Machine Classification of Prosodic Control in Dysarthria

Thomas M. DiCicco, Ph.D

Harvard-MIT Division of Health Sciences & Technology—Program in Speech and Hearing Bioscience and Technology, Cambridge

Rupal Patel, Ph.D

Department of Speech-Language Pathology and Audiology, Northeastern University, Boston

Harvard-MIT Division of Health Sciences & Technology—Program in Speech and Hearing Bioscience and Technology, Cambridge

Keywords: machine classification; prosody; dysarthria; voice-driven AAC aids

Recent studies suggest that speakers with dysarthria may be able to manipulate prosodic features sufficiently to convey information. Leveraging prosodic cues as an alternative or augmentative communication (AAC) signal may allow some individuals with dysarthria to use their vocalizations to engage in richer and more efficient interactions. As an initial step towards building voice-driven communication aids, the performance of three machine classification algorithms was compared to determine which algorithm(s) was most accurate and efficient for classifying a dataset of prosodic manipulations. Our findings suggest that machine classification of dysarthric productions is feasible using preexisting machine learners and rather minimal training data. Highly accurate classification of categorical duration control was achieved for all speakers with dysarthria; however, classification of pitch categories and simultaneous duration-pitch control varied widely across speakers. These findings have implications for harnessing the residual vocal abilities of individuals with dysarthria for machine-mediated AAC interactions.

MACHINE CLASSIFICATION OF PROSODIC CONTROL IN DYSARTHRIA

Individuals with severe motor speech impairment are often unable to rely on speech as their primary communication modality. In such instances alternative and augmentative communication (AAC) aids are used to supplement communication. Current state-of-the-art AAC devices rely on pointing and/or scanning input to select icons or words and phrases from graphical interfaces in order to construct messages which are displayed on a screen and/or produced by a speech synthesizer. AAC aids enable communication that would otherwise be limited; however, they fail to afford users the natural, complex, and efficient exchange that speech-based communication provides (Beukelman & Mirenda, 1998; Mathy-Laikko, West, & Jones, 1993;

Disclosure: Preliminary results of this work were presented as a poster at the 2007 American Speech-Language-Hearing Association (ASHA) annual convention.

Journal of Medical Speech-Language Pathology
Volume 18, Number 4, pp. 35–39

Shein, Brownlow, Treviranus, & Parnes, 1990; Vanderheiden, 1985). Utilizing residual vocal control as an input modality has the potential to improve the quality and efficiency of communication.

In this study, we focused on building and evaluating machine classifiers capable of identifying distinct prosodic categories produced by speakers with dysarthria. While speakers with severe dysarthria may exhibit inadequate articulatory control for producing speech sound segments, recent studies suggest many individuals have preserved ability to manipulate prosodic features as a means of conveying information (Ciocca, Whitehill, & Yin Joan, 2004; Le Dorze, Ouellet, & Ryalls, 1994; Patel, 2002, 2003, 2004; Patel & Campellone, 2009; Patel & Salata, 2006; Patel & Schroeder, 2007; Vance, 1994). Controlling duration and pitch requires less complex motor control with slower changes in the speech musculature occurring over a wider temporal frame (Patel, 2002, 2004). If speakers with dysarthria can consistently and precisely control prosodic features, these cues could be leveraged as an AAC signal serving as a rapid and more natural means to access and maneuver through options and menus on AAC aids (Patel, 2002; Patel & Roy, 1998).

