SPEECH PAUSES AND GESTURAL HOLDS IN PARKINSON’S DISEASE

*Francis Quek, 1Mary Harper, *Yonca Hacihmetoglu, 1Lei Chen, and 1Lorraine Ramig

*Vision Interfaces & Sys. Lab. (VISLab) CSE Dept., Wright State U., OH
1ECE Dept., Purdue University, IN , 1University of Colorado, Boulder

Correspondence: quek@cs.wright.edu

Abstract

Parkinson’s disease (PD) belongs to a class of neurodegenerative diseases that affect the patient’s speech, motor, and cognitive capabilities. All three deficits affect the multimodal communication channels of speech and gesture. We present a study on the changes in speech pause patterns and gesture holds before and after treatment. We present the results of a pilot study of two Idiopathic PD patients who have undergone Lee Silverman Voice Treatment (LSVT). We show that there was a consistent change in the location of pauses with respect to semantic sentential utterance units. After treatment, the number and duration of pauses within sentential units decreased while the inter sentential pauses increased. This indicates reduction in hesitation and increase in speech phrasing. We also found a decrease in the number of sentence repairs and the time spent in repairs. For gesture, we found that non-rest holds intermixing with within-sentence pauses appears to decline after treatment, as does the ratio of rest holds during speech against rest holds between sentences. While the work is preliminary, these patterns suggest that multimodal discourse characteristics might provide access to the underlying cognitive state under the load of narrative discourse.

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].

Earlier studies of voice onset time [8] and the effects of PD on naming deficits and language comprehension [9] suggest that language deficits may provide a handle on understanding the neuropathology of the disease. Such studies seek to understand these component effects in isolation. We believe that there are effects of PD that reveal themselves in individuals only under the load complexity of natural discourse. This paper tests the proposition that these effects are accessible via the analysis in multimodal narrative discourse involving gesture and speech. Specifically we study the effects of the disease on speech pause and gestural hold distribution. We analyzed the before and after treatment videos of two PD patients who underwent a treatment program known as Lee Silverman Voice Treatment (LSVT) [10]. 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. Our earlier studies have shown that there is an improvement in overall ‘liveliness’ motion quality [11] and in the dynamics of both motion and speech intensity variation [12, 13].

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, breathlessness, 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 [14, 15].

The initial treatment aim of LSVT is to improve the phona
tory 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 [16, 17, 18].

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 quasi-automatic fashion [19, 20]. Loudness scaling is a task that humans engage in all their lives [21, 22]. 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 loudness 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 multi-level 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 [23], and the results of
a PET brain imaging study of individuals with IPD pre- and post-LSVT [24] suggest that training loud phonation may also promote the recruitment of the right insular cortex and the anterior cingulate cortex.

Taken together, this suggests 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 gestural activity ‘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 [25, 26].

3. THE EXPERIMENT

Our pilot dataset was obtained from two patients, both male, who underwent an 8-week LSVT protocol involving therapy for 1 hour 2 days a week over the period. We shall designate the our two pilot study subjects 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. Both 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 [27]. 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 an 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. P1 and P2 produced datasets of 275.3 sec (8259 frames) and 336.033 sec (1081 frames) in before-treatment narration video for P1 and P2 respectively as does the duration, inter-sentential pauses exist in normal speech, such pauses per unit level time alignment. We also perform an analysis for sentential pause times. We first transcribe the speech data manually and run the Entropics in after-treatment narration video for P1 and P2 respectively.

Table 2: PD 1 and 2 Speech Repairs

<table>
<thead>
<tr>
<th>Pre-TX</th>
<th>Post-TX</th>
<th>Change %</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.6782</td>
<td>0.9521</td>
</tr>
<tr>
<td>P2</td>
<td>0.4530</td>
<td>0.4881</td>
</tr>
</tbody>
</table>

4. SPEECH PAUSE ANALYSIS

We first transcribe the speech data manually and run the Entropics word aligner on it. The are imported to the Praat phonetics analysis tool [28] and adjusted to obtain an accurate word, and syllable level time alignment. We also perform an analysis for sentential boundaries, disfluencies, and speech repairs.

