Audiovisual Phonologic-Feature-Based Recognition of Dysarthric Speech

Mark Hasegawa-Johnson, Jon Gunderson, Thomas Huang, and Adrienne Perlman

Automatic dictation software with reasonably high word recognition accuracy is now widely available to the general public. Many people with gross motor impairment, including some people with cerebral palsy and closed head injuries, have not enjoyed the benefit of these advances, because their general motor impairment includes a component of dysarthria: reduced speech intelligibility caused by neuromotor impairment. These motor impairments often preclude normal use of a keyboard. For this reason, case studies have shown that some dysarthric users may find it easier, instead of a keyboard, to use a small-vocabulary automatic speech recognition system, with code words representing letters and formatting commands, and with acoustic speech recognition models carefully adapted to the speech of the individual user. Development of each individualized speech recognition system remains extremely labor-intensive, because so little is understood about the general characteristics of dysarthric speech. We propose to study the general audio and visual characteristics of articulation errors in dysarthric speech, and to apply the results of our scientific study to the development of speaker-independent large-vocabulary and small-vocabulary audio and audiovisual dysarthric speech recognition systems.

Scientific Merit

This project will research word-based, phone-based, and phonologic-feature-based audio and audiovisual speech recognition models for both small-vocabulary and large-vocabulary speech recognizers, designed to be used for unrestricted text entry on a personal computer. The models will be based on audio and video analysis of phonetically balanced speech samples from a group of speakers with dysarthria. Analysis will include speakers with reduced intelligibility caused by dysarthria, categorized into the following groups: very low intelligibility (0-25% intelligibility, as rated by human listeners), low intelligibility (25-50%), moderate intelligibility (50-75%), and high intelligibility (75-100%). Interactive phonetic analysis will seek to describe the talker-dependent characteristics of articulation error in dysarthria; based on analysis of preliminary data, we hypothesize that manner of articulation errors, place of articulation errors, and voicing errors are approximately independent events. Preliminary experiments also suggest that different dysarthric users will require dramatically different speech recognition architectures, because the symptoms of dysarthria vary so much from subject to subject. We propose to develop and test at least three categories of audio-only and audiovisual speech recognition algorithms for dysarthric users: phone-based and whole-word recognizers using hidden Markov models (HMMs), phonologic-feature-based and whole-word recognizers using support vector machines (SVMs), and hybrid SVM-HMM recognizers. The models will be evaluated to determine, first, overall recognition accuracy of each algorithm, second, changes in accuracy due to learning, third, group differences in accuracy due to severity of dysarthria, and fourth, dependence of accuracy on vocabulary size. The results of this research will contribute to scientific and technological knowledge about the acoustic and visual properties of dysarthric speech.

Broader Impacts

This research will provide the foundation for constructing a speech recognition tool for practical use by computer users with neuromotor disabilities. Tools and data developed in this research will all be released open-source, and will be designed so that, if successful, the technology developed for this proposal may be easily ported to an open-source audiovisual speech recognition system for dysarthric users.

Audiovisual Phonologic-Feature-Based Recognition of Dysarthric Speech

Mark Hasegawa-Johnson, Jon Gunderson, Thomas Huang, and Adrienne Perlman

1 Background and Objectives

1.1 Broader Impacts

Speech and language disorders result from many types of congenital or traumatic disorders of the brain, nerves, and muscles [6]. Dysarthria refers to the set of disorders in which unintelligible or perceptually abnormal speech results from impaired control of the oral, pharyngeal, or laryngeal articulators. The specific type of speech impairment is often an indication of the neuromotor deficit causing it, therefore speech language pathologists have developed a system of dysarthria categories reflecting both genesis and symptoms of the disorder [16, 17, 21]. The most common category of dysarthria among children and young adults is spastic dysarthria [48]. Symptoms of spastic dysarthria vary from talker to talker, but typical symptoms include strained phonation, imprecise placement of the articulators, incomplete consonant closure resulting in sonorant implementation of many stops and fricatives, and reduced voice onset time distinctions between voiced and unvoiced stops.

