Article:

Alghamdi, N., Maddock, S., Marxer, R. et al. (2 more authors) (2018) A corpus of audio-visual Lombard speech with frontal and profile views. Journal of the Acoustical Society of America, 143 (6). pp. 523-529. ISSN 0001-4966

10.1121/1.5042758

A corpus of audio-visual Lombard speech with frontal and profile views

Najwa Alghamdi,[1,2,a)] **Steve Maddock,**[1] **Ricard Marxer,**[1,3] **Jon Barker,**[1] **and Guy J. Brown**[1]

[1)] *Department of Computer Science, University of Sheffield, UK*

[2)] *Information Technology Department, King Saud University, Saudi Arabia*

[3)] *Université de Toulon, Aix Marseille Univ, CNRS, LIS, Marseille, France*

amalghamdi1@sheffield.ac.uk, nalghamdi@ksu.edu.sa,
s.maddock@sheffield.ac.uk,
marxer@univ-tln.fr,
j.p.barker@sheffield.ac.uk,
g.j.brown@sheffield.ac.uk

(Dated: 8 June 2018)

Abstract: This paper presents a bi-view (front and side) audiovisual Lombard speech corpus, which is freely available for download. It contains 5,400 utterances (2,700 Lombard and 2,700 plain reference utterances), produced by 54 talkers, with each utterance in the dataset following the same sentence format as the audiovisual Grid corpus (Cooke et al., 2006). Analysis of this dataset confirms previous research, showing prominent acoustic, phonetic, and articulatory speech modifications in Lombard speech. In addition, gender differences are observed in the size of Lombard effect. Specifically, female talkers exhibit a greater increase in estimated vowel duration and a greater reduction in F2 frequency.

[a] Author to whom correspondence should be addressed.

1. Introduction

The Lombard effect (Lombard, 1911) is a reflexive adaptation to speech production which occurs when communicating in adverse conditions. Lombard speech is characterized by a collection of acoustic and phonetic modifications, including an increase in fundamental frequency (F0) and signal energy, a shift in the centre frequency of the first and second formants (F1 and F2), a tilt of the speech spectrum, and an increase in vowel duration (Junqua, 1993; Lu and Cooke, 2008). In the visual domain, greater face and head motion (Vatikiotis-Bateson et al., 2007) and a greater global change in the movement of the jaw and lips (Garnier et al., 2010) have been reported. When presented at the same signal-to-noise ratio, Lombard speech (uttered in the presence of noise) is usually more intelligible than plain speech (uttered in quiet)(Cooke et al., 2014).

Although studies of Lombard speech have been consistent in their general characterisation of the effect, there have been widely varying reports of even the most basic characteristics, e.g., reports of the level increase when speaking in 80 dB of noise vary (Pittman and Wiley, 2001; Summers et al., 1988; Tartter et al., 1993). Some of this variability is due to the manner in which individual speakers respond to noise. However, previous studies have typically used small numbers of speakers, making it hard to get a good characterisation of these across-speaker effects. Pooling results across studies is not typically valid because the Lombard reflex is sensitive to the characteristics of the communication environment, including noise type (Lu and Cooke, 2008), the noise immersion method (Garnier et al., 2010), noise level (Šimko et al., 2016), communication task (Garnier et al., 2010), and communication modality (Fitzpatrick et al., 2015), variables which typically vary from one study to the next.

This paper aims to provide a more detailed characterisation of the across-speaker variation in the Lombard effect by collecting and analysing a corpus of plain and Lombard speech from a total of 54 speakers uttering a total of 5400 utterances. The amount of data collected significantly exceeds that used in previous controlled Lombard studies. It is also the first collection that has been designed with precise video analysis in mind. In particular, the collection uses head-mounted cameras that allow highly accurate measurement of the visual Lombard effect from both a frontal and profile view.

The data are being made publicly available for the benefit of other researchers. In particular, the dataset is an extension of the audio-visual Grid corpus (Cooke et al., 2006) that has been widely used in the study of speech intelligibility in noise and the perception of simultaneous speech signals. The data are also suitable for development of novel speech processing algorithms. In particular, the Lombard effect has major implications for the design of automatic audio/audiovisual speech recognition systems. Such systems are typically trained on clean speech datasets or on datasets to which noise has been artificially added. The performance of these systems can then deteriorate under real Lombard conditions that have not been observed during training. Although there are audio-video speech datasets that have been recorded in noise, e.g., AVICAR (Lee et al., 2004), these datasets lack controlled non-Lombard reference signals against which to make accurate measurements of the adaptation.