Hesitation and silence in speech have been applied to the study of the cognitive aspects of speech production [29]. We posit that the distribution of empty pauses intra- and inter-sentence units constitute metrics of the degree of hesitation and deliberation. While intra-sentential pauses exist in normal speech, such pauses per unit time or utterance should remain relatively constant for an individual discussing a particular topic. Increased intra-sentential pauses would indicate greater hesitation and disfluency in the speech. On the other hand, since PD suffers tend to compensate for speech deficits by rushing their utterances, we expect that better speech phrasing and deliberation will manifest themselves in increased intersentential pause times.

We define $EP = \sum_T EP$ to be the set of all pauses between words greater than some pause determination threshold (we use a value of 150ms). $|EP|$ is the number of such pauses, $T(EP)$ is the total duration of such EP. We define $EP \cap S$ to be the set of empty pauses within sentences, and $EP \cap SP$ to be the intersentential pauses.

From Table 1 we can see that the percent of time speaking (i.e. word utterance time / total narration time) increased by 14.20% and 1.30% for PD1 and PD2 respectively.

Table 1: PD 1 and 2 Speech Characteristics

<table>
<thead>
<tr>
<th>Pre-TX</th>
<th>Post-TX</th>
<th>Change %</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>55.70%</td>
<td>63.61%</td>
</tr>
<tr>
<td>P2</td>
<td>79.05%</td>
<td>80.08%</td>
</tr>
</tbody>
</table>

5. SPEECH REPAIRS

We marked disfluencies in the aligned speech, labeling speech repairs, and aborted sentences (utterances that do not terminate and are not repaired), and filled pauses (such as <um>). Repairs are marked as tules of reparandum (region of the mistake, editing words (words that indicate a repair will happen), and alterations (the repaired speech). An example of a disfluent sentence taken from our PD data is: “[third time] [uh] [second time] [excuse me] [second time] [he goes up the drainpipe . . . ]”

1: This is the reparandum; 2: This is an editing word; 3: This is the alternation for ‘l’ (i.e. the fixup), but this also becomes the reparandum for ‘l’; 4: These are the editing words for the 2nd repair, and 5: This is the final reparandum. Note that the ‘<um>’ in ‘l’ would have been marked as a filled pause if it were not judged to perform the role of an editing word in the utterance.
We measured the number of repairs pre-TX and post-TX and their durations. As can be seen in Table 2, both the total number of repairs and time taken in repairs improved post-TX.

6. GESTURE ANALYSIS

We apply a fuzzy image processing approach known as Vector Coherence Mapping (VCM) [30] to track the hand motion. VCM applies spatial coherence, momentum (temporal coherence), speed limit, and skin color constraints in the vector field computation by using a fuzzy-combination strategies, and produce good results for hand gesture tracking. We apply an iterative clustering algorithm that minimizes spatial and temporal vector variance to extract the moving hands. The positions of the hands in the stereo images are used to produce 3D motion traces describing the gestures. To detect holds, we employed a motion energy-based detector to locate places where there was low motion energy [31]. We have observed that there is a high degree of cohesion between holds and pauses [32].

We divided holds into three kinds: rest holds where both hands were at the rest position, pre-stroke and post-stroke holds and stationary points in a motion sequence which are holds with durations lower than 200 msec (pre- and post-stroke holds are typically holds that take place before or after a motion stroke, and stationary points are typical in oscillatory gestures where the hand stops momentarily at the stroke extrema), and long hold strokes (non-rest holds greater than 200 msec).

We posit that long non-rest holds have different meanings depending on whether they intersect with pauses or words. If they intersect with words, they are more likely to be ‘hold-strokes’ (information-laden parts of a gesture that includes a effortful hold). If they intersect with empty pauses within sentences they are likely to reinforce the concept of hesitation. Some silences occur within sentences when the hand is moving, this may serve a coordinating function between gesture and speech (as when one says “When you pause while hand moves up to prepare a throwing action”).

