Audio and Vision-Based Evaluation of Parkinson's Disease from Discourse Video

Francis Quek, Robert Bryll,
Vision Interfaces & Sys. Lab.
CSE Dept., Wright State U.
Dayton, OH

Mary Harper, Lei Chen,
Purdue University
West Lafayette, IN

Lorraine Ramig
University of Colorado, Boulder
Boulder, CO

Correspondence: quek@cs.wright.edu

Abstract

Parkinson's disease (PD) belongs to a class of neurodegenerative diseases that affect both the patient's speech and motor capabilities. To date, PD diagnosis and the determination of disease progress and treatment efficacy is based entirely on the subjective observation of a trained physician. We present the results of a pilot study of two Idiopathic PD patients who have undergone Lee Silverman Voice Treatment (LSVT). It has been observed subjectively that gestural performance of patients improve in tandem with speech improvements after LSVT. It is hypothesized that these improvements are taking place at a neurological level. Measurements of speech and gesture suggest that LSVT improves the quality of both gesticulation and speech.

1 Introduction

Parkinson's disease (PD) belongs to a class of neurodegenerative diseases that affect both the patient's speech and motor capabilities. Currently there are one and a half million sufferers of the disease and this number is expected to rise fourfold by 2040 [1]. Parkinson's Disease (PD) is a progressive neurodegenerative disorder in which death of dopaminergic cells in the substantia nigra results in a variety of changes in motoric function such as delay in initiation of movement, increase in resting muscle tone, slowness of movement, and resting tremor [2, 3]. Cognitive changes, such as slowed information processing (bradyphrenia) [4], and loss of postural reflexes may also occur [5]. Speech changes are common; approximately 70% of Parkinson's patients have speech problems. Parkinson's patients often exhibit monotonous pitch and loudness, reduced stress, variable speech rate, short rushes of speech, and imprecise consonant articulation. They also manifest hypophonia, start-hesitation or stuttering, or delay in the production of speech, which is often difficult to distinguish from bradyphrenia [6, 7].

There are currently no widely available tests for PD. Diagnosis and assessment of progression of the disease are done on the basis of serial clinical examinations, relying on the subjective evaluation of a trained physician, often using the modified Hoehn and Yahr staging and the United PD Rating Scale (UPDRS) [8, 2].

In this paper, we present research in which we analyzed the before and after treatment videos of two PD patients who underwent a treatment program known as *Lee Silverman Voice Treatment* (LSVT) [9]. This treatment applies voice therapy methods to improve the speech intelligibility of PD patients. Subjective observations indicate that not only did speech qualitatively improve, the accompanying gesticulation appeared to improve as well. We present a set of metrics that we employed to detect this subjective observation.

2 Background

The reduced ability to communicate is considered to be one of the most difficult aspects of Idiopathic Parkinson's Disease (IPD) by many IPD sufferers and their families. Soft voice, monotone, breathiness, hoarse voice quality, and imprecise articulation, together with lessened (masked) facial expression, contribute to limitations in communication in the vast majority of individuals with IPD [10, 11].

The initial treatment aim of LSVT is to improve the phonatory source in individuals with IPD. The treatment typically results in significant, long-term improvement in laryngeal valving and post-treatment changes in thyroarytenoid muscle activity, subglottal air pressure, maximum flow declination rate, voice sound pressure level (SPL), loudness and voice quality [12, 13, 14].

LSVT uses phonation as a trigger to increase effort and coordination across the speech production system through stimulating the global variable "loud." Speech production is a learned, highly practiced motor behavior, with many of its movements regulated in a quasiautomatic fashion [15, 16]; loudness scaling is a task that humans engage in all their lives [17, 18]. For example, it is common to increase loudness to improve speech intelligibility when speaking against noise or when the listener is far away. By targeting loud-

ness in treatment, well-established, centrally stored motor patterns for speech may be triggered; that is, intensive loudness training may provide the stimulation needed for the individual with IPD to activate and modulate appropriate speech motor programs that are still intact. Such multilevel upscaling across the speech system is likely to involve common central neural pathways. Further, post-treatment observations of increased facial expression accompanying improved loudness and intonation [19], and the results of a PET brain imaging study of individuals with IPD pre- and post-LSVT [20] suggest that training loud phonation may also promote the recruitment of the right insular cortex and the anterior cingulate cortex.

