyzed on the hand segmented vocalic portion of the signal using Praat for only one participant. For the correlation analysis, the auditory and visual data were aligned by fitting a cubic smoothing spline to the auditory data (either RMS energy or interpolated F0 curve) which was then resampled to precisely match the time of the visual data; correlation coefficients were calculated as the zeroth lag of the normalized covariance function. For spectral-energy by movement correlations, RMS energy was band-passed based on formant frequency values. This correlation was only performed for one participant.

3. Results

Auditory analysis: Analysis of the auditory data indicated a Lombard effect: Speech in noise was louder than that produced without, [$F(1,78) = 365.41$, $p < 0.05$]. The average size of this effect was 11 dB. The length of the renditions also increased for the in-noise conditions compared to the no-noise condition, [$F(1,78) = 33.4$), $p < 0.05$] with an average increased production time of 290 ms. Formant frequencies tended to rise in Lombard (**babbleP** and **whiteP**), with the largest increase in F3 (see Figure 2).

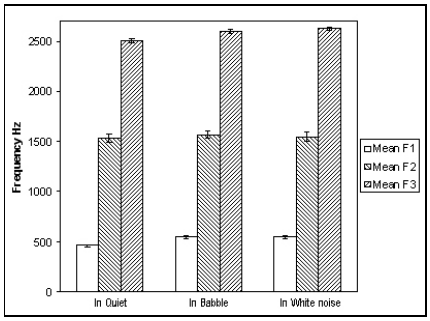

Figure 2. *Average formant frequencies (F1, F2 and F3) for the In-quiet, Babble earplug and White-noise earplug conditions (for one participant).*

The main effects of Formant and Speech Condition were significant [$F(2,260) = 1453.9$, $p < 0.05$; $F(2,26) = 39.17$, $p < 0.05$, respectively]. There was also a significant interaction between these effects [$F4, 260 = 6.9$, $p < 0.05$]. Lombard speech was relatively more intense in the upper speech formants (see Figure 3).

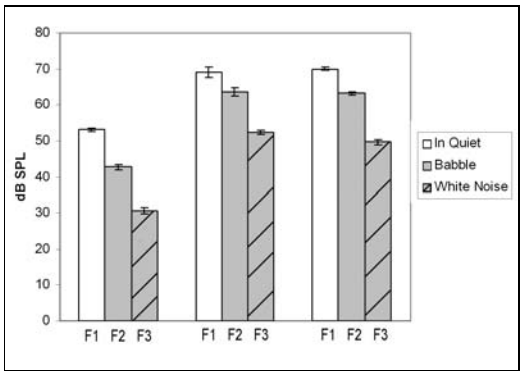

Figure 3. *Average formant energy (dB) as a function of Speech Condition (for one participant).*

This intensification was more marked for the **babbleP** Plug condition in F3 compared to the **whiteP** condition. All four speakers showed higher mean fundamental frequency values in the noise conditions compared to the In-Quiet condition. Compared to In-Quiet, all four speakers exhibited larger ranges of fundamental frequency in the **babbleP** condition and three of them had even larger range expansions in the **whiteP** condition (see Table 1).

Table 1: *Mean fundamental frequency (F0), standard deviation (sd), and range of F0 each speech condition for all four speakers.*

S	mean F0 (sd)			range F0		
	Quiet	babbleP	whiteP	Quiet	babbleP	whiteP
1	136.5 (23.8)	194.03 (26.3)	195.8 (27.6)	105.7	129.7	142.1
2	141.1 (22.7)	177.1 (25.4)	200.9 (28.6)	96.4	121.5	142.5
3	115.1 (26.0)	210.8 (45.2)	179.9 (46.3)	104.0	179.6	179.1
4	183.5 (26.5)	272.3 (44.3)	302.5 (53.4)	105.0	188.3	222.3

Table 2 provides a summary of accent and phrase commands for the four talkers. Overall the mean accent command amplitudes (*Aa*) are significantly lower in the In-Quiet condition compared to the other conditions. This effect can be observed when comparing the accent command amplitudes in Figure 6. Displayed are utterances of the female speaker in the two communicative environments: **Quiet** condition at the top and **babbleP** at the bottom. Each example displays from top to bottom: speech waveform, F0 contour (extracted: +-signs, model-based: solid line), word labels, and amplitudes *Ap* and *Aa* of phrase and accent commands underlying the *F0* contour. Mean phrase command amplitudes (*Ap*) are on average twice as high in the conditions **babbleP** and **whiteP**. The mean frequency of accent commands and phrase commands (rightmost columns - commands per second) were similar across the speech conditions.

