Whispered and Lombard speech: different ways to exaggerate articulation

Chris Davis and Jeesun Kim
The MARCS Institute, Western Sydney University
chris.davis/j.kim@westernsydney.edu.au

Abstract

When speaking in noise (Lombard speech) talkers exaggerate their articulation compared to speaking in quiet. The current study compared motion exaggeration in this speech style with whispered speech by measuring talker's lip and jaw, and eyebrow and head motion. Four talkers uttered sentences in quiet, noise or in whisper while their face and head movements were recorded with optical tracking. The results showed that both Lombard and whispered speech had movements of greater duration and amplitude than speech in quiet. For half the participants, whispered speech had greater motion than Lombard, whereas the other half showed the opposite pattern.

Index Terms: speech production, Lombard speech, whispered speech, exaggeration

1. Introduction

Speech style can have a considerable impact on the way that speech is articulated (e.g., hyper/hypo-articulation). Changes and variability in speech articulation are of interest to researchers from a range of disciplines (e.g., those researching speech production and recognition, speaker states, speech biometrics; multimodal speech, etc).

Exaggerated articulation is particularly interesting as it has been proposed to make speech more intelligible [e.g., 1]. Speech exaggeration has been observed in a number of different speech styles (e.g., clear speech, machine speech, foreigner directed speech, Lombard speech). In this study, we contrast a well-studied speech style that is known to produce hyper-articulation (speaking in noise, Lombard speech) with a less studied style that has also been suggested to exhibit hyper-articulation (whispering).

Comparing whispered and Lombard speech is interesting because part of the exaggerated articulation of Lombard speech may be simply due to speaking loudly, whereas for whispered speech it may be due to a strategy to produce clear visual speech. If this were the case, then whispered and Lombard speech may show slightly different motion patterns. For instance, exaggerated jaw and lip motion may be a necessary concomitant of Lombard speech but this may not be the case for eyebrow and head motion that are less directly coupled to speech articulation (although see [9]).

Before describing the current study in more detail, it is useful to note some aspects of whispered speech that were considered in planning the current study. Whispered speech is typically produced with an open glottis and without voicing. Whispered speech has been classified into low and high energy whisper [2]. Low-energy whisper is typically produced in situations where interlocutors wish to maintain local intelligibility but deliberately attempt to reduce the perceptually salience of speech to others (for example, whispering in a library or in a formal or solemn setting). High-energy whisper (also known as 'stage whisper') is produced in order to be intelligible at a distance and is characterized by greater pulmonic force and air flow. In this type of whisper, the voice is still not produced with active vocal cord vibration but occasional passive vibration can occur.

Studies of whispered speech have mainly concentrated upon characterizing the vocal mechanisms of whisper production (e.g., glottal configuration, [3; 4]) or examining the perception of whisper [5]). The few studies that have examined articulation have typically looked at the production of segments (e.g., /p/, /b/ or single vowels). For example, using an intra-oral pressure measure, Schwartz [6] found significantly longer bilabial closure and significantly greater whispered durations for /p/ and /b/ but not /m/. Using a spectrographic analysis method Parnell et al [7] found that closure and constriction durations for /t/, /s/, and /z/ were significantly longer in whispered vowel environments compared to voiced ones. Higashikawa et al [8] used a video-based analysis of reflective markers positioned on both the upper and lower lip and to the protuberance of the mandible. These authors found differences between lip movements for bilabial plosives during normal and whispered speech. That is, lip opening was significantly faster when whispering /b/ compared with whispering /p/ or non-whispered /b/. In summary, it has been demonstrated that at least for some segments, whispered production speech may have greater duration and involve greater articulatory motion.

The current study examined high energy whisper (i.e., the interlocutor was separated from the speaker by several meters due to restriction imposed by the recording method). In this regard we note that Solomon et al [2] showed that low-energy and high-energy whispering could be differentiated by supraglottal constriction and to lesser extent by vocal-fold adjustments, (although individual differences tended to be considerably larger than any systematic effects due to the type of whisper). It also seems plausible, given their different functional roles, that the degree of visual articulation will differ between the two types of whisper.