In order to use prosodic vocalizations to control an AAC aid, accurate and efficient methods to categorize prosodic manipulations are needed. In this study we chose three popular machine learning methods: *k*-nearest neighbor (*k*-NN), support vector machines (SVMs), and a supervised clustering (SC) scheme, in order to determine the algorithm(s) with the highest accuracy. These algorithms were chosen because they span a range of complexity and computational/data demands. *k*-NN assigns an unseen data point the label associated with a majority of the *k* closest known (training) points (Cover & Hart, 1967). *k*-NN is simple to implement and does not require retraining with the addition of new training examples. SVMs map a data set to a higher non-linear dimension with the goal of discovering a separator within that space (Burges, 1998). SVMs have proven extremely powerful for a variety of classification tasks but training can prove computationally demanding and require retraining with additional training exemplars. Finally, SC is a clustering technique that allows for novel visualization of data and lies in between *k*-NN and SVM in terms of computational complexity (Eick & Zeidat, 2005). Traditional unsupervised clustering involves grouping data so that the distance between clusters is minimized. Alternatively, SC maximizes the number of cluster members with the same class label.

METHOD

Prosodic Control Dataset

The present study utilized speech recordings from a database of 5 children with severe dysarthria (2 males and 3 females; ages 6–13 years old, mean age = 9 years, 7 months) secondary to cerebral palsy and 5 gender-matched healthy children (ages 7–8 years old, mean age = 7 years, 7 months), that were collected in a previous study (Patel & Salata, 2006). The children with dysarthria were severely impaired or essentially non-verbal and most relied on a combination of communication modalities including two speakers who used AAC devices. The healthy controls were included in the analysis to provide a baseline measure for comparison.

Spoken utterances in the database consisted of three different experimental tasks, each involving sustained production of the vowel /a/. The duration control task required speakers to produce a short, medium, or long duration /a/ while trying to maintain a constant pitch. In the pitch control task, speakers produced the vowel at a low, medium, or high pitch while maintaining a constant duration. Lastly, in the simultaneous control task, speakers manipulated the duration and pitch of the vowel simultaneously (for a total of 9 possible distinctions). In each task, prosodic categories were requested using a computer game interface in which animated characters were used to elicit a given distinction. For each speaker, 15 repetitions of each category from each of the duration and pitch protocols and 8 repetitions of each distinction from the simultaneous control protocol were utilized. The Praat speech analysis software package (Boersma & Weenink, 2007) was used to extract the appropriate acoustic features (duration and/or average fundamental frequency (F_0)) from the recordings.

Machine Learners

This study sought to build computationally efficient machine learners capable of recognizing prosodic manipulations as an initial step towards developing assistive communication technologies that utilize prosodic control as an input signal.

The duration and/or average F_0 values extracted from the recordings of the Patel-Salata Database were classified using three different machine learning algorithms: *k*-nearest neighbor (*k*-NN), support vector machines (SVMs), and a supervised clustering (SC) scheme. Given large individual differences in vocal abilities among the speakers with dysarthria, machine classifiers were built for each individual speaker. Recognition error for each classifier was estimated using leave-one-out cross-validation. For each speaker, the average test error of each classifier type (*k*-NN, SVM, SC) was calculated for each feature set (duration, pitch, duration and pitch).

k-nearest neighbor. The *k*-NN rule assigns an unseen data item a class label based on the "majority vote" of the *k* closest (according to a given distance metric) prototypes, i.e. training examples whose labels are known (Duda, Hart, & Stork, 2001). In this case, the squared Euclidean distance was used as the distance metric. The optimal value for *k* was found using an additional round of cross-validation.

Support vector machines. In SVM analysis, input data are transformed into a higher dimensional space via a non-linear mapping function. An SVM then finds the separating linear hyperplane with maximal margin in this new dimensionality (Burges, 1998).

Supervised clustering. Clustering is a knowledge discovery technique that attempts to group similar objects based on a given metric. Whereas unsupervised clustering ignores classifications of data items, SC focuses on maximizing cluster purity (i.e. cluster members with the same class label) (Eick & Zeidat, 2005). SC aids in building a classifier that can assign an unseen, unclassified data point the label of the cluster that best shares its attributes. This classification scheme is a 1-nearest neighbor classifier where the cluster representatives serve as the known prototypes.