We are interested in spastic dysarthria because it is the most common type of severe, chronic speech disorder experienced by students at the University of Illinois, as well as being one of the most common types of dysarthria generally [48]. Spastic dysarthria is associated with a variety of disabilities such as, but not limited to, cerebral palsy and traumatic brain injury [16, 17, 21]. 0.26% of all seven-year-old children in the United States have moderate or severe cerebral palsy, and an additional 0.2% are reported to have mild cerebral palsy [44]. Adults with cerebral palsy are able to perform most of the tasks required of a college student, including reading, listening, thinking, talking, and composing text: in our experience, their greatest handicap is their relative inability to control personal computers. Typing typically requires painstaking selection of individual keys. Some students are unable to type with their hands (or find it too tiring), and therefore choose to type using a head-mounted pointer. Many students with noticeable dysarthria are less impaired by their dysarthria, in daily life, than by their inability to use computers.

Several studies have demonstrated that adults with dysarthria are capable of using automatic speech recognition (ASR), and that in some cases, human-computer interaction using speech recognition is faster and less tiring than interaction using a keyboard [7, 8, 9, 20, 38, 73, 31]. With few exceptions, the technology used in these studies is commercial off-the-shelf speech recognition technology. In the early 1990s, commercial speech recognizers were speaker-dependent systems, meaning that the recognition models were completely retrained for each new user. Since the mid-1990s, most commercial systems have been speaker-adaptive systems, meaning that recognition models are initialized using data from hundreds of different speakers, then adapted to the speech of each user with an algorithm such as MLLR [43] or MAP [42] adaptation. Rhagavendra et al. [61] compared recognition accuracy of a speaker-adaptive system and a speaker-dependent system. They found that the speaker-adaptive system adapted well to the speech of speakers with mild or moderate dysarthria, but the recognition scores were lower than for an unimpaired speaker. The subject with severe dysarthria was able to achieve better performance with the speaker-dependent system than with the speaker-adaptive system.

Dysarthric speakers may have trouble training ASR systems, especially speaker-dependent systems, because of the great amount of training data required. Reading a long training passage can be very tiring for a dysarthric speaker. Doyle et al. [20] asked six dysarthric speakers and six unimpaired speakers to read a list of 70 words once in each of five training sessions. They found that the word recognition accuracy of a speaker-adaptive ASR increased rapidly after the first training session, then increased more gradually during training sessions two through five. Chen et al. [11]

Table 1: Phoneme production errors in dysarthria, as reported in [36], listed together with the distinctive feature or features changed by the given phoneme substitution [68]. Phoneme labels are given in ARPABET notation [83].

Articulatory Deficit	Distinctive Feature(s)	Examples
tongue positioning	blade	T vs. K
tongue blade position	anterior	SH vs. S
oral-laryngeal timing	spreadglottis	T vs. D
degree of closure	continuant	T vs. S
degree of closure	sonorant	P vs. M, K vs. NG
vowel articulation	advancedtongueroot	UW vs. UH
vowel articulation	reduced, front	AE vs. AX

studied the speech of a subject with intelligibility (as rated by human listeners) of only 15%, and found that after ten iterations of each word in a ten-word vocabulary, automatic word recognition accuracy was raised to 90%.

Most studies of speech recognition for dysarthric talkers have focused on small-vocabulary applications, with vocabulary sizes ranging from ten to seventy words. Sanders et al. [63] studied the effect of vocabulary size on word recognition accuracy, using phone-based speaker-dependent recognizers trained on the speech of two dysarthric and two unimpaired speakers. In four small-vocabulary tasks, with perplexity ranging from 2 to 13, word recognition accuracy of the dysarthric talkers ranged from 87.8% to 100% (compared to 96.4-100.0% for unimpaired speakers). In a medium-vocabulary task (vocabulary of 516 words, with no language model), word recognition accuracy for the dysarthric talkers ranged from 0.0% to 79.4% (compared to 35.1-100.0% for unimpaired speakers). No paper in the literature reports an attempt to train large-vocabulary speech recognition for a dysarthric talker.

To our knowledge, there is not currently any commercial or open-source product available that would enable people in this user community to enter unrestricted text into a personal computer via automatic speech recognition. Our proposed experiments will result in, first, a multi-microphone, multi-camera audiovisual database of dysarthric speech, and second, programs and training scripts that could form the foundation for an open-source speech recognition tool designed to be useful for dysarthric speakers. Human subjects participating in this research will have the option of allowing their own recordings to be released to interested researchers at other research institutions; all recordings from willing talkers will be released following the same protocol we have used to distribute our AVICAR multimodal speech corpus [41].

