Teachable Interfaces for Individuals with Dysarthric Speech and Severe Physical Disabilities

Rupal Patel[†] and Deb Roy[*]
[†]Department of Speech-Language Pathology, University of Toronto
[*]The Media Laboratory, Massachusetts Institute of Technolog
r.patel@utoronto.ca, dkroy@media.mit.edu

1 Abstract

Standard interfaces including keyboards, mice and speech recognizers pose a major obstacle for individuals with severe speech and physical disabilities. A person with insufficient control of their hands or voice will be unable to efficiently use such interfaces. We are investigating teachable interfaces, which can adapt to the preferences and abilities of an individual user. The user provides active feedback to the system to guide the adaptation process. This paper reports on preliminary work, which will form the basis for building a teachable interface for individuals with severely dysarthric speech. Ultimately, the system would be capable of translating "unintelligible" vocalizations into effective actions or clearly articulated synthesized speech.

2 Introduction

Individuals with severe speech impairments such as dysarthria[1] have a compromised channel of communication. Slow, imprecise and variable speech imposes an information bottleneck, impeding efficient and effective communication. Assistive interface technologies have attempted to improve information transfer for individuals who are non-speaking by increasing the number of input modes [Shein91]. An aided communication system capable of recognizing the vocalizations produced by individuals with severe dysarthria would capture an additional information channel available to the user. Although we are currently focusing specifically on dysarthria, the research methodologies discussed here may also be applicable to people with other communication impairments.

We are interested in developing human-machine interfaces for individuals with severe speech and physical disabilities, which leverage the power of machine learning and perceptual computing. Our strategy is to create a set of sensor driven feature detectors which are attuned to various aspects of a person's behaviour such as vocalizations and hand gestures (including keyboard presses and touch pad interaction). Machine learning techniques based on reinforcement from the user are used to select salient sensory features for specified human-machine communication tasks. The goal is to provide a sensor-rich interface, which over time learns to recognize salient human behaviours such as words and gestures and take appropriate actions. We assume the user is a willing participant in the adaptation process and actively provides feedback to the machine to guide its learning.

[1] Dysarthria is a neurogenic motor speech disorder. Slow, weak, imprecise and/or uncoordinated movements of the speech production musculature result in unclear speech output [Yorkston88].

In this paper we describe current research in our two laboratories with plans to merge our work in the near future. At the University of Toronto we are conducting clinical experiments with several subjects with severe dysarthria to determine their ability to control prosodic aspects of their vocalizations. This work is motivated by previous unsuccessful attempts to use commercial speech recognizers due to inherently high intraspeaker phonetic variability. Using a phonetic speech recognizer developed at the MIT Media Lab, we conducted pilot experiments with speech samples from subjects with severe dysarthria. On a 15-utterance recognition task based on phonetic matching, we found only chance performance. The potential for using prosodic aspects of dysarthric speech as an information-carrying channel has not been documented in the literature. Pilot experiments suggest that subjects may in fact be able to control the pitch, duration and intensity of their vocalizations to varying degrees. We report on these findings and further experiments that we are conducting to verify these results.

At the MIT Media Lab, we are developing a framework for multimodal adaptive interfaces [Roy98a]. The approach uses machine learning to find correlations between human behaviours and machine actions. Appropriate sensors to detect the user's actions will be selected based on the clinical experiments conducted at the University of Toronto. We emphasize that this paper describes work in progress. Although we have performed initial clinical experiments with the target disabled population and we have also implemented a prototype teachable interface, these research areas have not yet been combined.

This paper begins with a brief review of the field of augmentative and alternative communication (AAC) including previous work on using automatic speech recognition for individuals with dysarthria. In Section 4 we present results of our pilot clinical experiments aimed at determining vocalization parameters for individuals with dysarthria and methodologies for future experiments. Section 5 summarizes our work on a teachable interface prototype.

3 Background

3.1 Alternative and Augmentative Communication

Individuals with severe speech impairment such as severe dysarthria typically compensate by using alternative and augmentative communication[2] (AAC) aids including physical objects, picture symbols, alphabet boards, and sign language. AAC systems may include iconic symbol displays or adapted keyboards with alphanumeric symbols. Direct selection[3] and scanning[4] are among several methods to select the desired symbol or key on a communication device. A voice output communication aid (VOCA) may frequently be part of an individual's AAC system. The VOCA translates keypad/keyboard selections into digital synthesized speech.