Taken together, this suggest that there is an element of spreading of the gains accrued in improved phonation to other emotional and motoric activity at a neuronal level. This suggests that the observed gesticulatory 'improvement' may be a result of this neuronal level spreading. This is congruent with the basic assertion in our gesture research that gestural behavior and speech share a common semantic source and are tightly integrated [21, 22, 23, 24].

3 The Experiment

Our pilot dataset was obtained from three patients (two males and a female) who underwent an 8-week LSVT protocol involving therapy for 1 hour 2 days a week over the period. Patients have been stabilized on their anti-Parkinsonian medication prior LSVT so that behavioral changes are not attributable to medication adjustments. Each patient performed a videotaped narration before and after the treatment. The experimental protocol employed was the standard 'Tweety and Sylvester' narrations that are a staple of gesture-speech research [25]. The patient viewed a series of Tweety-and-Sylvester cartoon clips and were instructed to convey the contents of the clips to an interlocutor so that she can then go and tell it to someone else. The narration was videotaped with a single superVHS camera for analysis. The audio was recorded using a AKG C451 EB boom mounted microphone approximately 2 feet from the subject. The audio was recorded through an amplifier connected to a Panasonic SVHS AG-1960 VCR. The video was digitized along on an Silicon Graphics O2 in SGI-MJPEG format. Audio was digitized at the same time at 44.1KHz and then downsampled to 14.7 KHz for analysis.

Of the three patients, one (the female) turned out to have a Parkinson's Plus condition (instead of IPD) and did not respond to treatment. In fact, she deteriorated so much that there was hardly any motion in the 'after' video. We shall designate the other two patients P1 and P2. They were 61 and 60 years of age respectively. P1 and P2 has a Hoehn & Yahr stage II and stage III diagnoses respectively. Both men have been diagnosed with IPD for 5 years.

Figure 1: Vectors extracted in an experiment video frame

P1 and P2 produced datasets of 275.3 sec (8259 frames) and 336.033 sec (1081 frames) in before narration video for P1 and P2 respectively; and 219.1 sec (6573 frames) and 291.867 sec (8756 frames) in after-treatment narration video for P1 and P2 respectively. Figure 1 shows a video frame of our experimental video (for P2 after treatment in this case) after the video processing to be described subsequently. The patient's face has been blacked out for privacy reasons.

We will evaluate whether there are differences in the speech and gestures of P1 and P2 before and after LSVT.

4 Speech Analysis

Each subject's speech before and after treatment was processed to extract a variety of measurements. Since Parkinson's patients often exhibit monotonous loudness and pitch, we obtained the average and standard deviation of the intensity and F_0 of the speech signal using the Praat tool for each speech sample [26]. Intensity was extracted from the speech signal with a time step of 0.01 seconds. The F_0 values were extracted using autocorrelation with a time step of 0.01 seconds, 75 Hz as the minimum pitch, 600 Hz as the maximum, and the default settings for the thresholds [27]. The perception of monotonous pitch may result from the fact that there is little variation in pitch over a limited time window; hence, we also calculated the average change in pitch from one frame to the next (where a frame is 0.01 seconds) and the average change in pitch over an interval of voiced speech (where change is defined as the difference between the minimum and the maximum F_0 value in the interval of speech). Since Parkinson's patients are perceived to produce short rushes of speech, we measured the average length of voiced speech intervals and unvoiced speech intervals and their standard deviation. Finally, we measured the average jitter of each speech sample, as well as the stan-

Table 1: Before and after treatment measurements on the speech of P1 and P2.