To examine the articulator effects of whisper and Lombard speech, we used active infrared optical tracking (Optotrak) and measured articulatory motion from the lips and jaw (movements directly related to speech articulation as in Higashikawa et al [8]), as well as eyebrow and rigid head motion. Also, we examined motion at the sentence (unlike [8]) rather than segment level. This was done to obtain a more global impression of the effects of exaggeration on articulation. We also examined speech production in quiet in order to determine a baseline against which hyper-articulation could be judged.

2. Method

2.1. 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.

2.2. Materials

10 sentences selected from the 1965 revised list of phonetically balanced sentences (Harvard Sentences, [10]).

2.3. Apparatus

Two Northern Digital Optotrak machines were used to record the movement data. The configuration of the markers on the face and head rig is shown in Figure 1.

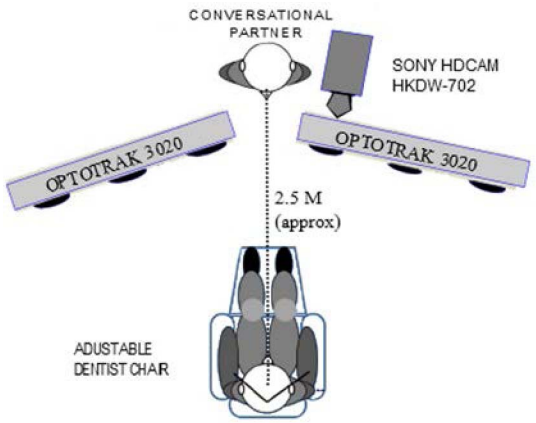

Figure 1: *The layout of the capture session. The participant whose head and face movement were recorded sat in an adjustable chair facing a "conversational partner" who stood behind the two NDI Optotrak 3020 machines.*

2.4. Procedure

Each session began with the placement of the movement sensors (see Figure 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 quiet 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 2.5 meters (See Figure 1). The participant then repeated the ten sentences.

This basic procedure was repeated several times, once for each speech mode (in quiet, whispered, Lombard). Lombard speech was induced by participants speaking while hearing multi-talker babble through a set of ear phones (at approximately 80 dB SPL, a similar level to [1]). For whispering, participants were instructed to whisper the sentences at a level judged loud enough for the conversational partner to hear.

2.5. 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. To guard against the over-representation of particular marker configurations, a movement threshold quantification procedure was employed to keep only the frames that were sufficiently different. *Guided* Principle Component Analysis (gPCA, see [11]) was used to reduce the dimensionality of the data; this procedure uses linear decomposition to constrain the PCA to motion planes that are relevant to articulation using *a priori* defined markers (jaw, lips, eyebrows).

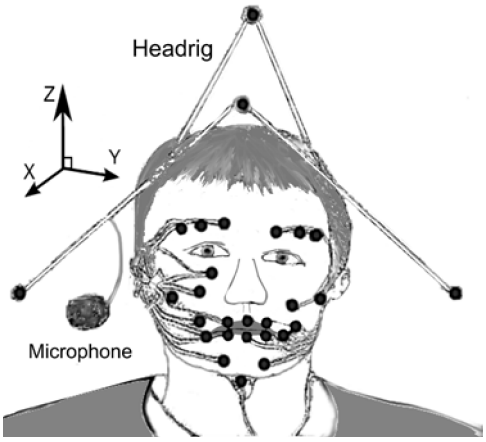

Figure 2: *The location of the 24 optical sensors on the face (the size of the sensors have been exaggerated for clarity). Four additional sensors were positioned on a head-rig to measure rigid movements around the centre of rotation.*

For the current study we report the following gPCA components. Direct speech articulation: Jaw: Jaw Opening; Jaw Protrusion; Mouth: Lip Opening; Lip Rounding. Indirect articulation: Eyebrow: Eyebrow Raising; Eyebrow Pinching. Head translation: Forward / Backwards; Up / Down.