[2] Alternative and augmentative communication is an "area of clinical practice which attempts to compensate (either temporarily or permanently) for the impairment and disability patterns of individuals with severe expressive communication disorders (i.e., the severely speech-language and writing impaired)" (ASHA, 1989, p. 107).

[3] Direct selection is an input method. An individual indicates a choice using methods such as finger pointing, eye gaze or use of an infrared pointer.

[4] Scanning is an additional input method. Individuals who are not able to use direct selection may use scanning systems to select the desired symbol from a scanning array. Scanning selection requires activation or deactivation of one or more switches in order to scan through the symbol array and select the desired symbol.

Although an AAC system that does not include speech may serve vocational and educational needs, it may not sufficiently satisfy the individual's social communication needs [Beukelman92]. Moreover, methods such as direct selection and scanning are slow and often fatiguing, especially in the presence of poor motor control [Ferrier92; Treviranus92]. Automatic speech recognition of distorted speech would offer an alternative interface to touch-typing, pointing or scanning for control of a VOCA or computer.

3.2 Severe Dysarthria Secondary to Cerebral Palsy

Cerebral palsy is a non-progressive developmental neuromotor disorder that results from an abnormality in the developing brain [Hardy83]. Postural and motor impairments vary according to the location and degree of the lesion. Primary motor impairments often manifest as severe speech and writing difficulties that require AAC intervention [Beukelman92 [5].

Speech intelligibility is a common measure of the severity of dysarthria. The Assessment of Intelligibility of Dysarthric Speaker [6] is a standardized assessment tool used to determine the level of severity of dysarthria[7]. Individuals would who score in the severe range on this battery are likely to achieve functional communication through AAC approaches.

Many individuals who have severe dysarthria and use AAC may have normal or exceptional intellects, reading skills and language skills and many would prefer to use their residual speech capabilities despite its limitations [Ferrier92; Fried-Oken85; Treviranus92]. Many individuals will exploit whatever residual speech available to them to express emotions, gain attention and signal danger [Beukelman92].

3.3 Automatic Speech Recognition for Dysarthric Individuals

Automatic speech recognition technology is intuitive, hands-free and encourages face-to-face interaction. It has the potential to be faster and less physically fatiguing than direct manual selection or scanning [Treviranus91].

Current commercially available automatic speech recognition products are designed for individuals whose speech is not impaired. Commercial systems may be able to recognize the speech of individuals with mild impairments and/or individuals who have received sufficient training to alter their articulatory patterns to achieve improved machine recognition rates [Carlson87; Ferrier92; Freid-Oken85; Kotler97; Schmitt86]. Severely dysarthric speech poses a challenge for commercial systems. Although moderate success has been achieved by rebuilding acoustic models for specific dysarthric populations [cf. Deller91; Jayaram95], considerable interspeaker variability for this population remains a significant problem. These findings have motivated us to develop an interface, which will adapt directly to the abilities of each individual user.

[5] Reported incidence rates of dysarthria range between 31-88% [Yorkston88].

[6] The Assessment of Intelligibility of Dysarthric Speakers. Yorkston, Beukelman, & Traynor (1984) C.C. Publications, Inc., P.O. Box 23699, Tigard, Oregon 97223-0108

[7] Patients are asked to produce a list of 50 words that will be heard by three naïve listeners. Intelligibility is the ratio of words understood by the listener to the total number of words articulated. Based on listener judgements of intelligibility, a severity level is determined.

4 Pilot Studies in Analyzing Prosody of Dysarthric Speech

4.1 The “Noisy”Acoustic Information Channel

Despite the severity of speech impairment, communication exchanges that take place between an AAC user and a familiar communication partner (FCP) are generally complex and rich in information. A marked reduction in communication efficiency may result, however, when the same AAC user interacts with an unfamiliar communication partner. Vocalizations may be distorted in phonemic (referred to as “segmental”)[8] clarity and/or prosodic (referred to as “suprasegmental”)[9] characteristics. The communicative message is buried within the acoustic “noise” and must be decoded by the communication partner. While FCPs may use information from multiple communication channels (e.g., facial expressions, body language, emotional state, situational contextual cues, and acoustic cues) to decode the AAC user’s intended message, the vocalization recognition system will have access to only the acoustic channel.