Measure	P1 Before	P1 After	P2 Before	P2 After
Avg. Intensity	63.885	63.628	64.291	62.651
Std. Dev. Intensity	6.287	10.808	9.783	11.742
Avg. F_0	125.385	145.249	135.554	140.179
Std. Dev. F_0	18.535	22.614	36.257	22.576
% Voiced	49.104%	58.856%	44.768%	51.167%
Avg. Voiced Length	17.591	23.525	27.057	31.440
Std. Dev. Voiced Length	14.673	19.381	24.574	26.927
Avg. Unvoiced Length	18.214	16.446	33.309	30.090
Std. Dev. Unvoiced Length	24.743	27.064	60.847	54.566
Avg. Δ F_0 **Frame**	1.929	2.414	2.572	2.503
Std. Dev. Δ F_0 **Frame**	2.377	2.765	3.722	3.376
Avg. Δ F_0 **Interval**	19.135	29.844	33.249	36.507
Std. Dev. Δ F_0 **Interval**	12.941	16.975	20.698	20.620
Avg. Jitter	0.016	0.010	0.017	0.009
Std. Dev. Jitter	0.015	0.0008	0.014	0.008

dard deviation. Jitter is a measure of perturbation in the pitch period that has been used by speech pathologists to identify pathological speech [26, 28, 29]; a value of 0.01 corresponds to a jitter of one percent, which represents a lower bound for abnormal speech. The measurements of each patient before and after treatment are shown in Table 1.

We use an F-test to determine whether there are significant differences in the variance before and after treatment, and a two sample t-test to determine whether there is a significant difference in the means of the before and after samples. As can be seen in the table, there is no increase in the average intensity for P1 and P2 after treatment; however, there is an increase in the standard deviation of the intensity. An F-test shows that the variances of the before and after treatment samples are significantly different ($p < 0.001$), suggesting that the treatment increases the variation of loudness in speech. The average value of F_0 increases significantly after treatment for both of the patients ($p < 0.001$); however, there is no consistent pattern for the standard deviation. The percentage of voiced speech increases after treatment. Additionally, there is a significant increase in the average length ($p < 0.05$) of voiced speech and a consistent increase in standard deviation (although there is no significant difference in the variance of the before and after treatment samples for P2). There is a consistent (although insignificant) decrease in the average length of voiceless speech, although there is no consistent pattern on the standard deviation. The average change in F_0 over adjacent frames does not change consistently for both patients, although the standard deviation of the difference does increase (with the F-test indicating that the variances of the before and after treatment samples are significantly different ($p < 0.001$)). The average difference between the maximum F_0 value in running voiced speech and the minimum does increase significantly after treatment for both patients ($p < 0.05$), although there is no consistent pattern for the standard deviation. Finally the average jitter decreases significantly after treatment for both patients ($p < 0.01$), and the standard deviation also decreases consistently. An F-test shows that the jitter variances of the before and after treatment samples are significantly different ($p < 0.01$).

These measurement data suggest that LSVT increases the variability in loudness, increases the average F_0 value of speech, and increases the range of F_0 excursions in a running stretch of voiced speech. Additionally, there is a post-treatment increase in the percentage of voiced speech along with an increase in the average length of a voiced stretch of speech. This suggests that the speakers have increased the length of spoken vowels indicative of speaking more slowly and deliberately. Additionally, the speakers post treatment have much less jitter in their speech.

5 Gesture Analysis

5.1 Video Preprocessing

We processed each video dataset using our *Vector Coherence Mapping* (VCM) [30, 31, 32] system to extract the vector fields describing the subjects' hand motions. VCM applies a parallelizable fuzzy image processing algorithm, and is used to track unadorned hands in video sequences consisting of tens of thousands of frames in for our re-

search on the gestural correlates to natural discourse events [33, 24, 34, 21, 31, 22, 23, 35, 36, 32]. VCM combines the correlation and constraint-based smoothing processes into a set of fuzzy image processing operations. The algorithm is completely parallel and obviates the iterative post-process used by most optical flow algorithms. VCM performs a voting process in vector parameter space and biases this voting by likelihood distributions that enforce the spatial and temporal constraints. Hence, VCM is similar to the Hough based approaches [37, 38, 39, 40]. The difference is that in VCM, the voting is distributed and the constraints enforced on each vector are local to the region of the vector. Furthermore, in VCM the correlation and constraint enforcement functions are integrated in such a way that the constraints "guide" the correlation process with the likelihood distribution. Hough methods, on the other hand, apply a global voting space. Bulthoff et al [41] describe a neurally-inspired approach that is similar to VCM in that it uses a set of locally-summed correlations to compute a vector field at high contrast image points. The difference between their system and VCM is that VCM is based on a broader fuzzy model that facilitates the enforcement of a variety of constraints (e.g. momentum, color similarity etc.). Our results show that VCM has good noise immunity. The robustness of VCM lies in the fact that correlation errors owing to noise occur in image space, and have little support in the parameter space of the vectors. Please see [30, 31, 32] for detailed discussion of this algorithm.

