Automatic Landmark Analysis of Dysarthric Speech

Thomas M. DiCicco, M.S.

Harvard-MIT Division of Health Sciences & Technology–Program in Speech and Hearing Biosciences and Technology, Cambridge, Massachusetts

Rupal Patel, Ph.D.

Northeastern University–Department of Speech-Language Pathology & Audiology, Boston, Massachusetts

Harvard-MIT Division of Health Sciences & Technology–Program in Speech and Hearing Biosciences and Technology, Cambridge, Massachusetts

The present study sought to characterize dysarthric speech in terms of acoustic landmarks. Landmark analysis provides a means to relate acoustic events to underlying articulatory behavior thereby allowing for comparisons between highly intelligible speech and dysarthric speech along a set of distinct acoustic parameters. Automatic landmark detection algorithms were utilized to extract acoustic landmarks from recordings produced by nine speakers with dysarthria and one control. Findings indicated that speakers with dysarthria not only produced expected acoustic targets at lower rates than the control, they also inserted unexpected landmarks at higher rates. Thus the dysarthric speech stream not only contains noisy acoustic information but also additional acoustic cues that may serve to confuse listeners. Additionally, these data highlight the notion that intelligibility is more than merely the result of accurate production of acoustic-phonetic targets. Rather, intelligibility scores resulted from the cumulative effects of precise, imprecise, absent, and superfluous articulation. The present study suggests the utility of automatic landmark analysis in developing personalized dysarthria treatment by specifying the acoustic cues that a speaker produces accurately while also identifying cues that a speaker fails to produce or inserts unnecessarily. Implications of this work on designing semiautomatic diagnostic tools and computer-assisted interventions are discussed.

Dysarthria is a motor speech disorder characterized by weak, slow, and/or uncoordinated movements of the musculature involved in speech production (Duffy, 2005; Yorkston, Beukelman, Strand, & Bell, 1999). Severely dysarthric speech commonly appears unintelligible to unfamiliar listeners; however, those familiar with the speaker are often able to comprehend with high accuracy (Deller, Hsu, & Ferrier, 1991). This observation implies that the speaker is producing acoustic cues that, while seemingly unintelligible to the unfamiliar listener, are capable of conveying information. Deller et al. (1991) hypothesized that the dysarthric speech stream not only contains noisy acoustic information but also additional acoustic cues that serve to confuse the listener. The current study sought to provide quantitative evidence for this hypothesis.

Stevens' Lexical Access from Features (LAFF) paradigm (Liu, 1995, 1996; Slifka, Stevens, Man-

Journal of Medical Speech-Language Pathology
Volume 16, Number 4, pp. 213–219

uel, & Shattuck-Hufnagel, 2004; Stevens, 1992, 2002; Stevens, Manuel, Shattuck-Hufnagel, & Liu, 1992) was applied to perform automatic landmark detection. Based on distinctive feature theory, Stevens' model provides explicit definitions for landmarks, the acoustic correlates of articulator-free features (Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1952). Articulator-free features, which are also referred to as manner features, do not depend upon the position of the speech articulators. Instead, they provide a description of vocal tract constriction by classifying a speech segment as a vowel, glide, or consonant (sonorant or obstruent). Landmark detection provides a metric for comparing highly intelligible speech to dysarthric speech along a set of empirically derived acoustic features, thus serving as a lens for identifying accurate as well as inserted acoustic cues. Characterizing the differences between healthy and dysarthric speech in terms of vocal tract constriction also has clinical utility in that it relates acoustic events to underlying classes of articulatory behaviors.

METHODS

Nemours Database

Recordings used in this study consisted of productions from the Nemours Database of Dysarthric Speech (Menéndez-Pidal, Polikoff, Peters, Leonzio, & Bunnell, 1996; Polikoff & Bunnell, 1999). This database contained recordings from 11 young males (in their twenties and thirties; exact ages were not documented) with dysarthria of varying severities secondary to either cerebral palsy ($N_{CP} = 7$) or head trauma ($N_{HT} = 4$), and a single control speaker. Prior to data collection, individuals with dysarthria were examined by a speech-language pathologist using the Frenchay Dysarthria Assessment (Enderby, 1983). While diagnostic classification of motor control was noted, the dysarthria subtype was not documented in the original database. For each speaker, the clinician assigned an intelligibility rating on a scale of 0–8 (where 8 corresponded to highest intelligibility). Mean sentence intelligibility of the speakers with dysarthria was reported to be 2.9 (SD = 2.4). Nine speakers with dysarthria ($N_{CP} = 6$; $N_{HT} = 3$) who had complete data sets were included in the analysis.