For rest holds, we posit that a predominance of these holds during speech indicates a decrease in the ideation imagery. We computed the intersection of these hold phenomena with the between-sentence pauses, within-sentence pauses, and sentence durations.

We define $\text{N RH} = \sum \text{N RH}$ to be the set of all non-rest holds greater than our 0.2 sec threshold, and $\text{R H} = \sum \text{R H}$ to be the set of all rest holds. $\mathcal{T}[\text{N RH}]$ and $\mathcal{T}[\text{R H}]$ are the total duration of such $\text{N RH}$ and $\text{R H}$ respectively. We define $\text{N RH} \cap \text{E Pw}$ to be the set of intersections of $\text{N RH}$ with empty pauses within sentences, $\text{R H} \cap \text{S}$ to be the set of intersections between rest holds and sentences, and $\text{R H} \cap \text{S P}$ to be the set of intersections between rest holds and inter-sentence pauses.

For PD1, $\mathcal{T}[\text{N RH}]$ as a fraction of total discourse time was 34.025 msec per sec pre-TX. This declines to 9.57 msec/sec (or 71.87%) post-TX. For PD2, this value was 21.43 msec/sec pre-TX, but it increased slightly to 24.43 msec/sec (3%) post-TX.

To obtain the relationship of rest holds within and between sentences, we took the quotient:

$$\frac{\mathcal{T}[\text{RH} \cap \text{S P}]}{\mathcal{T}[\text{R H}] \cap \text{S}}$$

For PD1, this quotient increased from 0.181 pre-TX to 0.337 post-TX (an increase of 86.19%). For PD2, this quotient increased similarly from 0.364 pre-TX to 0.770 post-TX (an increase of 111.54%). The strong decrease in $\mathcal{T}[\text{N RH}]$ as a fraction of total discourse time in PD1 matches our finding in improved fluency in the speech. The problem with PD2 is that there was a shift in the experimental conditions pre-TX and post-TX. The subjects were seated in a therapy chair with adjustable armrests (they could be engaged or not). In the pre-TX case, the arm rests were not engaged, and the subjects gestured with the rest position on their laps. Hence each excursion from rest required greater effort. In the post-TX case, the arm rests were erroneously engaged. PD1 still used his laps as the rest position while PD2 rested his elbows on the arm rests and clasped his hands in front of his mid-section. This resulted in a markedly different dynamic in his gesticulation. Since the gesture and rest spaces were very close, it was difficult to tell when he was at rest and when the mid-length holds were rest or non-rest holds. Furthermore, because of the reduced effort required, PD2 post-TX engaged in many more beat gestures directly from rest position. Hence, we have to conclude that his results were inconclusive for the non-rest hold comparison.

The consistency in pre-TX to post-TX change in our second rest hold measure indicates that there was more gesticular activity post-TX as opposed to before.

7. CONCLUSIONS

We have presented the results of our pilot work on speech and vision-based gesture metrics to access the cognitive change in PD patients before and after LSVT. We employed a speech pause and gesture hold analysis for this purpose. We stress that these results are preliminary, but they do suggest that the study of change in multimodal language characteristics in narrative discourse may be of utility in understanding the broad cognitive changes in individuals with PD.

The temporal relation between pauses and sentences was consistent across the two subjects in our study even though there was a large disparity between the speech rates and disease conditions of the two patients. The gestural analysis of rest and non-rest holds with respect to speech pause and sentences in PD1 agreed with our basic hypotheses. In PD2, the confounding situation of the arm rests contributed to mixed results.

The purpose of this pilot work was to show the promise of multimodal linguistic analysis to study cognition and language in PD patients. In our ongoing work, we are extending our study to more subjects and devising more robust metrics for such cognitive evaluation.

8. REFERENCES