4.2 Need to Determine the Prosodic Characteristics of Severely Dysarthric Speech

Perceptual speech characteristics of various types of dysarthria have been documented thoroughly [cf. Darley69; Rosenbeck78]. Existing literature focuses on the degradation of the clarity, flexibility and precision of dysarthric speech. With advances in automatic speech recognition technologies, phonetic variability and acoustic distortion of the residual speech channel of individuals with severe dysarthria have been investigated [Deller91; Doyle95; Ferrier95; Jayaram95]. There is a paucity of literature, however, documenting the prosodic features of the residual speech channel of these individuals.

4.3 Utility of Documenting Control of Prosodic Parameters

Most commercially available speech recognition systems have explicit algorithms that factor out variations in pitch, intensity and duration [Rabiner93]. They assume that pitch, loudness and duration do not carry any linguistically salient information. Although highly variable phonetic characteristics typically impede accuracy rates for current automatic speech recognition systems, technologies capable of interpreting control of suprasegmental features such as pitch and loudness contour may enable individuals with severe dysarthria to control their VOCA using their vocalizations.

4.4 Pilot Study aimed at Determining Control of Prosodic Features

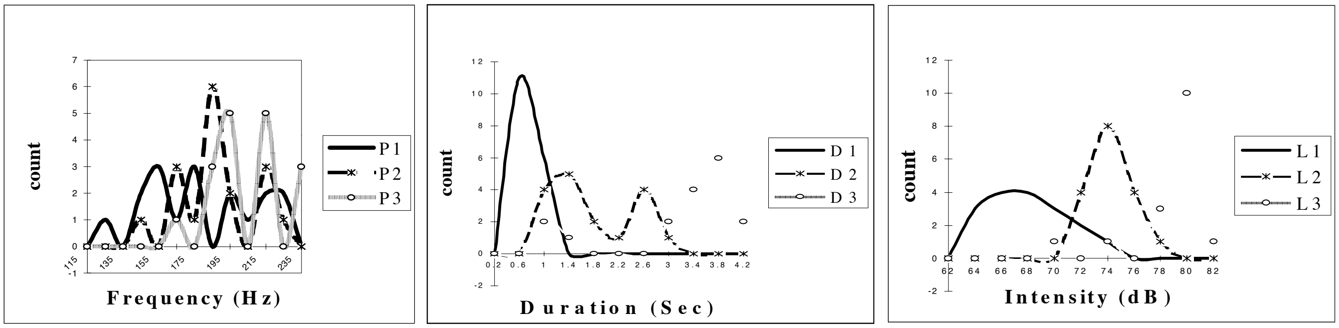
A pilot experiment aimed at determining prosodic control was undertaken in May 1997 at the University of Toronto. An adult male, age 52, with cerebral palsy and severe dysarthria participated in the study. There were three identical protocols for examining control of pitch, loudness and duration. In each protocol, the subject was instructed to sustain production of the vowel /a/ at three levels: low, medium and high. For each protocol, 51 vocalizations (i.e. 17

[8] Segmental features of a speech signal include transcription of the sample into the basic phoneme units [Shriberg95].

[9] Suprasegmental features of a speech signal include intonational markers, stress pattern, pitch variation, loudness variation, pausing and rhythm [Shriberg95].

vocalizations at each level) were requested[10]. The average level of frequency, intensity and duration were collected for each utterance and tagged to the level requested. The data were analyzed to determine the number of distinct non-overlapping categories that the speaker was able to produce. Descriptive statistic and informational analysis calculations were used to determine the number of levels of control within pitch, loudness and duration (figure 1). Results for this individual indicated little to no consistent control of frequency, greater control for duration and the most control for intensity. These findings provide the impetus to further investigate the role of prosody control in this population.