The database contained 74 nonsense utterances of the form "The *noun1* is *verb*-ing the *noun2*" produced by each speaker. The lexicon consisted of 74 monosyllabic nouns and 37 disyllabic verbs (counting the *ing*). The combinations of nouns and verbs were unique to each speaker with dysarthria (SWD). A single control speaker produced a corresponding set of utterances for each SWD. The database included time-aligned phonetic labels of the recordings produced by each SWD. Labeling was performed using a discrete Hidden Markov Model (HMM) labeler, followed by manual correction when necessary. The phonetic sequence was specified by the underlying text. Phonetic labeling of the utterances produced by the control speaker was performed using a similar procedure. Interlabeler reliability, in terms of time alignments, was not measured. Therefore, there may have been interlabeler variation in terms of time localization of phonetic boundaries but the phonetic sequences across dysarthric and control productions was the same.

Landmark Analysis

Speech sounds can be classified into one of three broad articulator-free classes: vowel, glide, or consonant (Stevens, 2002). The consonant class is further divided into sonorant and obstruent. Each broad class is associated with a set of corresponding acoustic correlates known as landmarks. Vowels and glides each have only one landmark. Vowel landmarks (V) are characterized by local maxima in the first formant (F_1) and waveform amplitude. Conversely, glide landmarks (G) are characterized by decreases in F_1 and waveform amplitude. For the consonantal class there are three landmark types: glottis (*g*), sonorant (*s*), and burst (*b*); and associated with each type is a sign (±). Glottis landmarks indicate a transition to (+*g*) or cessation (–*g*) of free vocal fold vibration. Sonorant landmarks occur during a voiced region in which there is a closure (+*s*) or release (–*s*) of a nasal or /l/. Burst landmarks denote the presence of a constriction resulting in acoustic discontinuity. A stop or affricate burst is denoted by +*b* landmarks, while –*b* landmarks signify cessation of frication or aspiration noise. In the present study, we extracted only vowel (Howitt, 2000a, 2000b) and consonantal landmarks (Liu, 1995, 1996). Glide landmarks were not examined because a reliable automatic glide landmark detector has not yet been implemented and validated.

A software implementation of the Lexical Access from Features (LAFF) paradigm (Howitt, 2000a, 2000b; Liu, 1995, 1996; Stevens, 2002) was used to automatically extract landmarks from recordings of the nine speakers with dysarthria and the

corresponding utterances produced by the control speaker. For all recordings the detection process was the same: consonantal landmark detection was performed first, followed by vowel landmark detection.

Consonantal landmark extraction followed a multistep process. The speech waveform of an utterance was preemphasized by 3 dB, and a broadband (6 ms Hanning window) spectrogram was taken every 1 ms (Liu, 1995). This short window provided good temporal resolution, and the high frame rate allowed for accurate tracking of rapid acoustic changes. Next, the spectrogram was divided into six frequency bands (Band 1: 0–0.4 kHz, Band 2: 0.8–1.5 kHz, Band 3: 1.2–2 kHz, Band 4: 2–3.5 kHz, Band 5: 3.5–5 kHz, and Band 6: 5–8 kHz) and the energy waveform in each band was constructed. The temporal derivative of the energy waveforms was computed and peaks in these derivate waveforms were extracted using a peak detection algorithm (Mermelstein, 1975). A $+g$ was associated with an abrupt (6 dB or more) increase in Band 1 energy, while a $-g$ corresponded to an abrupt decrease in Band 1 energy (i.e., frequencies below 400 Hz). Within a voiced region (i.e., between a $+g$ $-g$ pair), increases or decreases, on the order of ±9 dB, in Bands 2–5 were labeled as sonorant landmarks. In voiceless regions, increases or decreases, on the order of ± 9 dB, in Bands 2–6 were labeled as burst landmarks.