1.2 Scientific Merit

The speech impairments resulting from spastic dysarthria are neither arbitrary nor unpredictable; indeed, van Santen and his colleagues demonstrated a dynamic systems model of vowel distortion under dysarthria [74]. Table 1 lists a number of specific phoneme substitutions errors attested in the literature [36]. As emphasized by the organization of the table, most of the specific impairments reported in the literature can be characterized as imprecision in the implementation of one or two distinctive features; e.g., /t/→/k/ is a mistake in the place of articulation of the stop. This proposal will use the term "phonologic feature" to mean any binary classification of the set of all possible phonemes or prosodic contexts (including prosodic features, and including the presence vs. absence of articulatory gestures [5]), while the term "distinctive feature" will refer to the particular set of features proposed by Miller & Nicely [51], as renamed and expanded by Stevens [68].

As shown in Table 1, phoneme production errors reported in the literature seem to be primarily errors in the production of one distinctive feature. Table 1 is much too small to draw such a conclusion with any confidence, so in preparation for this proposal, we phonetically transcribed four long recordings from [2]: a phonetically rich read paragraph (the "grandfather passage"), and three diadokinesis sequences (each consisting of twenty to thirty intended repetitions of the same

Table 2: Pronunciation errors found in paragraph reading and diadokinesis, one male talker from [2], phonemically transcribed at the University of Illinois. Phonemes are labeled using ARPABET notation [83].

Phonemes	Count	Distinctive Features	Phonemes	Count	Distinctive Features
P → B	31	spreadglottis	NG → N	1	blade
T → D	24	spreadglottis	Z → N	1	sonorant, continuant
K → G	19	spreadglottis	K → NG	1	sonorant, continuant
P → M	2	sonorant	F → H	1	sonorant, continuant, lips
S → Z	1	spreadglottis	AA → AX	1	reduced
T → N	1	sonorant	AE → AX	1	reduced
Z → D	1	continuant	D → DX	1	reduced
ZH → Z	1	anterior	IH → AX	1	reduced

syllable; the intended syllables were "puh," "tuh," and "kuh"). All four passages were read by one male talker diagnosed with moderate spastic dysarthria. All words in the grandfather passage whose pronunciation differed from that given in a standard American English pronunciation dictionary [37] were marked as "errors;" likewise, all consonant closures produced during diadokinesis as anything other than the target unvoiced stop were marked as errors. Table 2 lists all substitution errors found in this corpus; deletion errors are not listed, and there were no insertion errors.

In 1955, Miller and Nicely [51] showed that errors in the perception of different distinctive features are independent: when a phoneme with distinctive features $[d_1, \ldots, d_N]$ is produced, the probability that a listener will hear a phoneme with distinctive features $[\hat{d}_1, \ldots, \hat{d}_N]$ is

$$p(\hat{d}_1, \ldots, \hat{d}_N | d_1, \ldots, d_N) \approx p(\hat{d}_1 | d_1) \ldots p(\hat{d}_N | d_N) \quad (1)$$

One of the implications of equation 1 is that errors in perception of one distinctive feature are far more common than errors in perception of two features. The results in Tables 1 and 2 show a pattern similar to the Miller and Nicely results: based on patterns of phoneme error under dysarthria reported in the literature, and based on our own analysis of data from the Aronson recordings [2], it appears plausible that the probability of a phoneme substitution error may be factored into the independent probabilities of distinctive feature substitutions. If true, this hypothesis could have important implications for theories of speech production, just as the finding of Miller and Nicely continues to drive research in speech perception down to the present day [23, 1]. One of the goals of the proposed research, therefore, will be to test, by manual transcription and interactive phonetic analysis, the hypothesis that the probability of a phoneme substitution error in dysarthria may be factored into independent probabilities of individual distinctive feature errors.