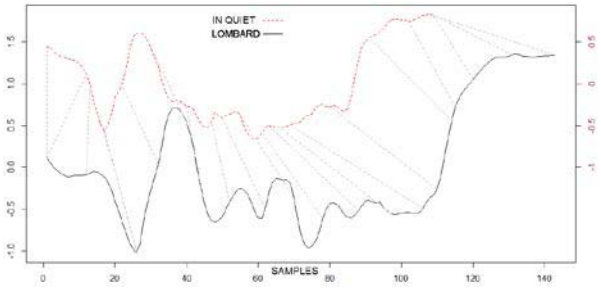

Figure 3: *An example of two time series showing the contribution of a guided principle Component (here Lip opening) for speech 'in quiet' and 'in noise'. Also shown is the DTW between them (the vertical axis represents cm, left scale in noise condition).*

In order to quantify the degree of hyper-articulation of the production component we used Dynamic Time Warping (DTW) [12]. Dynamic time warping (DTW) is a procedure

that provides a measure of comparison between series of data points (inherent distance or warping cost). For example, DTW can expand or compress one time series to resemble another one and by summing the distances of individually aligned elements an inherent distance between the two can be computed (Figure 3).

We compared the warping cost of time-series of the contribution of the gPCs over each utterance for the speech in quiet condition compared to either the Lombard or the whispered speech conditions. Note that these time-series were mean-centered to avoid the effect of off-sets.

In order to have an index of the power of the contribution of a PC, we used the standard deviation of the mean-centered time-series.

3. Results

Figure 4 shows the mean inherent distance scores (warping costs) for whispered or Lombard speech compared with speech produced in quiet for the following motion types: jaw (opening and protrusion) and lips (opening and rounding), as well as eyebrow (up-down and pinched) and rigid head motion (forward–back and up-down).

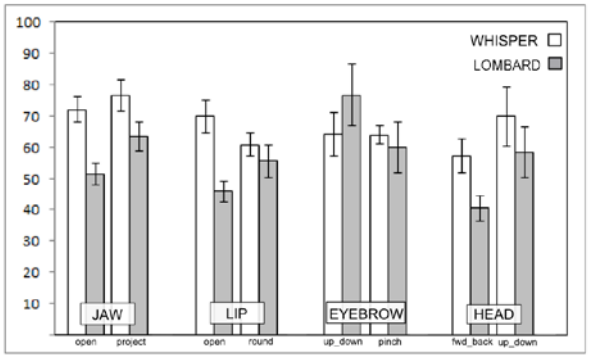

Figure 4: *Mean inherent distance scores (warping cost in arbitrary units) for the different motion types as a function of speech type. Error bars show Standard error.*

It is clear from the magnitudes (and error bars) of the warping cost scores that both the whisper and Lombard speech motion curves differed from those for speech in quiet (the baseline against which the distance scores were determined).

The warping cost scores were analyzed using a linear mixed model (LMM) analysis (random-intercept) using the LmerTest package to approximate degrees of freedom [13]. This analysis indicated that the effect of speech type (whisper vs. Lombard) was significant, $F(1,626) = 9.877$, $p < 0.01$. This effect was not significant if random slopes were included in the model, indicating a possible interaction effect with participants (although [14] have also argued that maximal models can be unduly conservative).

Figure 5 shows the average warping cost scores for all motion types as a function of participants. As can be seen in the figure, there was considerable variation across participants in terms of whether there was more motion in articulating whispered speech or Lombard speech. Indeed, the interaction between the speech type and participant was significant, $t = 4.564$, $p < 0.05$.