Automatic vowel landmark detection involved monitoring only one energy band, spanning 0–650 Hz (Howitt, 2000). In order to extract vowel landmarks, syllable boundaries were first localized. Syllable boundaries were identified as local minima in the energy waveform according to intensity (more than 2 dB difference between syllable boundary and peak) and durational (at least 80 ms between syllable boundaries) constraints (Mermelstein, 1975). A maximum within a pair of syllable boundaries was labeled as a vowel landmark if the peak was less than 25 dB below the utterance's maximum energy within the 0–650 Hz band.

Hypothetical Landmark Sequences

A phoneme-to-landmark mapping algorithm was used to automatically define expected sequences of landmarks. For each utterance, the mapping algorithm utilized the time-aligned phonetic transcriptions to hypothesize the corresponding sequence and timing of acoustic landmarks. The hypothetical landmark sequences allowed for the calculation of the detection, deletion, substitution, and insertion rates. If an observed landmark was the same type (and if applicable sign) and within 30 ms of a hypothesized landmark, then the hypothesized landmark was judged a detection. Previous work has shown that a 30 ms analysis window is sufficiently long for allowing a majority of automatically extracted landmarks to be associated with corresponding hand-labeled landmarks while being succinct enough that observed landmarks were not paired with hand-labeled landmarks from neighboring acoustic-phonetic events (Liu, 1995). A hypothesized landmark for which no landmark of the same type (and if applicable sign) was extracted within the acceptable analysis window was marked a deletion. If an observed landmark was the same sign but different type and within the analysis window of a hypothesized landmark, then the hypothesized landmark was judged a substitution. Note that substitutions are only applicable for consonantal landmarks. Finally, if a landmark was extracted from the waveform but did not correspond to a hypothesized landmark then it was marked an insertion.

RESULTS

Landmark detection, substitution, deletion, and insertion rates for the group with dysarthria and the control speaker are provided in Figure 1. Detection and error rates were calculated by dividing event counts by total number of hypothesized landmarks.[1,2] Speakers with dysarthria (SWDs) produced the expected acoustic targets 57% of the time, nearly one-third less often than the control (80%). SWDs also produced more than double the number of substitutions (15% vs. to 7%) and deletions (29% vs. 13%) and, most notably, inserted more than six times as many unexpected landmarks (113% vs. 19%) compared to the control. Analyzing the results by landmark type (right side of Figure 1) revealed similar trends. For all landmark types, the detection rate was lower, while the substitution, deletion, and insertion rates were higher for

[1]*Detection rate = 100% – Deletion rate – Substitution rate*

[2]Insertion rate was not bounded at 100% because it was possible to extract more insertions than number of hypothesized landmarks.