Figure 1. Distinct levels of pitch, loudness and duration



4.5 Identifying and Harnessing Salient Prosodic Parameters Using a Perceptual Model

The purpose of this investigation is to determine which parameters, if any, of the acoustic channel are salient to human listeners when decoding severely dysarthric speech. In light of highly variable and distorted phonemic features, it is hypothesized that prosodic features such as frequency and intensity contour may be more stable parameters of severely dysarthric speech. These parameters are temporally longer and therefore may offer greater vocal stability. Systematically removing specific prosodic features from the acoustic signal and measuring the corresponding effect on discrimination accuracy can be used to deduce the relative saliency of the features manipulated. It is also possible to determine the relative importance of frequency and intensity contour for assisting with discrimination judgements. The differentially facilitative role of prosodic features in making discrimination judgement among "unintelligible" vocalizations can also be determined for FCPs and lay listeners.

Modeling a vocalization recognition system based data gathered from human perceptual judgements is in accord with contemporary approaches to interface design. Ultimately, the communication aid of the future, capable of attending to the salient prosodic features of dysarthric speech, would act as a "familiar facilitator" between the AAC user and an unfamiliar communication partner. In such a system, vocalization of severely dysarthric speech would sere as an alternative access, an adaptive automatic vocalization recognition system would be the interface, and intelligible synthesized speech or an orthographic display, would be the output.

[10] A training period was allowed where the speaker practiced making "high", "medium" and "low" levels of frequency, intensity and duration. Visual cues were provided using the oscilloscopic display of the VisiPitch (Model 6087 DS, KayElemetrics Corporation, 12 Maple Ave, Pinebrook, NJ 07058).

5 Multimodal Adaptive Interfaces

5.1 Overview of Approach

At the MIT Media lab we are developing interfaces which combine multimodal sensing of the human with machine learning. Interfaces will be capable of adapting to the preferences and abilities of individual users. A collection of sensors and associated signal processors produce a set of sensory primitives. Input modalities might include haptic (touch), visual (gesture, facial expressions) and auditory (speech, vocalizations). In addition, some sensors are able to detect aspects of the environment that are relevant to the communication task. The system also has a channel to receive feedback from the user.

For the machine learning component of the system, the goal is to learn to recognize communicative primitives and domain-specific semantics from the user. Once trained, the interface can take action when it detects that the user has performed a salient behaviour. An example of a communication primitive might be a spoken word and the result might be to turn on a light switch or dial a phone number.

5.2 Sensing the User

We have developed feature detectors aimed at representing speech and hand gestures. For speech we have developed a speaker independent continuous phoneme recognition system [Roy98b]. To track hand gestures we are using a computer vision system similar to Azarbayejani (1996). The system uses to overhead cameras to locate the user's hands using skin color. The views from the two cameras are merged to estimate the 3-D position.

Application of this technology to individuals with severe speech impairment and physical disabilities will require modifications. As mentioned in Section 4, the current phoneme recognition system is not suitable for dysarthric speech. Appropriate prosodic feature detectors will be added to the system based on the results from clinical experiments that are in progress at the University of Toronto. Depending on the user's physical abilities, many standard input devices including touch pads, keyboards, mice, and joysticks may also be used.

5.3 A Prototype: Toco the Toucan

We have created a prototype adaptive interface, which is embodied as an animated synthetic character called Toco the Toucan. Toco uses the vision and speech systems described above to observe the user's actions. In a word learning application, Toco can acquire acoustic models of spoken words regardless of the speaker's language and accent. The current task is a simple block world in which colors and shapes of the blocks are internally encoded and provide a grounding for word semantics. The current interface relies on learning from examples rather than direct feedback from the user. In a typical interaction the user can point to virtual objects on a large screen and name them. Toco learns the words and infers their meanings. For more details see [Roy98a].

Toco could function as a graphical interface for an adaptive vocalization recognition system for individuals with severe dysarthria. Toco would recognize vocalizations and the produce clearly articulated synthesized speech. The speaker would provide Toco with feedback for learning through any one of various physical switches (e.g. footpedals, head-mounted switches, joysticks).

6 Future Directions

We have reported on two research projects that we plan to merge in the near future. Our goal is to build an adaptive interface for individuals with severe dysarthria for the purpose of enhancing communication efficiency. A set of salient input sensors will be identified based on results of vocalization control experiments and assessment of other physical control abilities of potential users. These feature detectors will then be added to the adaptive interface framework. Last, usability tests will be performed with subjects from the target population.