Two characteristics of the VCM feature selection process are pertinent to this paper. First, VCM uses a variant of the moving edge detector in [42] that emphasizes moving edge points. This detector takes the product of the spatial and temporal gradients to find image points where both are high. In our experiments, we found that this detector fails where the image gradient is low, as would happen when the hand moves over regions with similar image intensity. Taking a product of this image gradient with the temporal gradient depresses influence of the temporal gradient. Our modified detector takes a weighted sum of spatial and temporal greyscale gradients:

$$F(I) = w_S|S(I)| + w_t\frac{\partial I}{\partial t} \quad (1)$$

Where $|S(I)|$ is the magnitude of the Sobel gradient of the greyscale image, and w_S and w_t are the spatial and temporal gradient weights respectively. This allows us to bias our computation to dynamic image regions where flow may be detected.

Second, to further focus our computation on hands and face, we accept only the interest points that belong to moving edges of skin-colored image regions. We achieve this by creating a *skin similarity map* for each frame and selecting only the interest points (moving edge points) for which the skin similarity is larger than a specified threshold. We begin by obtaining the color signature of the subject's skin from 20 to 50 points selected manually on the subject's skin regions in the video frames. We use the Normalized RGB color space where $\mathcal{N}(RGB)$ color space is obtained by: $\mathcal{N}(RGB) = (r, g) = \left(\frac{R}{R+G+B+1}, \frac{G}{R+G+B+1}\right)$. This gives us independence from the illumination color and the camera white balance settings. We experimented with a general "skin-color database" but we found that the detection results are much more robust when the samples are selected for each sequence. The $\mathcal{N}(RGB)$ vectors of the manually selected skin pixels are used as a reference cluster. For each pixel $p_i \in I$, the value in the *skin similarity map* ($\mathcal{S}(I)$) of image I is

$$\min\left(\max\left(|r_n - r(p_i)|, |g_n - g(p_i)|\right)\right) \forall n = 1 \ldots N \quad (2)$$

Where N is the number of samples in the reference cluster, r_n and g_n are the normalized r and g values of the n^{th} sample respectively, and $r(p_i)$ and $g(p_i)$ are the normalized r and g values of image point p_i respectively. Equation 2 yields the distance between the color of each image pixel to its nearest reference cluster value in $\mathcal{N}(RGB)$ space. Conservatively, we take the greater of the r and g differences as this distance.

Color similarity information is typically most stable within homogeneous color regions. At the boundaries, pixel averaging and motion effects make color detection difficult. Since we use color only to help identify interest points for tracking with VCM, it is not critical that we identify the exact color boundary. We must, however, ensure that the pertinent interest points are not elided. Hence, we grow the strong skin color similarity regions to cover the location of the moving edge pixels detected using equation 1. This ensures that moving edge points bordering regions of strong skin color similarity are not missed. We achieve this by applying an asymmetric morphological 'opening' operation to the strong skin similarity regions of the image. An opening operator comprises morphological erosion using some structuring element, followed by a dilation operation. Morphological 'opening' operators shrink regions of interest. To prevent this, we use a dilation kernel that is twice the size of the erosion kernel. This achieves the effect of growing larger high skin similarity regions while removing small regions with high skin color similarity, creating a mask that covers all skin color region edge pixels, but excludes other edge pixels.

Together, these feature selection criteria ensures that the number of vectors detected varies positively with the amount and degree of motion of skin color regions – particularly the hands – in the video. The vectors extracted for P2's moving right hand are shown in figure 1.

5.2 'Motion Quality' Indices

We tested two sets of motion-related measures from the vector fields extracted from the experimental video datasets. The goal is to compare the quality of gesticulation before and after treatment. We shall discuss these and their performance on our pilot data in turn. We shall present statistics on the measure that we found most reliable in the Statistical Tests section.