Besides its scientific implications, the factorability of distinctive feature production errors suggests a radically new approach to the problem of automatic speech recognition for dysarthric users. In preliminary results described in Section 3.4, we demonstrate that, for some dysarthric users, it may be possible to achieve better word recognition accuracy using a bank of binary feature-based SVMs in place of the usual hidden Markov model (HMM)-based speech recognition algorithm. Our preliminary results also seem to imply that different dysarthric users benefit most from the use of different speech recognition algorithms: apparently because of differences in the symptoms exhibited, we find that some speakers achieve the highest word recognition accuracy using HMMs, while others achieve higher accuracy using SVMs. We propose to train and test a large variety of word-based, phoneme-based, and phonologic-feature based speech recognizers. By testing these recognizers with data recorded by a wide variety of dysarthric talkers, we propose to explore the relationship between the symptoms of a dysarthric speaker (the specific types of speech variability and articulatory error that he or she produces) and the acoustic observations, machine learning algorithms, and vocabulary size most appropriate to give that speaker an accurate and usable individualized automatic speech recognizer.

Human speech perception is essentially a multimodal process, which encompasses not only the audio, but also other information sources, including observations of the lip and facial motions of the speaker [49]. To our knowledge, recognition of dysarthric speech using both audio and visual information has never been attempted. We propose to record subjects using synchronized audio and video recordings, and to determine whether speech recognition models using audiovisual information can outperform audio-only recognition of dysarthric speech.

2 Results of Prior NSF Support

2.1 CAREER: Landmark-Based Speech Recognition in Music and Speech Backgrounds

Mark Hasegawa-Johnson; July 1, 2002 to June 30, 2007.

"Landmarks" are articulatory events with noise-robust, relatively invariant acoustic correlates. Stevens [66, 67, 69] claimed that lexical access, in speech perception or speech recognition, requires the listener or computer to detect and classify only four types of landmarks: consonant releases, consonant closures, syllable nuclei, and intervocalic glides. Stevens defined two tasks that a landmark-based speech recognizer must perform. First, a landmark-based recognizer must *detect* the landmarks, i.e., it must determine that a landmark of a particular type has occurred. Second, a landmark-based recognizer must *classify* the landmark, i.e., it must determine the distinctive features present at a landmark. Both of these are binary classification tasks: each landmark *detector* is trained to determine whether a particular type of landmark is present vs. absent, and each landmark *classifier* is trained to perform a particular type of binary distinctive feature classification. Dr. Hasegawa-Johnson's research has developed two classes of noise-robust landmark-based speech recognition. First, we have modeled consonant releases in noise using a variety of Bayesian methods, including HMMs [60], factorial HMMs [19, 18], particle filters [22], and a generalized maximum-likelihood acoustic feature transformation designed to alleviate the problem of model-feature mismatch [55, 56, 54, 58, 57, 59]. We have found, however, that purely Bayesian methods fail to adequately represent the acoustic correlates of distinctive features at landmarks, because landmarks are rare: a typical monosyllabic word contains 20-30 centisecond frames, but only three landmarks. Therefore, since 2003, research has focused on a machine learning technology designed to learn from a very small number of training tokens: the support vector machine.

A support vector machine represents every training token as a point in N-dimensional space, where N is the length of the observation vector. The goal of classification is to minimize the expected error rate of the classifier on some unknown future test data. Unfortunately, all that we know about the future test data is its source (e.g., we may know the talker, or we may not; in worst case, we know only that the source will be speech). In order to learn all that we can about the test data, it is helpful to record a large training database; test data will not be identical to training data, but should be similar. If the training database is large enough, we can minimize *test* error rate by minimizing *training* error rate; this is the approach taken by neural networks and discriminatively trained HMMs. If the training database is not very large (as in landmark-based speech recognition), it is helpful also to estimate an upper bound on the possible difference between training error rate and test error rate. Support vector machines are trained in order to minimize $S = G + E$, where E is the error rate of the classifier on training data, and G is an upper bound on the expected difference between training and test error rates. Because of the term G, it is possible to train an SVM using as few as one labeled training token per class: the best current speaker verification systems include SVMs that have been trained using only one positive example (one recording from the target speaker), and many negative examples (recordings from other speakers) [65].

In summer 2004, at the Johns Hopkins workshop WS04, Dr. Hasegawa-Johnson led a team of six faculty, four graduate students, and two undergraduates in a six-week effort to achieve commercially viable landmark-based large vocabulary speech recognition [26, 27]. The final system used 72 SVMs trained to detect and classify different categories of landmark. SVMs were used to rescore the word lattice output of the SRI speech recognizer [71, 70]. SVMs were either used directly, to choose the best possible transcription from a list of alternative transcriptions, or indirectly, as part of

a hybrid SVM-DBN (dynamic Bayesian network) architecture [46, 47]. During WS04, distinctive feature classification error rates of some SVMs were reduced by as much as 50% (through extensive experimentation), but word recognition accuracy of the complete recognizer did not beat the baseline. Since WS04, we have continued this development effort, and have achieved binary classification error rates below 10% for 33 different landmark detectors and classifiers. In spring 2005, a hybrid SVM-HMM speech recognizer was constructed, using the outputs of the SVMs as observations in an HMM; using SVMs instead of regular spectral features (MFCCs) resulted in a small but significant error rate reduction [4].