The paper first describes the design and collection of the new dataset. It then presents an initial analysis of the acoustic, phonetic, and articulatory speech modifications under Lombard conditions across the dataset talkers. Results of this analysis are compared to previous research conducted on a smaller numbers of talkers (Junqua, 1993; Junqua et al., 1999; Lu and Cooke, 2008; Pisoni et al., 1985; Vatikiotis-Bateson et al., 2007), in which clear modifications in Lombard speech were reported. Finally, the larger number of speakers also enables us to report on the gender differences for both the audio and visual aspects of Lombard speech.

2. Corpus

2.1 Sentence design

The sentences in the corpus conform to the Grid corpus syntax (Cooke et al., 2006). These are six-word sentences, for example ‘bin blue at A 2 please’, with the following structure: <command: bin, lay, place, set> <color: blue, green, red, white> <preposition: at, by, in, with> <letter: A-Z (excluding W)> <digit: 0-9> <adverb: again, now, please, soon>. Three of these words – color, letter, and digit – are considered to be “keywords,” while the remaining words are “fillers.” The original Grid corpus was collected from 34 talkers reading 34,000 sentences selected from 64,000 possible combinations of the Grid word sequences. For the new Lombard Grid corpus, 55 talkers[1] uttered sets of sentences from the pool of the remaining 30,000 Grid word-sequence combinations (i.e., those that were not used in the original Grid corpus). Each talker was assigned to a unique set of 50 sentences featuring a uniform representation of Grid keywords, including twelve to fourteen instances of each color, two instances of each letter, five instances of each digit, and representative coverage of the Grid filler words[2].

Following other studies, e.g. Lu and Cooke (2008), speech-shaped noise (SSN) was used to induce the Lombard effect. In this study, SSN was created by filtering white noise to match the long-term spectrum of a speech corpus that includes 1,000 Grid sentences of a selected talker (ID = 1). Linear predictive coding was used to obtain the spectral envelope of the speech corpus. In previous Lombard-related studies, noise has been presented to talkers at a variety of levels, including 80 dB SPL (Summers et al., 1988), 85 dB SPL (Junqua, 1993), and 89-96 dB SPL (Lu and Cooke, 2008). For the current study, 80 dB SPL was chosen as the noise level: this is loud enough to induce a robust Lombard effect while still being at a level low enough to avoid hearing damage or undue vocal/auditory fatigue.

2.2 Talker population

The talkers who participated in the experiment consisted of 55 native speakers of British English (both male and female), all of whom were staff or students at the University of Sheffield in the 18 – 30 year age range. The hearing of the talkers was screened using a pure-tone audiometric test. All participants were paid for their contributions; ethics permission was obtained by following the University of Sheffield Ethics Procedure.

2.3 Collection

The recordings were made in a single-walled acoustically-isolated booth (Industrial Acoustics Company [IAC]). The speech material was collected at a sampling rate of 48,000 Hz and a resolution of 24 bits using a C414 B-XLS AKG microphone placed 30 cm in front of the talkers and digitized using the MOTU 8-pre 16 $\times$ 12 Audio Interface. The talkers wore Sennheiser HD 380 pro headphones. The SSN was mixed with the audio signal of their speech to provide self-monitoring feedback at a level that compensated for headphone attenuation.

The level of playback of the talkers' speech was carefully adjusted so that their perception of talking with and without the headphones would be comparable. The process was subjectively measured; the talker wore one headphone over one ear while the other ear remained uncovered. The talker was requested to speak while the playback of his/her voice was presented at gradually increasing levels via the headphones. The talker was asked to indicate the level at which balanced auditory feedback was received across his/her left and right ears. This level (which had relatively little variation amongst participants) was then recorded and used to present the self-monitoring feedback in the headphones. The noise presentation level was adjusted to 80 dB SPL using a Cirrus Optimus Yellow Class 2 sound level meter. In this process, a MATLAB routine automatically tuned the level of the Lombard inducing noise until a reading of 80 dB was achieved. This level was then recorded and fed to a MATLAB routine that controlled the presentation of the SSN during the recording experiment.

