Freyman, R.L., Helfer, K.S., McCall, D.D., Clifton, R.K., 1999. The role of perceived spatial separation in the unmasking of speech. J. Acoust. Soc. Am. 106, 3578 – 3588.

Helfer, K.S., Freyman, R.L., 2008. Aging and speech-on-speech masking. Ear Hear. 29, 87–98.

Li, L., Daneman, M., Qi, J.G., Schneider, B.A., 2004. Does the information content of an irrelevant source differentially affect speech recognition in younger and older adults? J. Exp. Psych.: Hum. Per. Perf. 30, 1077 – 1091.

Litovsky, R.Y., Colburn, H.S., Yost, W.A., Guzman, S.J., 1999. The precedence effect. J. Acoust. Soc. Am. 106, 1633 – 1654.

Nabelek, A.K., Robinson, P.K., 1982. Monaural and binaural perception in reverberation for listeners of various ages. J. Acoust. Soc. Am. 71, 1242 – 1248.

Rakerd, B., Aaronson, N.L., Hartmann, W.M., 2006. Release from speech-on-speech masking by adding a delayed masker at a different location. J. Acoust. Soc. Am. 119, 1597 – 1605.

Schneider, B.A., Li, L., Daneman, M., 2007. How noise interferes with speech comprehension in everyday listening situations? J. Am. Acad. Audiol. 18, 578-591.

Wu, X.-H., Wang, C., Chen, J., Qu, H.-W., Li, W.-R., Wu, Y.-H., Schneider, B. A, & Li, L. (2005). The effect of perceived spatial separation on informational masking of Chinese speech. Hear. Res., 199, 1 - 10.

INTRINSIC RELATIONSHIP BETWEEN FACIAL MOTION AND ACOUSTICS IN INDIVIDUALS WITH PARKINSON'S DISEASE

Luyao Ma[1,2], Pascal van Lieshout[1,2,3,5], Huawei Colin Li[1,3,4], and Akiko Amano-Kusumoto[1,6]
[1]Oral Dynamics Lab, Department of Speech-Language Pathology, University of Toronto; [2]Department of Psychology, University of Toronto; [3]Institute of Biomaterials and Biomedical Engineering, University of Toronto; [4]Department of Electrical and Computer Engineering, University of Toronto; [5]Toronto Rehabilitation Institute; [6]OGI School of Science & Engineering, Oregon Health & Science University, Portland, USA
luyao.ma@utoronto.ca; p.vanlieshout@utoronto.ca; huawei.li@utoronto.ca; akusumoto@bme.ogi.edu

Abstract

Parkinson's disease (PD) is closely associated with the death of dopaminergic neurons in the basal ganglia, which results in a reduction of facial dynamics during speech production. In young adult speakers, the relationship between facial motion and acoustics is robust. It can be hypothesized that this relationship between facial motion and speech acoustics is reduced in PD given the limitations in facial expression; however, virtually no study had addressed this relationship yet. The current project was designed to address this issue using a 3D video system in combination with Blacklight illumination to record facial motion with time aligned acoustic data. Findings show that in comparison to age-matched control speakers, PD subjects have significantly lower correlations for speech gestures, except for upper lip movement. The findings of this study have implications for the future development of facial motion based speech recognition software and rehabilitation tools.

In speech production abstract gestural goals are mapped onto time varying inputs of muscular activation levels that control the movement and positioning of biomechanical structures (Goldstein, Byrd, & Saltzman, 2006). The resulting shape of the vocal tract acts as an acoustic filter to produce resonance in specific frequencies, which is largely responsible for the various sounds that we hear (Stevens, 1989). Gestural patterns shape vocal tract dynamics in both visual and auditory ways, so its perception is not bound to a single (auditory) modality. This is captured in the notion of "audio-visual (AV) speech perception". From the perceivers' perspective, the coherence of observed visual information and acoustic signals is very important for comprehension. For example, a recent study indicated that subjects who had lip-read a speaker for one hour subsequently recovered speech in noise better when the acoustic signal was from the same talker as opposed to being from a different talker (Rosenblum, Miller, & Sanchez, 2007). Given the multi-modal nature of speech perception, facial motility is an important factor in how clearly a person produces speech and how others perceive the intended message. The study described here focuses on one particular population where facial motility is an issue, namely individuals with Parkinson's disease.