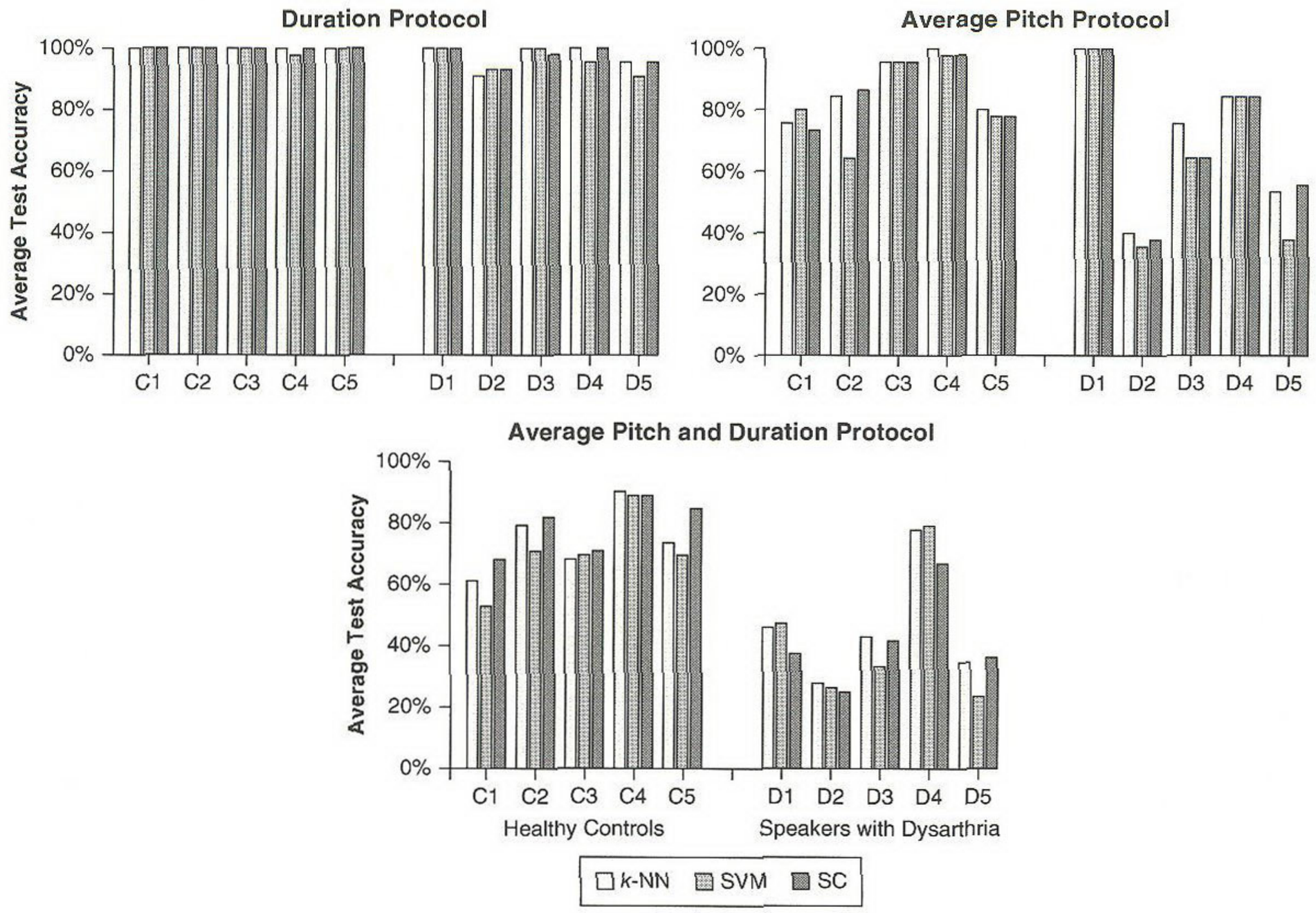

Figure 1. Per-Speaker, Average Test Accuracies of Each Optimized Classifier for the Duration (Upper Left-Hand Panel), Average Pitch (Upper Right-Hand Panel), and Simultaneous (Bottom Panel) Protocols