In addition to the audio recordings, simultaneous audiovisual recordings were made using a custom-made helmet rig system that was worn by the talkers. The system consisted of a lightweight bicycle helmet on which were mounted two Logitech HD Pro USB Webcam C920s connected using 8-inch GoPole Arm Helmet Extension armatures. This allowed one camera to be positioned directly in front of the face and one at a fixed position to the side of the face. Head-mounting ensured that the viewing angles remained fixed regardless of head motion thus allowing for more precise comparison of Lombard and non-Lombard visual speech. Four light sources were positioned so as to produce roughly uniform illumination across each talker's face; a plain white background was placed behind and at the right side of the talker's seat.

The audiovisual recordings from the webcams were collected onto two computers via USB 2.0 interfaces. The audiovisual stream from the front webcam was collected at 480p resolution (720 x 480), in full frame, at a variable frame rate fluctuating around 24 frames per second (mean FPS = 23.93; mean bitrate = 2817.82 kb/s). The recording software encoded the video stream using the built-in H.264 encoder and the audio stream using the AAC encoder at a sampling rate of 44,100 Hz. The video stream from the side webcam was collected at 480p (864 x 480) and in full frame at 30 FPS. The recording software encoded the video stream using the WMV encoder and the audio stream using wmav2 at a sampling rate of 48,000 Hz.

Each talker produced 100 utterances by reading his/her sentence list in both plain and Lombard conditions. The collection of the utterances in each condition was made in 5 blocks of 10 utterances. The plain and Lombard blocks were presented in an alternating order. Each block of 10 utterances was preceded by 5 'warm-up' utterances that were used to allow talkers to attune to the change in condition (i.e., from noise present to noise absent and vice versa). These initial utterances were discarded after recording. The Lombard-inducing noise was controlled by a computer (using a MATLAB routine as previously described) and was present throughout the Lombard blocks and turned off during the non-Lombard blocks.

The talkers read the sentences to the researcher, who acted as a listener. Having a listener was necessary because the Lombard effect is triggered both as an unconscious reaction to noise and by the need to maintain intelligible communication in noise (Lu and Cooke, 2008). The talkers sat inside a booth facing a screen, where the sentences were presented; the listener sat outside the booth listening to the talkers' speech, presented at 60 dB SPL, via a pair of Panasonic RP HT225 headphones connected to the audio interface. The presentation of the prompt sentences, as well as the listener's messages to each talker, were both controlled by a MATLAB script. The talkers were instructed to speak at a normal pace and in a natural style and were given 5 seconds to read each sentence. To aid this process, the talkers were prompted by a progress bar on the screen with a duration of 5 seconds. If the talker misread the prompt, then the listener presented the same sentence again. During the Lombard blocks, the listener asked the talkers to repeat an utterance every 5 to 7 sentences by indicating that she could not hear the talker. The purpose of this step was to maintain the public Lombard loop, which is driven by communication needs (Lu and Cooke, 2008).

2.4 Post-processing

First, the audio and visual signals were temporally aligned. This was achieved automatically by comparing the high quality audio (i.e., as captured by the desk microphone) and the audio embedded in the front and profile video signals. Specifically, for each of the two video channels, a search was made for the temporal offset that maximised the correlation between the hiqh quality audio signals and the audio in the video channel.

Second, each utterance was automatically end-pointed (delimited in time). For each session, an analysis of the speech energy envelope was employed to make an initial estimate of the utterance and end times. The automatic end pointing was then reviewed by a human annotator who corrected any gross end-pointing errors. The Kaldi toolkit (Povey et al., 2011) was then used to automatically determine vowel boundaries and end-points. A typical GMM-HMM setup was employed to force-align the acoustic recordings to phonetic transcriptions of the utterances. Training was performed using maximum likelihood linear transform (MLLT) model adaptation and feature-space maximum likelihood linear regression (fMLLR) speaker-adaptive training[3].

Finally, for each speaker, the 100 non-warm-up utterances were automatically extracted from the continuous audio and video signals using an extraction tool based on the FFMPEG [4] framework. Prior to extraction, a 200 ms margin was added by the extraction tool to the start and end times to capture the immediate context (i.e., so that pre-emptive visual cues are preserved). The audio stream was downsampled to 16 kHz and the start and end times were used to extract each utterance. The corresponding segments were also extracted from the video sequences (using H.264 codec) by adjusting the timings to compensate for the computed audio-visual offsets. In cases where the subject spoke the utterance multiple times (e.g. due to being asked to repeat or because of a reading error) the first correct rendition of the utterance was extracted and the repeats were discarded.