7 References

American Speech-language-Hearing Association. (1991). Report: Augmentative and alternative communication. Asha, 33 (Suppl. 5), 9-12.

Azarbayejani, A, Wren, C. & Pentland, A. (1996). Real-time 3-D Tracking of the Human Body. Proceedings of IMAGE'COM 96, Bordeaux, France, May 1996.

Beukelman, D. R. & Mirenda, P. (1992). Augmentative and alternative communication: management of severe communication disorders in children and adults. Baltimore: Paul H. Brookes.

Carlson, G.S. & Bernstein, J. (1987). Speech recognition of impaired speech. Proceedings of RESNA 10th Annual Conference, 103-105.

Darley, F. L., Aronson, A.E., & Brown, J.R. (1969). Differential diagnostic patterns of dysarthria. Journal of Speech and Hearing Research, 12, 246-269.

Deller, J. R., Hsu, D., & Ferrier, L. (1991). On hidden Markov modelling for recognition of dysarthric speech. Computer Methods and Programs in BioMedicine, 35(2), 125-139.

Doyle, P.C., Raade, A.S., St. Pierre, A. & Desai, S. (1995). Fundamental frequency and acoustic variability associated with production of sustained vowels by speakers with hypokinetic dysarthria. Journal of Medical Speech-Language Pathology, 3(1), 41-50.

Ferrier, L. J., Jarrell, N., Carpenter, T., & Shane, H., (1992). A case study of a dysarthric speaker using the DragonDictate voice recognition system. Journal for Computer Users in Speech and Hearing, 8 (1), 33-52.

Ferrier, L. J., Shane, H. C., Ballard, H. F., Carpenter, T., & Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. Augmentative and Alternative Communication, 11, 165-173.

Fried-Oken, M. (1985). Voice recognition device as a computer interface for motor and speech impaired people. Archives of Physical Medicine and Rehabilitation, 66, 678-681.
Hardy, J. (1983). Cerebral palsy. Englewood Cliffs, NJ: Prentice Hall.

Jayaram, G., & Abdelhamied, K. (1995). Experiments in dysarthric speech recognition using artificial neural networks. Journal of Rehabilitation Research and Development, 32(2), 162-169.

Kotler, A. & Thomas-Stonell, N. (1997). Effects of speech training on the accuracy of speech recognition for an individual with speech impairment. Augmentative and Alternative Communication, 13, 71-80.

Rabiner, L. & Juang, B. (1993). Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall.

Rosenbeck, J. & LaPointe, L. (1978). The dysarthrias: description, diagnosis, and treatment. In D. Johns (ed.), Clinical Management of Neurogenic Communicative Disorders. Boston: Little-Brown and Company, pp. 251-310.

Roy, D. & Pentland, A. (1998a). Multimodal Adaptive Interfaces. Proceedings of the AAA Spring Symposium Series, Stanford, March 1998.

Roy, D. & Pentland, A. (1998b). Word Learning in a Multimodal Environment. To appear in ICASSP '98, Seattle, May 1998.

Schmitt, D.G. & Tobias, J. (1986). Enhanced Communication for the severely disabled dysarthric individual using voice recognition and speech synthesis. Proceedings of RESNA 9th Annual Conference, 304-306.

Shein, F., Brownlow, N., Treviranus, J., & Parnes, P. (1990). Climbing out of the rut: the future of interface technology. In Beth A. Mineo (Ed.), 1990 ASEL Applied Science and Engineering Laboratories: Augmentative and Alternative Communication in the Next Decade p. 17-19. Wilmington: University of Delaware Press.

Shriberg, L. D. & Kent, R. D. (1995). Clinical Phonetics: Second Edition. Needham Heights, MA: Allyn & Bacon.
Treviranus, J., Shein, F., Haataja, S., Parnes, P., & Milner, M. (1991). Speech recognition to enhance computer access for children and young adults who are functionally non-speaking. Proceedings of RESNA 14th Annual Conference, 308-310.

Yorkston, K.M. , Beukelman, D.R., & Bell, K.R. (1988). Clinical Management of Dysarthric Speakers. Austin, TX: PRO-ED, Inc.