2.2 Multimodal Human Computer Interaction System: Toward a Proactive Computer

PI: Thomas Huang; July 1, 2000 to June 30, 2005.

The testbed of this ITR-funded project is an intelligent-tutoring environment for education in science and technology, using the Lego construction set, with children of primary and middle school age. The emphasis is on developing a proactive computer agent to encourage the interests of kids in science and technology. Communication between the computer agent and the user is via a multimodal spoken dialogue system, including audiovisual speech recognition, and audiovisual recognition of the cognitive state of the user (confused, confident, or frustrated) [78, 77, 62, 79, 80, 81].

In this grant and related grants over the past several years, research in our group on tracking of the head and lips has led to a robust 3D facial motion tracking system [72]. A 3D non-rigid facial motion model is used in a multi-resolution manner so that the speed of face movement tracking can be adjusted to match the computational resources available. Audiovisual speech recognition may be implemented by simply concatenating together observation vectors from the audio input (from the microphones) and observations from the visual input (from the cameras). We have obtained considerably improved performance, however, by fusing audio and visual data using an architecture we call the coupled hidden Markov model (CHMM) [14, 15]. A CHMM is a set of parallel HMMs, each with its own independent observations (audio or visual); state transition probabilities in one HMM depend on the current state values in all other HMMs. The topology of the CHMM ensures that learning and classification are based on the audio and visual domains jointly, while allowing asynchronies between the two information channels. The benefits of the CHMM have been confirmed by a series of experiments on audio-visual speech recognition [14, 15]. In one experiment, for example, white noise was added to the audio channel in a multimodal speech database. At 20dB SNR, with a 40-word vocabulary, speech recognition was 44% accurate using an audio-only HMM, and 43% accurate using a video-only HMM; when the audio and video HMM were merged into a CHMM structure, word recognition accuracy went up to 87%.

2.3 Audiovisual Speech Recognition in an Automotive Environment

Mark Hasegawa-Johnson, Thomas Huang, and Stephen E. Levinson; May 15, 2002 to May 14, 2006.

This project was funded by the Motorola Communications Center, not NSF, but is included here because we propose to use the hardware developed by this project in our research on Universal Access. A portable multimodal recording array was developed, composed of a horizontal array of eight microphones (in a wooden baffle), and a horizontal array of four cameras. The left side of Fig. 1 shows the transducer array deployed in an automobile; the camera array is attached to the dashboard, and the microphone array is attached to the sunvisor. The right side of Fig. 1 shows an image acquired by the same array, in an office environment. Using this array, 120 talkers have been recorded in an automotive environment, producing isolated digits, telephone numbers and phonetically balanced TIMIT sentences under five different noise conditions: engine idling, 35mph with the windows closed/open, and 55mph with the windows closed/open [41]. Sampler DVDs containing data from four talkers were distributed at ICSLP 2004 to roughly 50 interested researchers. The complete database (100G, on an external hard disk) has been distributed to four laboratories on three continents (Motorola, Northwestern University, Tsinghua University, and the University of Saarlandes). Our current research seeks to improve automatic speech recognition in these difficult environments using multi-microphone techniques [40], and using integration of the audio and visual speech information.

Figure 1: Left image shows the AVICAR transducer array deployed in an automobile: four cameras (on dashboard), eight microphones (embedded in wooden baffle, at level of the sunvisor). Right: the image of one dysarthric subject, acquired by deploying the AVICAR array on top of the monitor of a personal computer.