3. Analysis of the Lombard Effect

Acoustic, phonetic, and articulatory parameters were extracted from the plain and Lombard recordings of 54 talkers to study the Lombard effect. Three acoustic parameters from the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) (Eyben et al., 2016) were extracted using the openSMILE toolkit[5]. These acoustic parameters, calculated as means for each audio utterance, included a fundamental frequency-related parameter, namely the F0 mean, an energy-related parameter, namely the loudness mean, and a spectral parameter, namely the alpha ratio mean (Sundberg and Nordenberg, 2006) (the ratio between the energy from 50–1000 Hz and 1–15 kHz). Four additional parameters were estimated to characterise the vowels: the average of vowel duration, the ratio of total vowel duration to utterance duration, and the average first and second formant frequencies (estimated using Praat‘s (Boersma, 2006) formant tracker. Settings: default; max formant for female talkers = 5500 Hz; max formant for male talkers = 5000 Hz). One articulatory parameter, the vertical mouth aperture, was extracted using the Dlib toolkit (King, 2009); the standard deviation of this parameter across frames was calculated for each video utterance as a measure of ‘visual energy’. Each talker’s mean (i.e., the mean of these parameters across utterances produced by that talker) was calculated.

Figure 1 shows the talkers’ means in plain and Lombard conditions for each of the eight parameters. Table 1 shows across-talker means and standard deviations (SDs). Paired-samples t-tests were employed to determine the significance of differences between the across-talker means, across-female-talker means, and across-male-talker means in plain and Lombard conditions. Table 1 also summarizes the results of the statistical analysis.

The Lombard speech adaptations reported in previous studies (see Section 1) were observed in the Lombard recordings of this corpus. All parameters, except for the F2 frequency, demonstrated significant increases. The mean F1 frequency is expected to increase under the Lombard effect (Junqua, 1993; Lu and Cooke, 2008; Pisoni et al., 1985; Summers et al., 1988; Kirchhuebel, 2010). Mixed findings, however, have been reported regarding F2 adaptation to noise: Junqua (1993) reported an increase by female talkers; Pisoni et al. (1985) and Lu and Cooke (2008) reported a decrease by both genders; Kirchhuebel (2010) found variable effects. In this paper, the mean F2 frequency showed a non-significant overall decrease, a similar finding to Pisoni et al. (1985) and Lu and Cooke (2008)[6], but this decrease was significant for female talkers.

Consistent with Junqua et al. (1999)'s findings, individual differences in coping with the SSN noise were found. Gender differences were also noticed in the size of Lombard effect. For example, female talkers showed greater increase in loudness, estimated vowel duration, estimated vowel-to-utterance ratio and mouth aperture, and a greater decrease in vowels F2 frequency. A one way MANOVA found a statistically significant difference in speech parameters' adaptations to noise based on talkers' gender ($F(8,\ 45)\ =\ 2.994,\ p\ =.009$): gender has a statistically significant effect on estimates of both vowel duration adaptation ($F(1, 52) = 4.96$; $p = 0.03$) and F2 frequency adaptation ($F(1, 52) = 6.68$; $p = 0.01$). Gender differences may have resulted from articulation differences between male and female talkers, as female talkers speak with a higher degree of articulation than male talkers (Koopmans van Beinum, 1980), a strategy that might be more exaggerated under the Lombard effect (Junqua, 1993). Junqua (1993) also found that Lombard speech produced in multi-talker noise by female talkers is more intelligible than male talkers. Gender difference has also been reported when the auditory feedback is delayed (Howell and Archer, 1984). This could suggest that male and female talkers may differ in their strategic responses to the auditory feedback that mediates the Lombard effect.

4. Corpus description

The corpus is being made freely available for download under a Creative Commons Attribution 4.0 International license. The download consist of 5400 utterances where for each utterance there is an audio file, front view video file and a profile view video file. The downloads are accompanied by a JSON format file storing associated metadata including the gender of each speaker and the utterance recording sequence. The corpus is available from http://spandh.dcs.shef.ac.uk/lombardgrid/.

5. Summary

This study has presented a bi-view audiovisual Lombard speech dataset collected under high-SNR levels. The dataset, which is an extension of the popular Grid corpus, includes audio, front-video, and side-video recordings of 54 talkers uttering 5,400 plain and Lombard sentences. Analysis of this dataset showed prominent acoustic, phonetic, and articulatory speech modifications in Lombard speech, which confirms previous research on the subject.