RESULTS

The performance of three classification algorithms was compared to determine which was most accurate and efficient for classifying prosodic categories produced by speakers with dysarthria. Average test accuracies for each of the three classifier types per speaker and feature set are shown in Figure 1. For the duration protocol (Figure 1, upper left-hand panel), the best performing classifier (*BPC,* defined as the classifier type that achieved maximum test accuracy for each speaker-feature set pair) was always 100% accurate for the HC group and ranged from 93% to 100% accurate for the DYS group. For all five DYS speakers the test accuracies were within a range of 5 percentage points for all three classifier types. For the pitch protocol (Figure 1, upper right-hand panel), BPC accuracy ranged from 80% to 100% for HC and 40% to 100% for DYS speakers. For speakers D1, D2, and D4 all classifier types produced similar (within 5 percentage points) average test accuracies. For speakers D3 and D5, depending upon classifier type, the average test accuracies ranged 10 and 18 percentage points, respectively. For four of five DYS speakers, the *k*-NN classifier was the BPC. For the simultaneous control task (Figure 1, bottom panel), BPC accuracy ranged from 68% to 90% for HC and 36% to 79% for DYS. The BPC accuracy was greater than 50% for only one speaker with dysarthria (D4). Compared to the previous two tasks, there was more variability between classifier types with no classifier type serving as the BPC for more than two speakers with dysarthria.

DISCUSSION

Results indicate that machine classification of DYS duration manipulations may be highly reliable and thus could be explored as a viable input modality for navigating through AAC interfaces. For all DYS speakers, all classifier types performed similarly well with limited variability in test accuracy. Classification accuracies of pitch adjustments were typically much lower than recognition rates of duration-controlled vocalizations, suggesting that further research into the feasibility of using pitch manipulation as an AAC input is required. Degraded classification was likely due to reduced separation between pitch production categories (Patel & Salata, 2006). While all DYS speakers were able to produce three non-overlapping (as observed using the mean and standard deviation) duration categories, only three of the five speakers (D1, D2, and D4) were able to produce three non-overlapping F_0 (pitch) categories. The remaining two speakers, D2 and D5, were able to produce only two categories. It should be noted that the two speakers with the highest recognition rates (D1 and D4) each produced three categories. For four of the five speakers with dysarthria the *k*-NN classifier was the BPC. For speaker D5, the accuracy of the *k*-NN classifier was within 2 percentage points of the BPC. These findings suggest that for average F_0 classification, *k*-NN may be the best classifier type to implement in a voice driven AAC device.

While the simultaneous control paradigm was the most difficult production task for both groups, the difference in performance between HC and DYS speakers was also largest. Additional data are required to determine whether paired adjustment of pitch and duration is too motorically complex, or whether it can be trained over time. It is encouraging, however, that for all DYS speakers BPC accuracies were at least 2.5 times higher than chance performance despite a highly restricted number of training tokens (15 for duration and pitch tasks, 8 for simultaneous control). Utilizing additional training tokens and/or reducing the number of prosodic categories could prove beneficial in increasing recognition accuracy in future research efforts.

CLINICAL IMPLICATIONS

The present findings suggest that machine classification of prosodic manipulations in dysarthric speech is feasible using preexisting algorithms that require minimal training data. Given the high classification accuracies, duration appears to be the most reliable prosodic cue to incorporate into voice-driven AAC interfaces. Furthermore, some speakers may be able to utilize both pitch and duration cues as input modalities. Although simultaneous control of pitch and duration may require additional speaker training, these initial findings are encouraging.