Parkinson's disease (PD) is a common neurological condition, characterized by muscle rigidity, tremor and slowness in physical movement and is particularly prevalent in people above 50 years of age (Pinto et al., 2004). One of the most severe consequences of PD is the loss of expressiveness in the body and especially in the face. As a result, the facial motility in patients affected by PD gradually diminishes and results in a mask-like status. This affects the natural facial dynamics, which accompany the production of speech (Smith, Ellgring, & Smith, 1996). In addition, PD motor impairment also often results in adverse acoustic changes which affect prosodic contrast in speech and this is evident even in earlier stages of disease progression (Cheang & Pell, 2007). The relatively weak voice in PD speakers in combination with reduced oral and facial movements makes their speech less intelligible and can have serious social consequences (Cheang & Pell, 2007; Miller et al., 2007; Tickle-Degnen & Lyons, 2004).

One way to study the congruency between speech acoustics and visual information is by building a model that maps the relationship between acoustic data and facial motion. If both signals are highly congruent, this model would provide an accurate prediction of facial motion from acoustic input. There are different approaches for building such a model (for a review see (Craig, van Lieshout, & Wong, Under Review). Most recently, our lab used (for various reasons) a linear multi-regression model with data from healthy young adults and found strong correlations (on average 0.70) between predicted and actual movements, suggesting a good correspondence between acoustic and visual speech related information in this group (Craig et al., Under Review). To date, it remains an open question to what extent this relationship is different in people with PD. Obviously, as PD patients are typically older subjects, one cannot simply compare their data to young adults. The proper control group has to be age-matched in order to deal with normal effects of aging.

The current study uses 3D motion data with time-aligned acoustics in a group of individuals with PD and age-matched control speakers. In line with previous work, we used a linear regression approach that has been shown to provide a good predictor model (Craig et al., Under Review; Yehia, Rubin, & Vatikiotis-Bateson, 1998). It can be hypothesized that the congruency between acoustic signals and facial motion in PD may be lower in comparison with age-matched control speakers. Apart from providing important theoretical knowledge about audio-visual relationships in speech production of these populations, the results from this study would be very important for the development of rehabilitation tools that use animation technology based on our understanding of the relationship between facial motion and acoustics in normal and disordered populations.

Methods

Participants

Nine subjects with idiopathic Parkinson's disease (mean age 62.9 years, SD = 9.1 years) were recruited from the Morton & Gloria Shulman Movement Disorders Centre at the Toronto Western Hospital. All PD subjects as part of a previous study received an oro-motor exam, were tested on the Unified Parkinson's disease rating scale (UPDRS) and the modified Hoehn & Yahr staging (0 = no disease to 5 = wheel chair bound or bedridden). Those patients who were medicated were measured on-medication only. They were compared to an age-matched control group (OC) consisting of eleven subjects (mean age 69.2 years, SD = 6.2 years). All individuals had no reported history of speech, language, hearing deficits or facial musculoskeletal abnormalities (congenital or acquired) that may have affected the motion of their face during speech, with the exception of PD in the case of the experimental group. Both groups consisted of native Canadian English speakers only.

Stimuli

The speech stimuli chosen in this experiment were the same as in a previous study done on young adults and consisted of 90 sentences selected from the TIMIT and HARVARD sentence database (Craig et al., Under Review). This combination of speech stimuli was considered to provide a representative sample of linguistic material used in daily speech.

*Procedure*s

Specific details on system calibration and data processing can be found elsewhere (Craig, van Lieshout, & Wong, 2007). Tiny glow-in-blacklight dots of face paint, approximately 2 millimetres in diameter, were applied to various locations on the face of participants. Chosen locations include midsagittal positions on forehead, the dorsum of the nose, upper lip, lower lip, jaw, and left and right position on the cheeks and lip corners (nine positions in total). Gestures are defined as the distance between two markers, which represent facial motion relevant to the production of sounds, like for example jaw opening and lip closure. Figure 1 shows a stylized representation of the speech articulators and related gestures as discussed in this paper. The Forehead (F) position is used as a reference point for head movement correction. With respect to the selected gestures, UL represents upper lip (U) motion relative to the forehead marker; BC represents upper lip (U) versus lower lip (L) motion (or bilabial closure); Jaw represents jaw (aka chin or Ch) motion relative to forehead marker; Cheeks represents motions of the left (C-l) relative to the right (C-r) cheeks and finally, BP represents bilabial protrusion as reflected in movements of the left lip corner (LC-l) and right lip corner (LC-r). Apart from the Cheeks gesture, these are all gestures described in a previous study on young adults (M. S. Craig et al., Under Review).