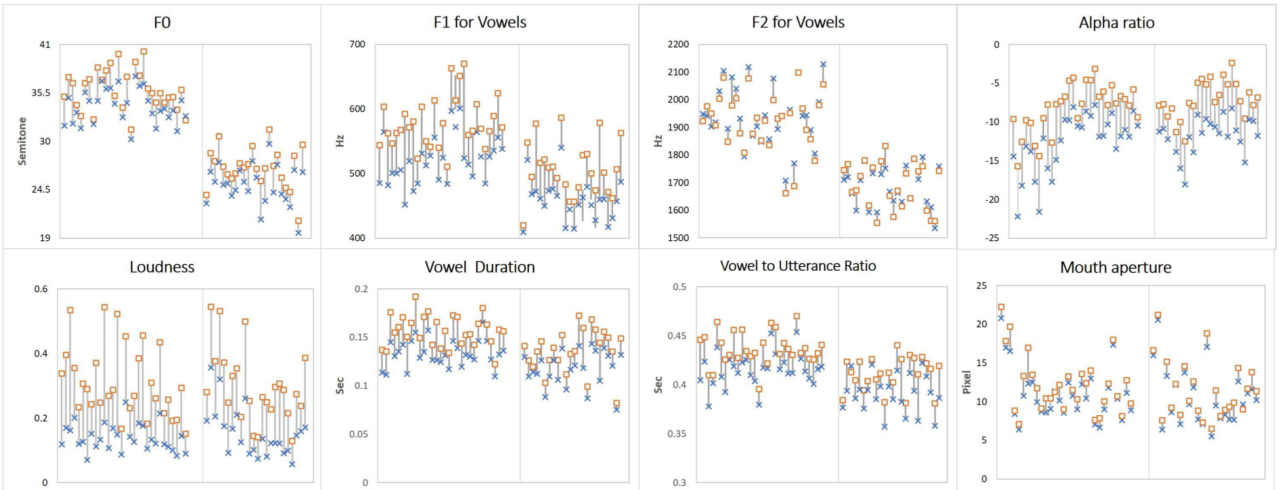

Fig. 1. Estimated acoustic, phonetic and visual features across talkers: Lombard (□); plain (×). In each sub-figure: female talkers (left); male talkers (right).

The large number of speakers has also enabled the testing of gender differences in the size of Lombard effect, with female speakers showing a greater increase in estimated vowel duration, and a greater decrease in F2 frequency. The complete dataset has been made publicly available for future research.

6. Acknowledgements

This research was funded by the UK Engineering and Physical Sciences Research Council (EPSRC project AV-COGHEAR, EP/M026981/1) and by the Saudi Ministry of Education, King Saud University.

References and links

[1]Recordings of the talker with ID 1 were subsequently excluded due to technical issues.

[2] Note, the Grid corpus has not been designed to be phonetically balanced and has limited coverage of the phonetic contexts occurring in English. This may be a limitation for some usages.

[3] A subset of the alignments generated from this process (10 pairs of utterances from the Lombard and non-Lombard conditions, 20 in total) were validated with human annotators. Findings showed that the ASR system consistently underestimated vowel duration by 0.029 ± 0.012 s compared to the human annotation. Importantly, however, the difference between human-estimated and ASR-estimated vowel durations was not affected by the experimental condition (i.e., the ASR showed no bias between the Lombard and non-Lombard speech conditions).

[4]https://www.ffmpeg.org/

[5]http://audeering.com/technology/opensmile/

[6] Although the shifts in *estimated* formant frequencies are in agreement with those observed in the literature, it should be acknowledged that the effect may be partly due to changes in alpha-ratio rather than changes to the actual formants.

P. Boersma. Praat: doing phonetics by computer. *http://www.praat.org/*, 2006.

M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

M. Cooke, S. King, M. Garnier, and V. Aubanel. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language*, 28(2):543–571, 2014.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, April 2016. ISSN 1949-3045.

M. Fitzpatrick, J. Kim, and C. Davis. The effect of seeing the interlocutor on auditory and visual speech production in noise. *Speech Communication*, 74:37–51, 2015.

M. Garnier, N. Henrich, and D. Dubois. Influence of sound immersion and communicative interaction on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 53 (3):588–608, 2010.