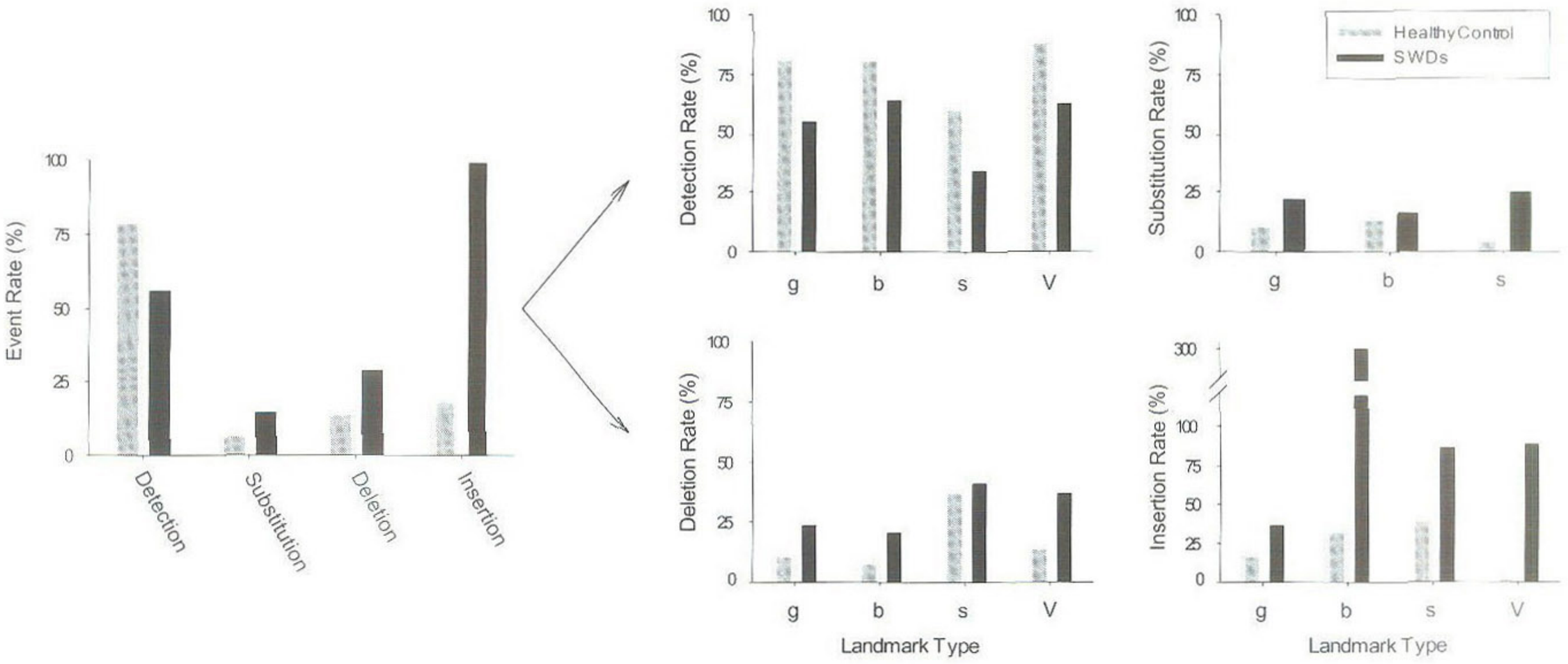

Figure 1. Landmark detection, substitution, deletion, and insertion rates for the control and the speakers with dysarthria. A breakdown of error rates by landmark type is provided on the right side.

the SWDs compared to the control. Burst (b) landmarks were deleted at the highest rate (42%) while sonorant (s) landmarks, corresponding to the closure or release of a nasal or /l/, were most frequently inserted (301%) by SWDs.

Given the heterogeneity of dysarthria etiology and severity within the Nemours Database, we also investigated the relationship between landmark rates and intelligibility. Since the database included a sentence production task, we utilized sentence intelligibilities from the Frenchay Dysarthria Assessment to establish this relationship. Scatter and regression plots of the landmark rates as functions of speaker intelligibility are shown in Figure 2. For regression analysis, the control speaker's intelligibility was assumed to be 8. Detection rate showed a positive correlation with intelligibility while substitution, deletion, and insertion rates all showed a negative correlation. The coefficients of determination, r^2, were similar for all landmark rates, ranging between 0.54 and 0.58. For all landmark rates the correlation with intelligibility was found to be statistically significant ($p < 0.05$).[3]

There were large cross-correlations between landmark event rates (Table 1, left side). The individual cross-correlations between detection rate and each of the error rates (deletion, substitution, and insertion rates) were all negative and ranged between –0.99 and –0.81. Thus, in general as detection rates decreased, error rates increased. The cross-correlations among the deletion, substitution, and insertion rates were all positive and ranged between 0.77 and 0.90, implying that the direction of change in these error rates was the same. All cross-correlations were significant ($p < 0.01$). To better account for the relationship between landmark rates and intelligibility, multiple linear regressions using all possible subsets of landmark rates were performed (Table 1, right side). When all four event rates served as dependent variables the coefficient of determination, r^2, was 0.63. A majority of this variation (98%) can be explained by the combination of detection and insertion rates alone.

DISCUSSION

Landmark analysis provided a means to relate acoustic-phonetic events to underlying articulatory behavior. A summary of the relationships between

[3]Due to the small sample size, correlations between intelligibility and substitutions as well as intelligibility and insertions were not statistically significant ($p = 0.08$ and 0.07, respectively) after the data from the control speaker was removed. However, correlations between intelligibility and detection and deletion rates remained significant ($p < 0.05$).