During the experiment, participants were seated in a darkened, UV illuminated room. The illumination was provided by two 250 W UV blacklights fixed on the ceiling, pointed towards the subjects from 2.5 meters away. In order to generate 3D representations of facial articulatory movements, two digital camcorders were used and located just to the left-of-
center and right-of-center of the face, approximately 1 meter away from the subject and at an angle of 45 degrees.

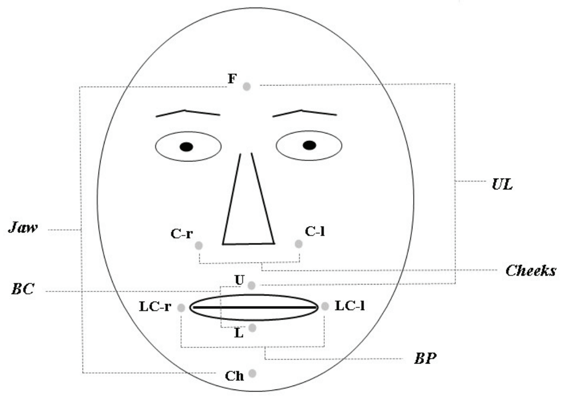

Fig. 1. Stylized depiction of marker positions for individual articulators and gestures. F = forehead; C-r = Cheek right; C-l = Cheek left; U = upper lip; L = lower lip; Ch = chin; LC-r = lip corner right; LC-l = lip corner left; UL = upper lip movement re to forehead position (see text for more details).

The ninety sentences were randomly presented to the participants on a laptop screen in three blocks of thirty sentences each. None of the sentences were repeated. Participants were instructed to read sentences in a normal speaking manner. Facial articulatory movements were recorded with the two cameras while speech acoustics were recorded with a separate digital state recorder (Marantz PMD670/U1B) at 22 kHz sampling rate.

The video/audio files were transferred to a computer. Acoustic data were down sampled to 7972 Hz to match the video sampling rate (59.95 Hz) and the two signals were synchronized using a custom designed MATLAB script. Video files from the 2 cameras were combined using software from the Ariel Performance Analysis System (APAS) to give a 3D representation of facial articulatory movements, which were also time aligned with the audio recordings. Each point in time of the articulatory gestures after synchronization was saved in x, y, z coordinates. Visual and acoustic data were represented in matrix form. The acoustic information was represented by an array of 65 parameters: The root mean squared (RMS) energy, 16th order Linear Predictor Coefficients (LPC), 16th order Line Spectral Pairs (LSP), and the first derivatives of the previous two sets. LSPs are derived from LPCs and are closely related to the formant values and bandwidths in the speech signal. First derivatives were used to capture formant transition information and RMS energy was included because acoustic energy during speech is highly correlated with facial motion (Yehia et al., 1998). RMS was also used to detect the sentences' boundaries (silent parts) when energy was below a certain threshold. In order to predict motion data from acoustic data in each time frame, an MLR (multiple linear regression) analysis was performed, which defined the appropriate transformation matrix (Craig et al., Under Review).

Dependent variables & Statistics

Acoustic data corresponding to one sentence was left for testing (test set), while remaining acoustic data and facial movement data (training set) were used to train the multi-linear regression model. Using a trained MLR model, a facial movement was predicted on a test set. Correlation coefficients (CC) were calculated between the predicted and acquired movement data similar to what was done with young adults (M. S. Craig et al., Under Review). This provides an objective index of the congruency between acoustic signals and facial motion. To handle missing data (5.5% OC; 22.2% PD, mostly for Cheeks and BP gestures) we tested differences in CC between the fixed effects of GROUP (PD versus OC), GENDER (males versus females) and GESTURE (Jaw, BC, BP, Cheeks, UL) using a maximum likelihood estimation method as provided in the Mixed Model Analysis of Variance procedure in NCSS (Hintze, 2007).

Results

The analysis showed no effect for gender (F < 1), so we will only report on GROUP and GESTURE factors. Figure 2 shows the data for CC values for PD and OC, separately for each gesture. It is clear from this figure that except for UL, all gestures show a lower CC value for the PD group. For this data set we also found a significant effect for GESTURE (F(4,52.7) = 2.64, p = 0.04), which was based on a general lower value for UL compared to other gestures, in particular for OC individuals. Given this apparent distinction between the UL gesture and the others, we performed an analysis without the UL gesture data and found a significant GROUP effect for the remaining 4 gestures (F(1,16.3) = 9.13, p = 0.008). PD subjects showed an average CC value of 0.48 (SD = 0.08) and OC showed an average CC value of 0.56 (SD = 0.07). No other effects were found significant.