This study serves as an initial step towards developing assistive communication technologies that utilize prosody as an input signal. We envision utilizing prosodic control as a quasi voice-driven mouse for AAC interfaces. Such a control

strategy has the potential to reduce fatigue in accessing menu items and to accelerate communication rate.

Acknowledgments This research was supported in part by funding from the National Institutes of Health (grant #T32 DC000038).

Address Correspondence to *Rupal Patel, Department of Speech-Language Pathology and Audiology, Northeastern University, 360, Huntington Ave, 102 Forsyth Building, Boston, MA 02115, PHONE: (617) 373-5842, FAX: (617) 373-2249*
e-mail: r.patel@neu.edu

REFERENCES

Beukelman, D. R., & Mirenda, P. (1998). *Augmentative and alternative communication: Management of severe communication disorders in children and adults* (2nd ed.). Baltimore: P.H. Brookes Publications.

Boersma, P. & Weenink, D. (2007). Praat: A system for doing phonetics by computer (Version 5.0.20). [Computer software]. Amsterdam: Institute of Phonetic Sciences. Available at www.praat.org.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 47.

Ciocca, V., Whitehill, T. L., & Yin Joan, M. K. (2004). The impact of cerebral palsy on the intelligibility of pitch-based linguistic contrasts. *Journal of Physiological Anthropology and Applied Human Science, 23*(6), 283–287.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on Information Theory, 13*(1), 21.

Duda, R. O., Hart, P. E., & Stork, D. (2001). *Pattern classification,* 2nd edition. New York: Wiley.

Eick, C. F., & Zeidat, N. (2005). Using supervised clustering to enhance classifiers. *Lecture Notes in Computer Science*, (Issue 3488), 248–256.

Le Dorze, G., Ouellet, L., & Ryalls, J. (1994). Intonation and speech rate in dysarthric speech. *Journal of Communication Disorders, 27*(1), 1–18.

Mathy-Laikko, P., West, C., & Jones, R. (1993). Development and assessment of a rate acceleration keyboard for direct-selection augmentative and alternative communication users. *Technology & Disability, 2*, 57–67.

Patel, R. (2002). Phonatory control in adults with cerebral palsy and severe dysarthria. *Alternative & Augmentative Communication, 18*, 2–10.

Patel, R. (2003). Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy, *Journal of Speech, Language, & Hearing Research, 46*(6), 1401–1415.

Patel, R. (2004). The acoustics of contrastive prosody in adults with cerebral palsy. *Journal of Medical Speech-Language Pathology, 12*(4), 189–193.

Patel, R., & Campellone, P. (2009). Acoustic and perceptual cues to contrastive stress in dysarthria. *Journal of Speech, Language, and Hearing Research, 52*, 206–222.

Patel, R., & Roy, D. (1998). Teachable interfaces for individuals with dysarthric speech and severe physical impairments. *AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology*, Madison, WI., p. 40–47.

Patel, R., & Salata, A. (2006). Using computer games to mediate caregiver-child communication for children with severe dysarthria. *Journal of Medical Speech Language Pathology, 14*(4), 279–284.

Patel, R. & Schroeder, B. (2007). Influence of familiarity on identifying prosodic vocalizations produced by children with severe dysarthria. *Clinical Linguistics and Phonetics, 21*(10), 833–848.

Shein, F., Brownlow, N., Treviranus, J., & Parnes, P. (1990). Climbing out of the rot: The future of interface technology. In B. Mineo (Ed.), *Augmentative and alternative communication in the next decade* (pp. 36–39). Wilmington, DE: Applied Science and Engineering Laboratories.

Vance, J. E. (1994). Prosodic deviation in dysarthria: A case study. *European Journal of Disorders of Communication, 29*(1), 61–76.

Vanderheiden, P. J. (1985). Writing aids. In J. Webster, A. Cook, W. Tompkins & G. Vanderheiden (Eds.), *Electronic aids for rehabilitation* (pp. 262–282). London: Chapman & Hall.