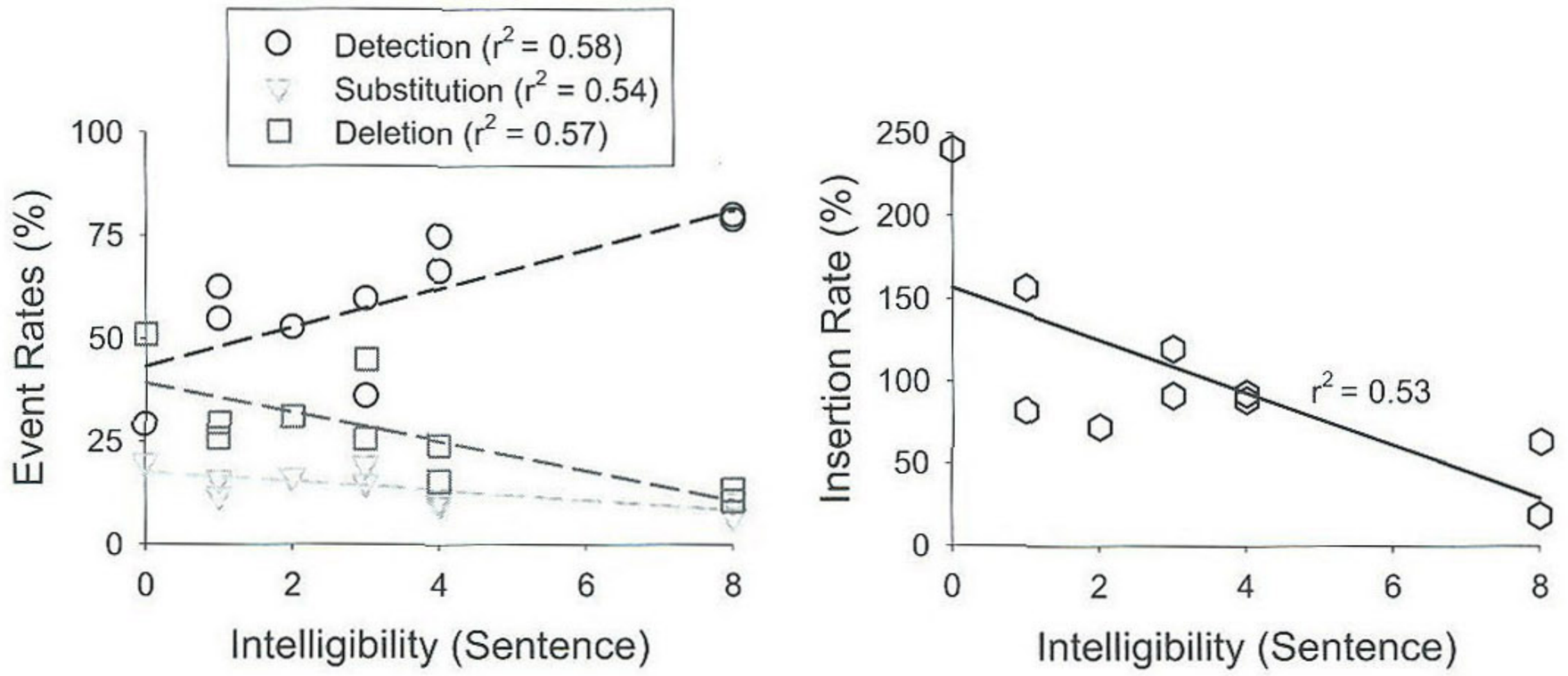

Figure 2. Landmark detection, substitution, deletion, and insertion rates as functions of speaker intelligibility. Linear regressions and coefficients of determination, r^2s, are included. Insertion rate is shown separately because it was not bounded at 100%.

TABLE 1. Cross-correlations ($\sigma_{i,j}$) between landmark event rates (left) and the r^2 and adjusted r^2 values from multiple linear regressions using intelligibility as the independent variable and the specified landmark *event rates* as the sets of independent variables (right).