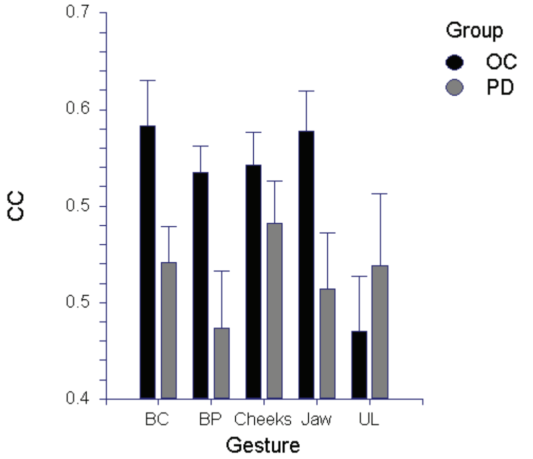

Fig. 2. Correlation coefficients (CC) between predicted and acquired facial movements for OC and PD, separately for the 5 gestures (error bars denote Standard Errors).

Discussion

The findings of this study show that individuals with Parkinson's disease in general show lower congruency between facial motion and acoustics when compared to age-matched control speakers. The only exception to this was found for the UL gesture, which also showed the lowest correlation values in control speakers (see Figure 2).

This reduction of congruency between acoustic parameters and facial motion observed in PD subjects relative to older control subjects may relate to several factors. First, acoustic output is typically reduced in PD subjects, even in the early stages and as a result could be lacking acoustic contract or detail (Cheang & Pell, 2007). This could have been a limiting factor in the development of a transfer matrix relating acoustic parameters to facial motion.

Second, PD patients suffer from a variety of motor problems at the same time (see introduction), including physical tremor and bradykinesia. Even though our PD subjects in general were in early stages of the disease, it is possible that facial motion is already more limited compared to age-matched controls. As with reduced acoustic quality, this would impede on the quality of input to the regression model and potentially reduce the predictability of facial motion. However, it is interesting to notice that upper lip movements show a reverse pattern, with higher CC values for PD subjects. This suggests that something in the way the upper lip is involved in speech production may be different (as the effect is not likely to be related to acoustic signals). Recent work in our lab on different populations with speech motor problems have found that the involvement of upper lip movements may increase in speech production as a potential source for movement stabilization through feedback entrainment (Namasivayam, van Lieshout, & De Nil, 2008; van Lieshout, Bose, Square, & Steele, 2007). It is possible that something similar is happening with our PD subjects, but further analysis will have to confirm this.

In conclusion, our study for the first time provides evidence that the relationship between speech acoustics and facial motion is reduced in people with PD relative to age-matched controls. It is likely that such a reduction in audio-visual congruency may have an impact on speech understanding and may form part of the well-known limitations in intelligibility found in this population (Miller et al., 2007). If future studies can confirm this relationship, it may provide new ways to test intelligibility in more objective ways and also, to develop better tools in support of facial motion enhancement as part of speech rehabilitation.

Acknowledgements

This research was supported in part, by the Canada Research Chairs Program awarded to the second author. The authors gratefully acknowledge the assistance of Aravind Namasivayam with data collection and analysis.

References

Cheang, H. S., & Pell, M. D. (2007). An acoustic investigation of parkinsonian speech in linguistic and emotional contexts. *Journal of Neurolinguistics, 20*(3), 221-241.

Craig, M., van Lieshout, P. H. H. M., & Wong, W. (2007). Suitability of a UV-based video recording system for the analysis of small facial motions during speech. *Speech Communication, 49*(9), 679-686.

Craig, M. S., van Lieshout, P. H. H. M., & Wong, W. (Under Review). A linear model of acoustic-to-facial mapping: Model parameters, data set size and generalization across speakers. *The Journal of the Acoustical Society of America*

Goldstein, L. M., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. A. Arbib (Ed.), *From action to language: The mirror neuron system* (pp. 215-249). Cambridge, UK: Cambridge University Press.

Hintze, J. (2007). *NCSS, PASS, and GESS*. Kaysville, Utah, USA.

Miller, N., Allcock, L., Jones, D., Noble, E., Hildreth, A. J., & Burn, D. J. (2007). Prevalence and pattern of perceived intelligibility changes in parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry, 78*(11), 1188-1190.

Namasivayam, A. K., van Lieshout, P., & De Nil, L. (2008). Bite-block perturbation in people who stutter: Immediate compensatory and delayed adaptive processes. *Journal of Communication Disorders, 41*(4), 372-394.