5.2.1 Total and Degree of Motion

Given the characteristics of the feature selection process in our VCM system, an obvious measure for the quantity of motion is the ratio of the total number of vectors detected in the entire sequence to the number of frames in the sequence. Since more motion produces more vectors, as does faster motion, this is a kind of *Total Motion Index*, TMI, where:

$$\text{TMI} = \frac{\text{Total number of vectors}}{\text{Total number of frames}} \tag{3}$$

The second measure we tested is related to the degree of motion. The average length of the vectors extracted gives the average speed of hand motions. Since hand holds are part of gesture activity, one has to remove stationary frames from the computation of this average. A frame is judged to have no motion if there is fewer than T_A vectors in it (we use a value of $T_A = 4$). Hence we define a *Degree of Motion Index*, DMI as:

$$\text{DMI} = \frac{\sum_{i=1}^{N_A} \overline{v}_i}{N_A} \tag{4}$$

Where $\overline{v}_i$ is the average vector length in frame i and N_A is the number of active frames.

Our third measure is the average number of vectors per active frame, $\overline{N}_A$. This is similar to the TMI, except that it considers only active frames. It measures the total activity in such active frames. This measure could be affected, for example, if the patient uses more two handed gestures or moves the entire hand instead of just changing the shape of a stationary hand.

The fourth measure we considered was the average acceleration per active frame, AAF, which is given by:

$$\text{AAF} = \frac{\sum_{i=1}^{N_{active}} \overline{a}_i}{N_{active}} \tag{5}$$

Where $\overline{a}_i$ is the average acceleration length in frame i computed as a difference in velocities of consecutive frames.

As can be seen in table 2, all these indices improved for P1 after treatment as compared with before.

The problem is that for P2, we got the reverse effect in all indices (table 5.2.1). Upon closer examination of the experimental video, the problem was traced to the inconsistent use of armrests in before and after narration experiments. The experiment was performed with the patient in a therapy chair with adjustable armrests (they may be engaged or disengaged). Figure 1 shows a frame of the after-treatment video for P2. Notice that the armrest are engaged. In the 'before' video, the armrests were disengaged (i.e. not used). The patient chose his laps as the 'rest' position for his hands, and consequently has to move up to the 'gesture space' to do any gesture. In the 'after' video, the armrests were engaged (i.e. in operation). The patient pivoted his elbows on the armrests and clasped his hands close to his mid-section/chest when not gesturing. As can be seen in figure 1, this resulted in movements to the gesture space being shorter, decreasing both amount and degree of motion.

P1 also had his armrests engaged in the 'after' condition and disengaged in the 'before' condition. His rest position for the no-armrest video was his lap. For the with-armrest video, his hands rested limply on the armrests. Hence, although the with-armrest rest position was closer to the gesture space, it was not as close as for P1. Because of this, we were able to detect the greater amount and degree of motion after. We suspect that the difference would have been more dramatic if both 'before' and 'after' videos employed the no-armrest condition.

P1	Before-tx	After-Tx	Change
No. of frames	8259	6573	
TMI	131.16	172.81	31.7%
N_A	3295	2669	
$\overline{N}_A$	328.756	425.59	29.45%
DMI	3.515	3.86	9.8%
AAF	0.816	1.00	22.5%

Table 2: Patient 1: Total and Degree of Motion

P2	Before-tx	After-Tx	Change
No. of frames	10081	8756	
TMI	485.73	354.88	-26.9%
N_A	7634	7168	
$\overline{N}_A$	641.43	433.50	-32.4%
DMI	5.10	3.17	-37.8%
AAF	1.28	1.04	-18.7%

Table 3: Patient 2: Total and Degree of Motion

5.2.2 Liveliness Index

Although our first set of measures produced conflicting results because of the different use of armrests, observers can nevertheless discern a difference in the gesticulation for

	Before	After	Change
P1	0.28095	0.369883	31.65%
P2	0.338495	0.409115	20.86%

Table 4: Liveliness indices for both patients

both patients. We needed an index that is more inoculated to the variation in distance between the patient's resting hand positions and the gesture space.