3 Preliminary Experiments: Data Acquisition and Analysis

3.1 Subjects

In preparation for this proposal, data were recorded from four subjects with self-professed speech disorders: three male (M01, M02, M03), and one female (F01). No formal speech pathology evaluation was performed prior to involvement in this preliminary study. Informal evaluation of recorded data found that subjects M01, M03, and F01 exhibit different symptoms of spastic dysarthria; specific symptoms are discussed below (Tables 3 and 4). Subject M02 exhibited symptoms of chronic stuttering, but no symptoms of dysarthria. Human subjects protocols were approved prior to start of research by the University of Illinois Institutional Review Board. All subjects were asked to specifically approve or disapprove five possible uses of the recordings: research at the University of Illinois, presentation of voice samples at professional conferences, presentation of video at professional conferences, distribution of voice samples to speech researchers at other institutions, and distribution of video to speech researchers at other institutions. All subjects voluntarily approved all five uses of their data.

3.2 Data Acquisition

Subjects were recorded using the array of microphones and transducers shown in Fig. 1. Cameras and microphones were mounted on top of a computer monitor. One-word prompts were displayed on the monitor using PowerPoint. Three of the four subjects were unable to control a keyboard or mouse, therefore an experimenter sat next to the monitor, advancing the PowerPoint slides after each word spoken by the subject. Each slide advance generated a synchronization tone, dividing the recording into one-word utterances. Four types of speech data were recorded. Isolated digits (zero through nine) were each recorded three times. The letters in the international radio alphabet (alpha, bravo, charlie,...) were each recorded once. Nineteen computer command words (line, paragraph, enter, control, alt, shift,...) were each recorded once. Finally, subjects read, one word at a time, in order, the words of a phonetically balanced text passage (the "Grandfather Passage," 129 words), and 56 phonetically balanced sentences (TIMIT sentences sx3 through sx59 [83]). Each subject recorded a total of 541 words, including 395 distinct words.

3.3 Intelligibility and Phonetic Analysis

Intelligibility tests were performed using 40 different words selected from the TIMIT sentences recorded by each talker. Selection was arbitrary, with the constraints that listeners should never hear two consecutive words from the same sentence, and that listeners should never hear the same word from two different talkers. Words selected in this way were presented to listeners on a web

Table 3: Three listeners (L1, L2, L3) attempted to understand isolated words produced by four talkers (F01, M01, M02, M03); percentage accuracy is reported here.

Listener	F01	M01	M02	M03
L1	22.5%	22.5%	90%	30%
L2	17.5%	20%	90%	27.5%
L3	17.5%	15%	97.5%	30%
Average	19.2%	19.2%	92.5%	29.2%

Table 4: Number of production errors of each type, out of a total of 289 words in error. DEL=deletion, INS=insertion, SUB=substitution, NS=erroneous number of syllables, WD=word deletion (labeler unable to guess the word).

	Initial Cons.			Medial Cons.			Final Cons.			Vowel		Word
	DEL	INS	SUB	DEL	INS	SUB	DEL	INS	SUB	SUB	NS	WD
All	37	6	83	16	7	63	45	32	64	74	29	87
F01	8	2	25	2	2	20	15	5	15	22	5	46
M01	3	0	40	11	4	20	18	17	19	34	14	27
M02	2	2	3	0	0	1	0	2	0	1	0	0
M03	24	2	15	3	1	22	12	8	30	17	10	14

page. Listeners were asked to listen with headphones, and to determine which word was being spoken in each case. The first listener (L1) is the PI; other listeners (L2 and L3) are students in his lab. Neither student was present when the data were first recorded, and neither student has formal training or extensive experience in the perception or judgment of dysarthria; it has been shown that listeners with formal training are usually able to understand dysarthric subjects with higher accuracy. Results are presented in Table 3. Several findings are apparent. First, talker M02 is much more intelligible than the other talkers. His stutter did not interfere with intelligibility to the same extent as the spastic dysarthria of other subjects. Second, inter-listener agreement is very high. Listener L1 was able to understand dysarthric subjects with slightly higher accuracy than the other two listeners, apparently because he had experience listening to these three dysarthric speakers. For this reason, average intelligilibility scores listed in the last row of the table may be a little too high; a more accurate estimate might be obtained by averaging the accuracies of listeners L2 and L3.