Pinto, S., Ozsancak, C., Tripoliti, E., Thobois, S., Limousin-Dowsey, P., & Auzou, P. (2004). Treatments for dysarthria in parkinson's disease. *Lancet Neurology, 3*(9), 547-556.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects: Research report. *Psychological Science, 18*(5), 392-396.

Smith, M. C., Ellgring, H., & Smith, M. K. (1996). Spontaneous and posed facial expression in parkinson's disease. *Journal of the International Neuropsychological Society, 2*(5), 383-391.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*(1-2), 3-45.

Tickle-Degnen, L., & Lyons, K. D. (2004). Practitioners' impressions of patients with parkinson's disease: The social ecology of the expressive mask. *Social Science and Medicine, 58*(3), 603-614.

van Lieshout, P. H. H. M., Bose, A., Square, P., & Steele, C. (2007). Speech motor control in fluent and dysfluent speech production of an individual with apraxia of speech and broca's aphasia. *Clinical Linguistics and Phonetics, 21*(3), 159-188.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1-2), 23-43.

FURTHER EVIDENCE OF TOP-DOWN GAIN CONTROL IN THE AUDITORY SYSTEM

Bruce A. Schneider[1] and Scott Parker[2]
[1]*Centre for Research on Biological Communication Systems*
University of Toronto Mississauga
Bruce.Schneider@utoronto.ca
[2]*Department of Psychology*
The American University
sparker@american.edu

Abstract

We have repeatedly proposed that a non-linear gain-control system under top-down governance regulates the encoding of auditory intensity. Here we examine the effects of individual differences and continued training on the identification of tones differing only in intensity. In a baseline condition, participants were asked to identify which one of four 1-kHztones (25, 30, 35, or 40 dB) was presented on a trial. In two other conditions an additional tone(50 or 80 dB) was added to the baseline set. Each of the three conditions was tested over the course of 20 sessions, each stimulus occurring 50 times per session. Performance improved over sessions with large individual differences among participants. However, in all cases, the pattern of change across conditions was consistent with the notion that the encoding of intensity along the decision axis is regulated by a non-linear gain control mechanism.

In previous papers (Gordon & Schneider, 2007; Parker, Murphy & Schneider, 2002; Parker & Schneider, 1994; Schneider & Parker, 1990) we have argued for the existence of a nonlinear gain control mechanism in the auditory system that is under top down control. This gain control mechanism has two functions: (1) to protect the auditory system from sensory overload; and (2) to maximize the discriminability of the acoustic stimuli in the listener's current soundscape (the range of sounds that the listener has recently or is currently experiencing, or expects to hear in the near future). Figure 1 presents a model of how this system might operate.

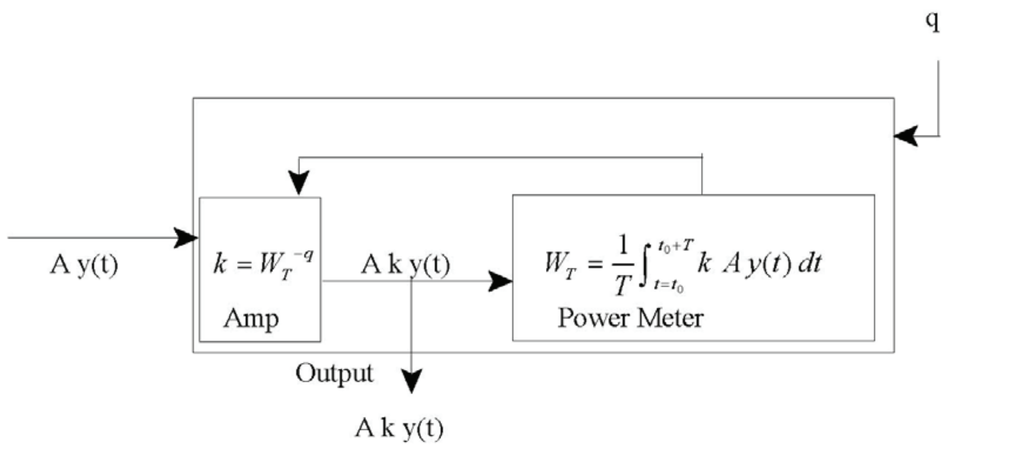

Figure 1. A non-linear gain control mechanism based on a feedback loop. The input signal is amplified by a factor k, whose value is controlled by the power at the output of the amplifier.