We define a *Liveliness Index*:

$$\mathrm{LI} = \frac{\sum_{i=1}^{N_A} \frac{\overline{a}_i}{\overline{v}_i}}{N_A} \tag{6}$$

This is essential the average of framewise ratio between the average acceleration per frame to the average velocity per frame. Intuitively, LI measures instantaneous acceleration against instantaneous velocity, giving a sense of the peppiness of the gesticulation.

Table 4 shows that we obtained a consistent increase in LI for both patients (31.65% for P1 and 20.86% for P2). We proceed to analyze the significance of this finding.

5.3 Statistical Tests

The variance of LI's for both patients before therapy (left column of Table 4) is 0.0016557, the variance after (right column in Table 4) is 0.00076957. The F-ratio of the two variances is $F = 2.15$, which is smaller than critical F for any confidence level for one degree of freedom in the numerator and denominator. Therefore, based on this F-test we can conclude that the variances of the before and after LI's for the two patients are not significantly different, so both indices come from these are population and a t-test can be performed on the sample.

The Before-After LI differences for both patients are:

P1	-0.088933
P2	-0.07062

Table 5: Before-After Liveliness Index differences for both patients

For the above data, the average Before-After change is: $\overline{\triangle} = -0.0797765$. The standard deviation $\sigma_{\triangle} = 0.012949$.

In a paired t-test the t value for the above data is:

$$t = \frac{\overline{\triangle}}{\sigma_{\triangle}/\sqrt{2}} = -8.71255 \tag{7}$$

Our null hypothesis H_0 is that the average Before-After change for the population $\mu_{\triangle} >= 0$, meaning no improvement in the LI. The H_1 hypothesis: $\mu_{\triangle} < 0$ (there is improvement in the LI).

¿From the t-table, for a left-tailed test, $t_{\alpha=0.05,1} = 6.314$. Since $t < -t_{\alpha=0.05,1}$ we can reject the null hypothesis and claim with 95% confidence that the therapy improves the Liveliness Index.

6 Summary and Conclusions

We have presented the results of our pilot work on speech and vision-based gesture metrics to assess improvement in IPD patients after LSVT. Changes in F_0, intensity, duration of voicing, and jitter were consistent with subjective judgements of improved speech quality post-treatment. Our *Liveliness Index* appears to provide a tool for comparison of gesticulatory motion accompanying speech before and after treatment. We believe this represents the promise of the application of vision in the study of IPD-induced and other motion disorders.

We stress that this is a pilot study with only data on two patients. Although we found statistically significant changes in the signals signals, we would like higher confidences. More data will permit us to make stronger statements. This is the subject of our continuing research as is the discovery of more robust metrics for diagnosis.

We are also proceeding to transcribe and perform detailed temporal analysis of the speech and gesture. This will allow us to examine the varying dissynchronies of gesture and speech as the disease progresses or as a result of treatment, using a spatially intensive narrative task. The natural variability of gesture output should not interfere with precise dissynchrony comparisons. We hypothesize that the study of such dissynchronies would permit us to develop a better understanding of the speech and gesture effect. We envision this research will lead to new quantitative ways to diagnose IPD or to measure medication/treatment response.

Acknowledgments

This research has been partially supported by the U.S. National Science Foundation STIMULATE program, Grant No. IRI-9618887, "Gesture, Speech, and Gaze in Discourse Segmentation" and the National Science Foundation KDI program, Grant No. BCS-9980054, "Cross-Modal Analysis of Signal and Sense: Multimedia Corpora and Tools for Gesture, Speech, and Gaze Research". This research was also supported in part by NIH-NIDCD R01 DC-01150. Appreciation is expressed to subjects who participated in this study.

References

[1] http://www.pdindex.com/, "PD INDEX: A directory of parkinson's disease information on the internet", Internet Publication.

[2] J. Jankovic, "Pathophysiology and clinical assessment of motor symptoms in parkinsons disease", in W. Koller, editor, *Handbook of Parkinsons disease*, pp. 99–126. Marcel Dekker, New York, 1987.

[3] M. Hallett and S. Khoshbin, "A physiological mechanism of bradykinesia", *Brain*, vol. 103, pp. 301–314, 1980.