Table 2: *Mean, standard deviation (sd) and total number (N) of accent commands Aa, as well as phrase commands Ap for the four speakers, total duration of speech material (Dur) as well as frequency of accent and phrase commands expressed as commands per second (cmds/s).*

Type	**Aa** mean/s.d.	**Ap** mean/s.d.	**Dur.** [s]	**Acc. cmds/s**	**Phr. cmds/s**
In-Quiet *N*	.30/.16 *309*	.43/.20 *110*	159.1	1.94	0.69
babbleP *N*	.40/.19 *612*	.85/.31 *231*	318.9	1.92	0.72
whiteP *N*	.43/.20 *357*	.80/.31 *139*	195.9	1.82	0.71

Visual analyses: In order to parameterize the contribution of the PCs, absolute values of each PC were summed to represent the amount that each PC contributed to head and face movement over time (see Fig 4). These data ("PC strength") were used as the dependant measure in a series of ANOVAs to determine whether there were differences in the amount of movement across speech modes, sentences and persons.

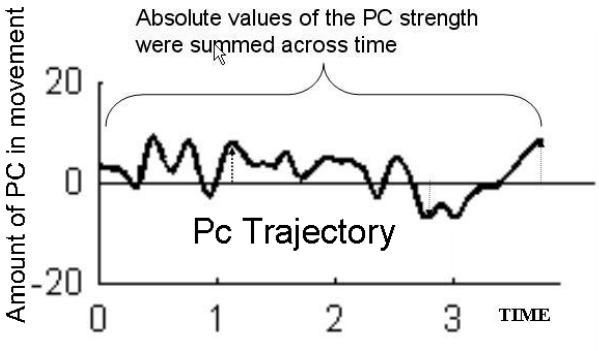

Figure 4. *The absolute value of each PCs contribution across time was summed to provide a PC strength measure.*

The analyses use the first 6 PCs since visualization of these showed that they largely represented specific head and face movement. That is, PC1 was jaw motion (and mouth opening); PC2 as mouth opening and eyebrow raising (without jaw motion), PC3 as head translation (towards the hearer), PC4 as lip protrusion, PC5 as mouth opening and eyebrow closure (cf PC2) and PC6 as rotation (pitch). Approximate 90% of the variance was captured by the first 6 PCs. In terms of these PCs, the effect of talking in noise (as measured from the In-Quiet condition) can be characterized as consisting of a marked increase in jaw (PC1) and mouth motion and eyebrow expansion (PC2), increases both in lip protrusion (PC4), mouth and eyebrow closure (PC5) and pitch head rotation (PC6). There was however no change in the translation of the head (PC3). Figure 5 (top) shows the increase in size of first 6 PCs for speech in noise (relative to each of the associated PCs for the In-Quiet condition).

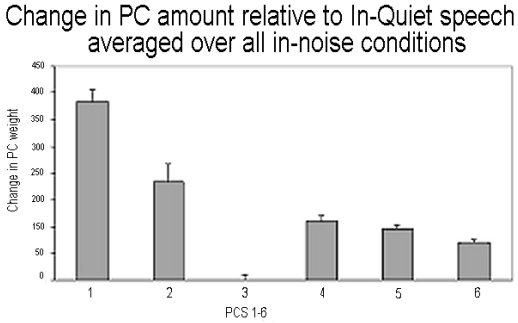

Figure 5. *The change in size of PCs 1–6 (relative to the In-Quiet condition) as an average of the four different speech-in-noise conditions*