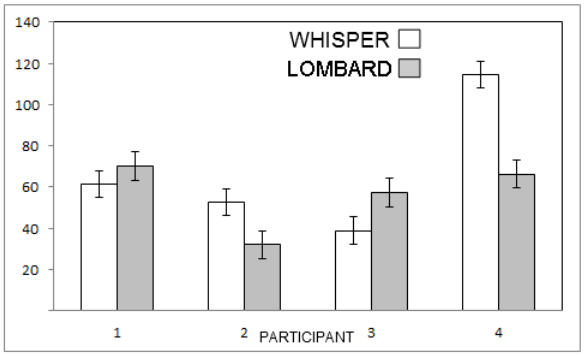

Figure 5: *Mean inherent distance for all motion types as a function of speech type and participant. Error bars show Standard error.*

Figure 6 shows a comparison of distance scores (relative to quiet speech) for whisper and Lombard speech broken down by eyebrow and head motion and jaw and lip motion. Across each individual participant, the pattern of motion for the jaw and lips (motion closely related to speech articulation) was similar to the pattern for the eyebrows and head. This was the case even though there was variability across participants as to whether Lombard speech had more motion than whispered speech or vice-versa.

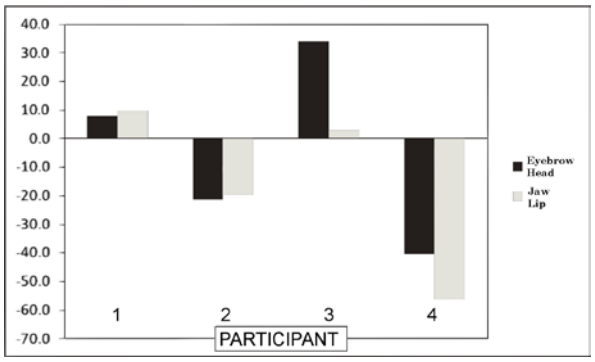

Figure 6: *Mean inherent distance scores for Lombard minus whispered speech (negative means larger warping costs for whispered speech) for Eyebrow and Head motion (black) and Jaw and lip motion (grey) for the participants (1 – 4)*

To determine whether there was a difference in the amount of the contribution of the gPCs, the standard deviation of the mean-centred time-series was calculated as a measure of the PC contribution. This value represents the amount to which the PC score deviated around the mean (which represents the initial configuration). The data for each of the motion PCs are shown in Figure 7.

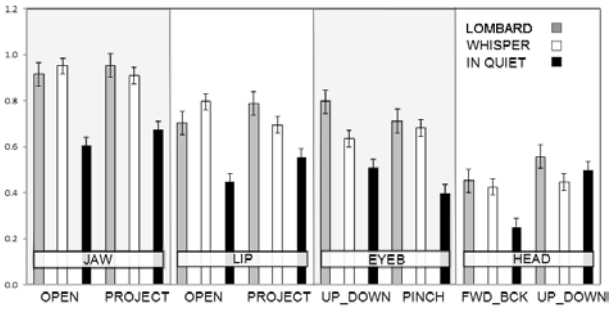

Figure 7: *Mean of the standard deviations (in arbitrary units) of the different motion PCs as a function of speech type. Error bars show Standard error.*

As can be seen in the figure, there was a more power (greater deviation) for the Lombard and whispered speech motion PCs compared to speech uttered in quiet. Two LMMs (random slopes, intercepts) were conducted to determine whether there was greater variation (amplitude) for the whispered and Lombard speech conditions each compared to the speech in quiet one. The LMM for Lombard versus in quiet was significant, $F(1,4) = 10.343$, $p < 0.05$; as was the LMM for whispered versus in quiet, $F(1,3) = 63.8$, $p < 0.01$. There was no difference between the Lombard and whispered speech scores, $F(1,3) = 0.3125$, $p > 0.05$.

4. Discussion

Speech related mouth and jaw articulation, along with eyebrow and head motion was measured for three speech styles (speech in quiet, in noise and whisper). Two methods for quantifying differences in motion of guided principle components were used. The inherent distance (warping cost) between motion PC curves (as given by DTW) for speech in quiet and two the other speech styles indexed temporal exaggeration. The other measure, the deviation of PCs from a centred mean value, indexed the power (amplitude) of the motion contribution. The results of both analyses support the claim that high-energy whispered speech is hyper-articulated compared to speech produced in quiet.