Cross-Correlations		**Multiple Linear Regressions**		
Landmark Event Rates	$\sigma_{i,j}$	**Landmark Event Rates**	r^2	r^2_{adj}
Detection-Substitution	–0.94	Detection, Substitution, Deletion, Insertion	0.63	0.45
Detection-Deletion	–0.99	Substitution, Deletion, Insertion	0.62	0.51
Detection-Insertion	–0.81	Detection, Substitution, Insertion	0.62	0.51
Substitution-Deletion	0.90	Detection, Deletion, Insertion	0.62	0.51
Substitution-Insertion	0.77	Detection, Insertion	0.62	0.57
Deletion-Insertion	0.80	Deletion, Insertion	0.61	0.57
		Detection, Substitution, Deletion	0.61	0.50
		Substitution, Insertion	0.61	0.56
		Substitution, Deletion	0.59	0.54
		Detection, Substitution	0.58	0.53
		Detection, Deletion	0.58	0.53
		Detection	0.58	0.58
		Deletion	0.57	0.57
		Substitution	0.54	0.54
		Insertion	0.53	0.53

landmark events and underlying articulatory behavior is shown in Table 2. Accurately detected landmarks signified production of expected acoustic targets, while acoustic-phonetic events that occurred at the expected time but were imprecisely articulated were considered substitutions. Deletions were failures to signal acoustic-phonetic events and insertions indicated excessive or poorly timed articulation.

TABLE 2. Relationship between landmark events and underlying articulatory behavior.

Landmark Type	Deletion	Insertion
+g	Hypoadduction	Hyperadduction
–g	Hypoabduction	Hyperabduction
+s	Either reduced velopharyngeal opening or failure to *create* an oral constriction associated with a nasal or a liquid	Either hypernasality or unintended nasal or liquid production
–s	Either inability to close the velopharyngeal port or failure to *release* an oral constriction associated with a nasal or a liquid	Either nasal emission or excessive oral constriction
+b	Failure to *initiate* frication or aspiration noise production	Unexpected turbulent noise production
–b	Failure to *cease* frication or aspiration noise production	Unexpected cessation of noise production
V	Failure to achieve an open configuration of the vocal tract during voicing	Either a failure to produce a consonantal constriction or unexpected open vocal tract configuration

Findings from the present study provide quantitative support for Deller et al.'s (1991) hypothesis that dysarthric speech contains not only malformed or missing acoustic cues but also erroneously inserted cues that may mislead or confuse listeners. SWDs produced expected acoustic targets at lower rates than the control speaker and also inserted more unexpected acoustic targets (see Figure 1). The lower the intelligibility of the speaker, the larger the magnitude of this phenomenon, as evidenced by the fact that the clinician-based intelligibility ratings and each of the landmark rates were significantly correlated (see Figure 2). Cross-correlations between landmark rates and results from the linear and multiple linear regressions suggest that intelligibility is more than merely the result of accurate production of acoustic-phonetic targets (see Table 1). Decreased sentence intelligibility resulted from the cumulative effects of imprecise (substitution), absent/failed (deletion), and superfluous articulation (insertion).

Linear regressions between landmark rates and intelligibility only accounted for 53 to 63% of variation. An additional source of variability may be associated with production of acoustic cues associated with articulator-bound or place features. Articulator-bound features describe the state of the lips, tongue blade, tongue body, soft palate, pharynx, glottis, and/or vocal folds. For example, a speaker with a detection rate located above the regression line (see Figure 2) may have been producing the broad manner classes of articulator-free features but was failing to achieve accurate articulatory placements associated with articulator-bound features. The failure to convey articulator-bound features may then have contributed to a decrease in overall perceptual intelligibility. While analyses based on articulator-bound features would, for instance, associate consistently substituted alveolar stops for velar stops with poor intelligibility, articulator-free analyses are not sensitive to this place of articulation error.