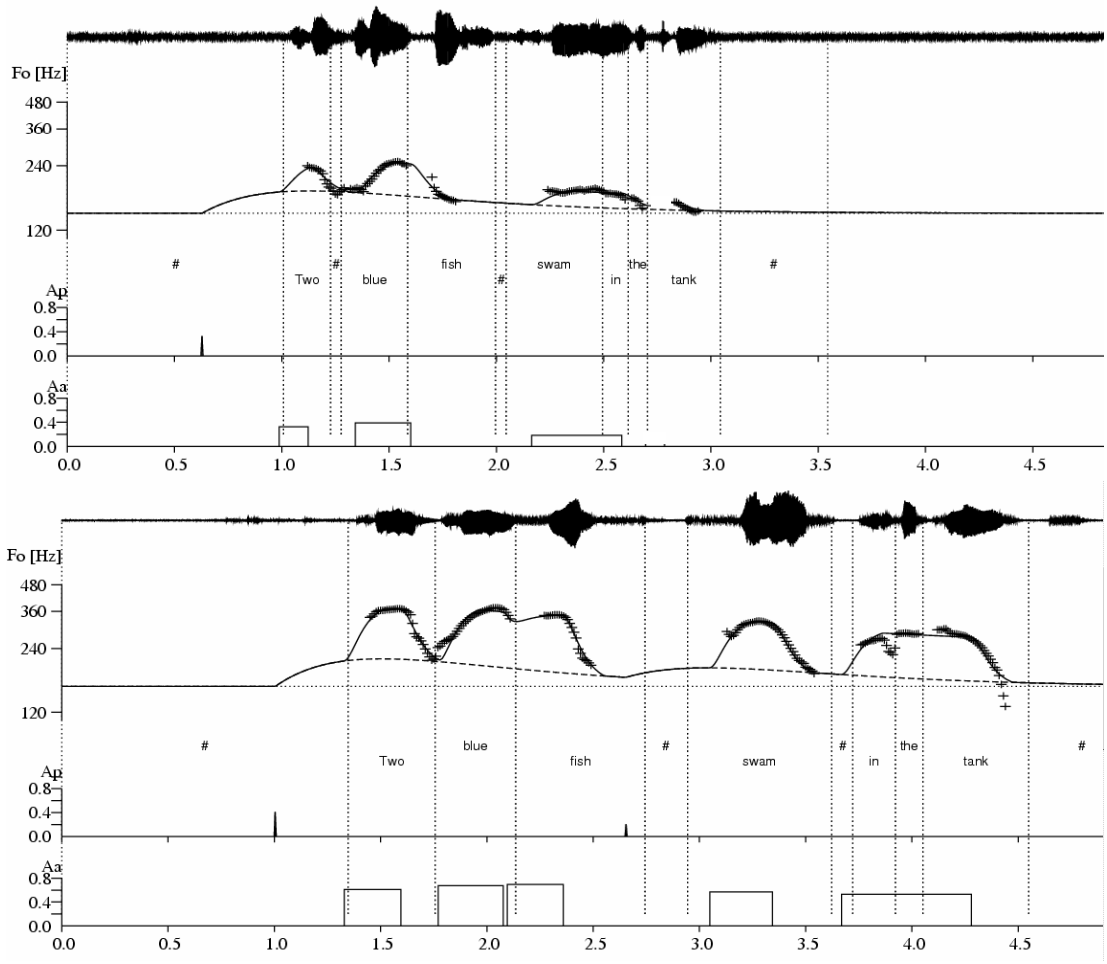

Figure 6**:** ***Two examples of the sentence "Two blue fish swam in the tank" spoken by the female speaker. Each example displays: Speech signal, F0, word labels, phrase commands, and accent commands. Top panel: normal, bottom panel: babbleP.***

The four speech-in-noise conditions were analyzed separately to determine if the type of noise affected the amount (main effect) and constitution (interaction effect) of head and face movement. The analyses indicated that there was a main effect of noise-type [$F(3,585) = 3.82$, $p < 0.05$] and that this interacted with the effect of the different PCs [$F(15, 585) = 2.86$, $p < 0.05$]. ANOVAs were conducted with noise-type and the mode of noise delivery (Plug vs. Loud Speaker) as factors to determine whether these properties altered the pattern of head and face movement. Analysis revealed that there was a significant effect of noise-type [$F(1,195) = 4.41$, $p < 0.05$, with the sum of PCs 1-6 being larger for babble noise than for white noise] and a significant effect of the mode of noise delivery (Plug vs. Loud Speaker) [$F(1,195 = 4.2$, $p < 0.05$, with the sum of PCs 1-6 being larger for the Plug than for Louder Speakers]. There was also an interaction between these effects and the pattern of PCs. For noise-type x PCs [$F(5, 195) = 4.46$, $p < 0.05$] and noise-delivery x PCs [$F(5,195) = 2.25$, $p = 0.05$].

Auditory-Visual analysis: The following analyses examined the relationship between the RMS of the auditory signal for wide band (WB) and two frequency sub-bands (in the F1 and F2 range) for PCs 1–6 for the different in speech conditions. These analyses were conducted using the data from a single participant (male, aged 45). Previous studies have shown that the correlation of auditory RMS and mouth movement change as a function of speech frequency band (e.g., [14]). To examine this, the auditory stimuli were band-pass filtered into two speech bands that should contain F1 (100-800 Hz) and F2 energy respectively (800-2200 Hz). Figure 7 shows the average correlation coefficients for F1 and F2 RMS energy and PCs 1-6 for In-Quiet speech and for an average of the speech in noise conditions.