The results concerning whether the degree of whispered speech motion differed from that produced when speaking in noise (Lombard speech) were less clear. The results taken over all speakers suggest that whispered speech may exhibit more hyper-articulation than Lombard speech. This finding is consistent with the proposal that speakers can strategically employ visual speech to aid communication when there is a barrier to auditory communication [15]. However, there was considerable individual variation in this pattern, with greater motion for whispered speech only found for two of the four participants (and two showing the opposite pattern). There was also no clear difference in the patterning of motion of the lips and jaw (motion tied to speech articulation) compared with that of the eyebrow and head. Further investigation is warranted.

Finally, it is interesting to point out practical implications of the current findings. For example, the finding that there is hyper-articulation for (high-energy) whispered speech (that, for some people, is even greater than that shown for Lombard speech) is relevant to a range of speech research topics: from the examination of methods of reliably detecting speech from face motion [16] to auditory-visual speech biometric systems, where whisper may be a useful speech style to use as hyperarticulation may be more distinctive.

5. Acknowledgements

We acknowledge the support of the Australian Research Council grant DP130104447.

6. References

[1] 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.

[2] Solomon, N. P., McCall, G. N., Trosset, M. W., & Gray, W. C. (1989). Laryngeal configuration and constriction during two types of whispering. Journal of Speech, Language, and Hearing Research, 32(1), 161-174..

[3] Esling, J. H. & Harris, J. G. 2003. An expanded taxonomy of states of the glottis. Proceedings of the 15th International Congress of Phonetic Sciences, 1049-1052. Barcelona, Spain:UAB.

[4] Mills, T. I. P. (2009). Speech motor control variables in the production of voicing contrasts and emphatic accent. Doctoral dissertation, University of Edinburgh.

[5] Tartter, V. C. (1994). Hearing smiles and frowns in normal and whisper registers. Journal of the Acoustical Society of America, 96(4), 2101-2107.

[6] Solomon, N. P., McCall, G. N., Trosset, M. W., & Gray, W. C. (1989). Laryngeal configuration and constriction during two types of whispering. Journal of Speech, Language, and Hearing Research, 32, 161-174.

[7] Parnell, M., Amerman, J. D., & Wells, G. B. (1977). Closure and constriction duration for alveolar consonants during voiced and whispered speech. Journal of the Acoustical Society of America, 86, 1678-1683.

[8] Higashikawa, M., Green, J. R., Moore, C. A., & Minifie, F. D. (2003). Lip kinematics for/p/and/b/production during whispered and voiced speech. Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP), 55(1), 17.

[9] Davis, C., Kim, J., Grauwinkel, K., & Mixdorff, H. (2006, May). Lombard speech: Auditory (A), Visual (V) and AV effects. In Proceedings of the Third International Conference on Speech Prosody (pp. 248-252).

[10] Harvard sentences: Appendix of: IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. IEEE Transactions on Audio and Electroacoustics. 17, 227-46, 1969.

[11] Maeda, S. (2005). Face models based on a guided PCA of motion capture data: Speaker dependant variability in /s/ - /z/ contrast production. ZAS Papers in Linguistics, 40, 95-108.

[12] Giorgino T (2009).Computing and Visualizing Dynamic TimeWarping Alignments in R: The dtw Package." Journal of Statistical Software, 31(7), 1{24. URL http://www.jstatsoft. org/v31/i07/

[13] Kuznetsova, P. B. Brockhoff, R. H. B. Christensen (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R-Version:1.1-0. http://cran.rproject.org/web/packages/lmerTest/index.html.

[14] Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2015). Balancing Type I Error and Power in Linear Mixed Models. arXiv preprint arXiv:1511.01864.

[15] Fitzpatrick, M., Kim, J., & Davis, C. (2011). The effect of seeing the interlocutor on speech production in different noise types. In Twelfth Annual Conference of the International Speech Communication Association.

[16] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. Speech Communication, 52(4), 270-287.