[4] B. Dubois. and B. Pillon., "Cognitive deficits in Parkinson's disease", *J Neurol.*, vol. 244, pp. 2–8, Jan. 1997.

[5] M. Morris, R. Iansek, F. Smithson, and F. Huxham, "Postural instability in parkinson's disease: a comparison withand without a concurrent task", *Gait Posture*, vol. 12, pp. 205–216, Dec. 2000.

[6] K. Forrest, G. Weismer, and G. Turner, "Kinematic, acoustic, and perceptual analyses of connected speech produced by parkinsonian and normal geriatric adults", *Journal of the Acoustical Society of America*, vol. 85, pp. 2608–2622, 1989.

[7] J. Logemann, H. Fisher, B. Boshes, and E. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson individuals", *Journal of Speech and Hearing Disorders*, vol. 42, pp. 47–57, 1978.

[8] M.M. Hoehn and M.D. Yahr, "Parkinsonism: Onset, progression and mortality", *Neurology*, vol. 17, pp. 427–442, 1967.

[9] L. Ramig, S. Countryman, L. Thompson, and Y. Horii, "A comparison of two forms of intensive speech treatment for parkinson disease", *Journal of Speech and Hearing Research*, vol. 38, pp. 1232–1251, 1995.

[10] T. Pitcairn, S. Clemie, J. Gray, and B. Pentland, "Nonverbal cues in the self-presentation of parkinsonian patients", *British Journal of Clinical Psychology*, vol. 29, pp. 177–184, 1990.

[11] T. Pitcairn, S. Clemie, J. Gray, and B. Pentland, "Impressions of parkinsonian patients from their recorded voices", *British Journal of Disorders of Communication*, vol. 25, pp. 85–92, 1990.

[12] C. Baumgartner, S. Sapir, and L. Ramig, "Perceptual voice quality changes following phonatory-respiratory effort treatment (LSVT) vs. respiratory effort treatment for individuals with parkinson disease", *Journal of Voice*, vol. in press, 2001.

[13] L. Ramig and C. Dromey, "Aerodynamic mechanisms underlying treatment-related changes in vocal intensity in patients with parkinson disease", *Journal of Speech and Hearing Research*, vol. 39, pp. 798–807, 1996.

[14] L. Ramig, S. Sapir, K. Baker, and M. Smith, "Electromyographic changes in laryngeal muscle activity following LSVT in idiopathic parkinson disease", *Journal of Speech, Language, and Hearing Research*, vol. in review, 2001.

[15] H. Ackermann, D. Wildgruber, I. Daum, and W. Grodd, "Does the cerebellum contribute to cognitive aspects of speech production? a functional magnetic resonance imaging (fMRI) study in humans", *Neuroscience letters*, vol. 247, pp. 187–90, 1998.

[16] S. Hirano, H. Kojima, Y. Naito, I. Honjo, Y. Kamoto, H. Okazawa, K. Ishizu, Y. Yonekura, Y. Nagahama, H. Fukuyama, and J. Konishi, "Cortical processing mechanism for vocalization with auditory verbal feedback", *Neuroreport*, vol. 8, pp. 2379–82, 1997.

[17] A. Ho, J. Bradshaw, R. Iansek, and R. Alfredson, "Speech volume regulation in parkinson's disease: effects of implicit cues and explicit instructions", *Neuropsychologia*, vol. 37, pp. 1453–60, 1999.

[18] AK Ho, JL Bradshaw, and T Iansek, "Volume perception in parkinsonian speech", *Movement Disorders*, vol. 15, pp. 1125–31, 2000.

[19] J. Spielman, L. Ramig, and J. Borod, "Preliminary effects of voice therapy on facial expression in parkinsons disease", *Journal of the International Neuropsychological Association*, vol. 7, pp. 244, 2001.

[20] M. Liotti, D. Vogel, L. Ramig, P. New, C. Cook, and P. Fox, "Functional reorganization of speech-motor function in parkinson disease following LSVT: A PET study", *Neurology*, vol. in review, 2001.

[21] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K-E. McCullough, "Gesture cues for conversational interaction in monocular video", *in ICCV'99 Wksp on RATFG-RTS.*, pp. 64–69, Corfu, Greece, Sep. 26–27 1999.