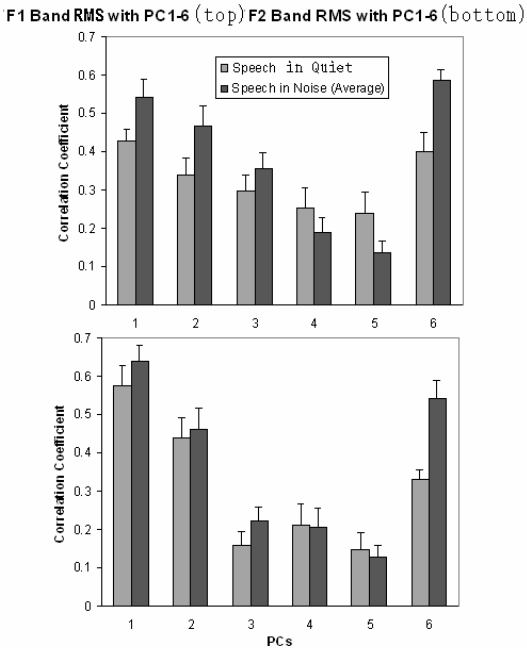

Figure 6. Average correlation coefficients between F1 band RMS (top) and F2 (bottom) and PCs1-6 for the different speech conditions

The general pattern of correlations shown for WB, F1 and F2 RMS band speech and the movement PCs were similar. However, the correlations for the first two PCs for the F2 data were considerably higher than those for the corresponding PCs in WB or F1 band signals.

In addition to examining correlations between spectral energy and visual speech we also investigated how F0 correlates with rigid head movements and the effect of Lombard speech. Three types of head movement were examined: translation in the Z and Y axes and rotation in the X axis (pitch). For this analysis only the In-Quiet and Babble-Plug conditions were contrasted. In general, the correlations were only weak to moderate and although some correlations increased for Lombard speech this was not a systematic effect. For example, although for two participants there was an increased correlation between F0 and head rotation in the X axis (pitch) for Lombard speech ($r = 0.31$ and 0.43, for the In-Quiet and Babble Plug conditions respectively), the other two participants did not show this effect. Interestingly, all participants showed an increased correlation between head translation in the Z axis and F0 for Lombard speech ($r = 0.31$ and 0.41, respectively). Only one participant showed an increased correlation between head translation in the Y axis and F0 for Lombard speech ($r = 0.29$ and 0.73, respectively).

4. Discussion

Lombard speech is louder, slower and more intense in the upper formants than speech in quiet. The analysis of accent commands has shown that all four speakers differ considerably between no-noise and speech-in-noise communicative environment. In terms of mean accent and phrase commands, the effect of talking in noise can be characterized as an increase in both amplitudes. Lombard visual speech has more mouth and jaw and rigid head movements than quiet speech. Lombard speech induced by babble noise was more intense in the F3 region than for the white noise inducer. Likewise, speech produced in babble tended to have more jaw and mouth movement than that produced in white noise. Lombard speech seems intended to aid communication in noise and the greater coupling between the AV signals for Lombard speech suggests that visual speech also plays a role.

5. References

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

[2] Junqua J-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. Speech Communication, 20, 13-22.

[3] Amazi DK, Garber SR. (1982). The Lombard sign as a function of age and task. Journal of Speech and Hearing Research, 25, 581-585.

[4] Sinnott, J. M., Stebbins, W.C. and Moody, D.B. (1975). Regulation of voice amplitude by the monkey. Journal of the Acoustical Society of America, 58, 412-414.

[5] Potash, L. M. (1972). Noise-induced changes in calls of the Japanese quail. Psychonom Science, 26, 252-254.

[6] Pick , H.L., Siegel, G.M., Fox, P.W., Garber, S.R. and Kearney, J.K. (1989). Journal of the Acoustical Society of America, 85, 894-900.

[7] Tonkinson, S. (1994). The Lombard effect in choral singing. Journal of Voice, 8, 24-29.

[8] Kim, J. et al (submitted)

[9] Harvard sentences: Appendix of: IEEE Subcommittee on Subjective Measurements. (1969) IEEE Transactions on Audio and Electroacoustics, 17, 227-246.

[10] Ramsay J.O. and Silverman, B.W. (1997) Functional Data Analysis. Springer

[11] http://www.praat.org

[12] Mixdorff, H. (2000) A novel approach to the fully automatic extraction of Fujisaki model parameters. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP2000, Vol.3, 1281-1284.

[13]www.tfh-berlin.de/~mixdorff/fujisaki_analysis.htm

[14] Kim, J. And Davis, C. (2003). Hearing foreign voices: does knowing what is said affect visual-masked-speech detection? Perception, 32, 111-120.