In summary, landmark analysis specifies the set of acoustic cues that a speaker produces accurately while also highlighting the cues that a speaker fails to produce or inserts unnecessarily. We envision using this information to drive personalized dysarthria treatment or in the development of computer-assisted interventions that account for speaker- or disorder-specific acoustic patterns. Further research aimed at characterizing dysarthric speech in terms of the articulatory-to-acoustic mapping is warranted. We are currently performing landmark analysis on a larger dataset of dysarthric speech (data from Patel, 2004, and Patel & Campellone, in review). This database contains a broader range of dysarthria severity than the current analysis in which eight of the nine speakers with dysarthria had sentence intelligibilities of four or lower. Also, in addition to relating landmark detection and error rates to standardized intelligibility ratings, we plan to ex-

amine the relationship between landmark event rates and intelligibility scores ascertained directly from the materials being analyzed. Finally, future efforts will concentrate on extracting articulator-bound features in order to account for additional sources of variation, and to provide a more detailed characterization of the dysarthric speech stream in terms of articulatory behaviors. The Lexical Access from Features paradigm currently does not support articulator-bound feature detection, and there is poor agreement among linguists on the acoustic parameters that robustly define these features. Thus, we plan to use more statistically driven methods that rely on machine classification algorithms to extract articulator-bound and articulator-free features (Juneja, 2004; Juneja & Espy-Wilson, 2003).

Acknowledgments This work was funded in part by NIH/NIDCD grant #T32DC000038.

Address correspondence to Rupal Patel, Ph.D., Department of Speech-Language Pathology and Audiology, Northeastern University, 360 Huntington Ave, 102 Forsyth Building, Boston, MA 02115 USA.
e-mail: r.patel@neu.edu

REFERENCES

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Deller, J. R., Hsu, D., & Ferrier, L. J. (1991). On the use of hidden Markov modeling for recognition of dysarthric speech. *Computer Methods and Programs in Biomedicine, 35*(2), 125–139.

Duffy, J. R. (2005). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis: Mosby.

Enderby, P. M. (1983). *Frenchay Dysarthria Assessment*. San Diego: College-Hill Press.

Howitt, A. W. (2000a). Author: provide full reference information.

Howitt, A. W. (2000b). Vowel landmark detection. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000), 4*, 628–631, Beijing, China.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis. The distinctive features and their correlates* (No. 13). Acoustics Laboratory, Massachusetts Institute of Technology, Technical Report.

Juneja, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks*. Unpublished doctoral dissertation, University of Maryland, College Park.

Juneja, A., & Espy-Wilson, C. (2003). Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. *Proceedings of the International Joint Conference on Neural Networks,* Portland, Oregon.

Liu, S. A. (1995). *Landmark detection for distinctive feature-based speech recognition*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Liu, S. A. (1996). Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America, 100*(5), 14.

Menéndez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996). The Nemours Database of Dysarthric Speech. *Proceedings of the Fourth International Conference on Spoken Language Processing*, 1962–1965. Pennsylvania, USA.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America, 58*(4), 880–883.

Patel, R. (2004). The acoustics of contrastive prosody in adults with cerebral palsy. *Journal of Medical Speech-Language Pathology, 12*(4), 189–194.

Patel, R.,& Campellone, X. (in review). Author: provide title, etc.

Polikoff, J. B., & Bunnell, H. T. (1999). The Nemours Database of Dysarthric Speech: A perceptual analysis. *Proceedings of the XIVth International Congress of Phonetic Sciences, 1*, 783–786. San Francisco, California.

Slifka, J., Stevens, K. N., Manuel, S. Y., & Shattuck-Hufnagel, S. (2004). A landmark-based model of speech perception: History and recent developments. *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, C85–C90. Massachusetts Institute of Technology, Cambridge.

Stevens, K. N. (1992). Lexical access from features. *Speech Communication Group Working Papers, Volume VIII*, 119–144. Research Laboratory of Electronics, Massachusetts Institute of Technology. .

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*(4), 1872–1891.

Stevens, K. N., Manuel, S. Y., Shattuck-Hufnagel, S., & Liu, S. (1992). Implementation of a model for lexical access based on features. *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1992), 1*, 499–502. Banff, Alberta, Canada.

Yorkston, K. M., Beukelman, D. R., Strand, E. A., & Bell, K. R. (1999). *Management of motor speech disorders in children and adults*. Austin, TX: Pro-Ed.