J-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1993.

J-C. Junqua, S. Fincke, and K. Field. The Lombard effect: A reflex to better communicate with others in noise. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 4, pages 2083–2086. IEEE, 1999.

D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. S. Huang. AVICAR: audio-visual speech corpus in a car environment. *INTERSPEECH*, pages 2489–2492, 2004.

E. Lombard. Le signe de l'élévation de la voix. *Ann Maladies de L'Oreille et du Larynx*, 37: 101–119, 1911.

F. J. Koopmans-van Beinum. Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions. *PhD thesis, Universiteit van Amsterdam*, , 1980.

Y. Lu and M. Cooke. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124(5): 3261–3275, 2008.

D. Pisoni, R. Bernacki, H. Nusbaum, and M. Yuchtman. Some acoustic-phonetic correlates of speech produced in noise. In *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 1581–1584, 1985.

A. L. Pittman and T. L. Wiley. Recognition of speech produced in noise. *Journal of Speech, Language, and Hearing Research*, 44(3):487–496, 2001.

J. Šimko, S. Beňuš, and M. Vainio. Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue. *The Journal of the Acoustical Society of America*, 139 (1):151–162, 2016.

W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

J. Sundberg and M. Nordenberg. Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The Journal of the Acoustical Society of America*, 120(1):453–457, 2006.

P. Howell and A. Archer. Susceptibility to the effects of delayed auditory feedback. *Perception & Psychophysics*, 36(3):296–302, 1985.

V. C. Tartter, H. Gomes, and E. Litwin. Some acoustic effects of listening to noise on speech production. *The Journal of the Acoustical Society of America*, 94(4):2437–2440, 1993.

E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia. Audiovisual Lombard speech:. *reconciling production and perception.* Auditory-Visual Speech Processing (AVSP), 2007.

Christin Kirchhuebel. The effects of lombard speech on vowel formant measurements. *São Paulo School of Advanced Studies in Speech Dynamics SPSASSD 2010 Accepted Papers*, page 38, 2010.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

Steve J Young, Julian J Odell, and Philip C Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312. Association for Computational Linguistics, 1994.

Table 1. The mean and standard deviation (M±SD) of acoustic, phonetic and visual features of all talkers, female (F) talkers and male (M) talkers. P: plain, L: Lombard. Columns t summarize the results of statistical analyses (t-tests) between plain and Lombard conditions. Symbols: increase: ↑ , decrease: ↓; All tests were significant ($p < 0.001$) except those marked with * ($p > 0.5$)

	F0 (semitones $0 \rightarrow 27.5Hz$)			Vowels F1 (Hz)			Vowels F2 (Hz)		
	P	L	t	P	L	t	P	L	t
All	30.0 ± 4.9	31.9 ± 4.9	↑	493 ± 46	547 ± 54	↑	1828 ± 158	1819 ± 149	↓*
F	34.0 ± 1.9	35.9 ± 2.3	↑	521 ± 36	579 ± 39	↑	1943 ± 105	1922 ± 102	↓
M	25.0 ± 2.2	27.0 ± 2.2	↑	458 ± 31	507 ± 42	↑	1683 ± 70	1689 ± 82	↑*

	Vowel duration (ms)			Vowel-to-utterance ratio			Alpha ratio		
	P	L	t	P	L	t	P	L	t
All	126 ± 17	148 ± 21	↑	0.4045 ± 0.021	0.4254 ± 0.021	↑	-12.17 ± 3.25	-7.67 ± 2.83	↑
F	133 ± 14	157 ± 16	↑	0.4153 ± 0.017	0.4367 ± 0.017	↑	-12.63 ± 3.74	-8.17 ± 3.05	↑
M	118 ± 18	136 ± 22	↑	0.3910 ± 0.019	0.4113 ± 0.017	↑	-11.59 ± 2.36	-7.037 ± 2.38	↑

	Loudness			Mouth aperture (pixel)		
	P	L	t	P	L	t
All	0.145 ± 0.058	0.306 ± 0.110	↑	10.777 ± 3.43	11.914 ± 3.66	↑
F	0.139 ± 0.041	0.313 ± 0.109	↑	10.967 ± 3.29	12.204 ± 3.61	↑
M	0.153 ± 0.074	0.298 ± 0.110	↑	10.540 ± 3.59	11.552 ± 3.69	↑