[22] F. Quek, McNeill, R. D., Bryll, C. Kirbas, H. Arslan, K-E. McCullough, N. Furuyama, and R. Ansari, "Gesture, speech, and gaze cues for discourse segmentation", *in Proc. of the IEEE Conf. on CVPR*, vol. 2, p. 247254, Hilton Head Island, South Carolina, June 13-15 2000.

[23] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K-E. McCullough, "Gesture and speech cues for conversational interaction", *ToCHI*, vol. in review, 2001, VISLab,

Wright State U., Tech. Report VISLab-01-01, http://vislab.cs.wright.edu/Publications/Queetal01.html.

[24] D. McNeill, F. Quek, K.-E. McCullough, S. Duncan, N. Furuyama, R. Bryll, X.-F. Ma, and R. Ansari, "Catchments, prosody and discourse", *in in press: Gesture*, 2001.

[25] D. McNeill, *Hand and Mind: What Gestures Reveal about thought*, U. Chicago Press, Chicago, 1992.

[26] Paul Boersma and David Weenink, "Praat, a system for doing phonetics by computer", version 3.4., 1996.

[27] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *in Proceedings of the Institute of Phonetic Sciences Computational Linguistics*, vol. 17, pp. 97–110, 1993.

[28] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, Inc., 2000.

[29] A. E. Rosenberg, "The effect of glottal pulse shape on the quality of natural vowels", *Journal of The Acoustical Society of America*, vol. 49, pp. 583–590, 1970.

[30] F. Quek and R. Bryll, "Vector Coherence Mapping: A parallelizable approach to image flow computation", *in ACCV*, vol. 2, pp. 591–598, Hong Kong, Jan. 1998.

[31] F. Quek, X. Ma, and R. Bryll, "A parallel algorithm for dynamic gesture tracking", *in ICCV'99 Wksp on RATFG-RTS.*, pp. 119–126, Corfu, Greece, Sep.26–27 1999.

[32] R. Bryll and F. Quek, "Vector-based tracking of conversational gestures", Technical report, Vision Interfaces and Systems Lab, Wright State U. OH, USA, 2001, VISLab Report: VISLab-01-08.

[33] R. Ansari, Y. Dai, J. Lou, D. McNeill, and F. Quek, "Representation of prosodic structure in speech using nonlinear methods", *in Wksp Nonlin. Sig. & Image Proc.*, Antalya, TU, 1999.

[34] F. Quek, R. Bryll, H. Arslan, C. Kirbas, and D. McNeill, "A multimedia database system for temporally situated perceptual psycholinguistic analysis", *Multimedia Tools & Apps.*, vol. In Press, 2001.

[35] D. McNeill and F. Quek, "Gesture and speech multimodal conversational interaction in monocular video", *in Measuring Behavior 2000*, p. 215, Nijmegen, NL, Aug. 15–18 2000.

[36] F. Quek and D. McNeill, "A multimedia system for temporally situated perceptual psycholinguistic analysis", *in Measuring Behavior 2000*, p. 257, Nijmegen, NL, Aug. 15–18 2000.

[37] G. Adiv, "Determining three-dimensional motion and structure from optical flow generated by several moving objects", *PAMI*, vol. 7, pp. 384–401, 1985.

[38] D. Ballard and O. Kimball, "Rigid body motion from depth and optical flow", *CVGIP*, vol. 22, pp. 95–115, 1983.

[39] T.Y. Tian and M. Shah, "Recovering 3D motion of multiple objects using adaptive Hough transform", *in Proc. 5th ICCV*, MIT, Cambridge, MA, June 20-23, 1995.

[40] M. Bober and J. Kittler, "Robust motion analysis", *in Proc. of the IEEE Conf. on CVPR*, Seattle, WA, June 21-23, 1994.

[41] H. Bulthoff, J. Little, and T. Poggio, "A parallel algorithm for real-time computation of optical flow", *Nature*, vol. 337, pp. 549–553, Feb. 9th, 1989.

[42] S.M. Haynes and R. Jain, "Time-varying edge detection", *CGIP*, vol. 21, pp. 345–393, 1983.