Listener errors (289 tokens) were phonologically analyzed; results are shown in Table 4. Three consonant positions were distinguished: word-initial cluster, word-final cluster, and others (word-medial). Consonants in each position could be deleted ("sport" heard as "port"), inserted ("on" heard as "coin"), or substituted ("for" heard as "bore"). Substitution errors were almost equally likely to be manner, place, or manner+place errors; obstruent voicing errors were less common. Three other types of errors were tracked. First, vowel substitutions were tracked (e.g., "and" heard as "end"). Second, the number of syllables could change ("NS"): 81 of the intended words were monosyllabic, 40 bisyllabic, 35 trisyllabic, and 4 quadrisyllabic. Third, the entire word could be deleted ("WD"). Listener L1 (the PI) never used the WD rating, but L2 and L3 used it whenever a word failed to sound like human language – a relatively frequent occurrence, as many words sounded more like a squeak or moan than a word. Table 4 shows that, although talkers M01 and F01 had similar intelligibility scores, the types of errors associated with their productions were very different. F01 suffered more "word deletions" than any other talker, meaning that her words were frequently not recognizably intended to be words. The speech of M01 exhibited a very slow and painstakingly enunciated stutter, and this slow stutter sometimes gave listeners the mistaken impression of inserted final consonants, or of inserted or deleted syllables. M03, by contrast, attempted to maintain a reasonable speaking rate, but in the process, deleted more word-initial consonants than any other speaker. Across all speakers, word-initial and word-final consonant errors were more frequent than word-medial consonant and vowel errors.

Table 5: Columns "H" report word recognition accuracy (WRA, in percent) of HMM-based recognizers if all microphone signals are independently recognized; columns "HV" report WRA if all microphones vote to determine final system output. "Word" reports accuracy of one SVM trained to distinguish isolated digits, treating each microphone signal independently. "WF" adds outputs of 170 binary word-feature SVMs. "WFV:" Like WF, but single-microphone recognizers vote to determine system output.

Vocabulary	395 Words		45 Words		10 Words (Digits)				
Algorithm	H	HV	H	HV	H	HV	Word	WF	WFV
F01	17	17	44	55	71	80	97	86	90
M01	17	22	42	49	86	95	70	69	70
M02	58	62	87	89	99	100	90	90	90
M03	40	44	77	80	99	100	97	100	100

3.4 Automatic Speech Recognition

In other research, we have used Dragon Dictate software to, rapidly and with minimum labor cost, acquire 95% accurate automatic transcriptions of speech produced by subjects without pathology [82]. The graduate research assistant who typically runs Dragon for our other research was asked to use the same protocol to produce a baseline automatic transcription of dysarthric speech data. It was impossible to train a Dragon recognition model for each dysarthric subject, because dysarthric subjects could not read through the standard training text, therefore, contrary to recommendations by the product manufacturer, we attempted to recognize these data using mis-matched models (models trained by a speaker without pathology). Resulting word recognition accuracy ranged from 0% (for subject M02: number of insertions equals the number of words correct) to -140% (for subject M01: no words were correctly recognized, and the number of insertions was 140% of the size of the reference transcription).

Better results have been reported with HMM-based systems that allow isolated-word recognition (e.g., Dragon Naturally Speaking), therefore our next experiment involved the design and test of HMM-based isolated word recognizers. Using the HTK toolkit [76], speaker-dependent speech recognizers were trained and tested. All systems used a relatively standard HMM architecture: monophone or clustered triphone HMMs [53], three states per phone, mixture Gaussian observation PDFs, PLP+energy+d+dd spectral observations [30]. Apparently because of the small training corpus, simple models outperformed complex models: monophone recognizers outperformed clustered triphones in all cases, and the optimum number of Gaussians in the mixture Gaussian PDF was always less than 10.

In the first experiment, models were trained using odd-numbered utterances, and tested using even-numbered utterances. The recognizer was constrained to recognize just one word per utterance (with optional silence before and after the word), with a vocabulary size of 395 (the number of distinct words in the database). In Table 5, columns "H" reports accuracy when every microphone recording is treated as an independent training or test utterance. Column "HV" implements a simple kind of multi-microphone combination: each microphone signal is independently recognized, and any word recognized by a plurality of the microphones is taken to be the final system output. This voting scheme was found to be more accurate, for these data, than training and testing a one-channel speech recognizer on the output of a delay-and-sum beamformer. Table 5 demonstrates that this configuration yields unacceptable accuracy for all four speakers.

In the second experiment, models were tested using a 45-word vocabulary that included the 19 computer command words, the 26 letters of the international radio alphabet, and the 10 digits. Test data included two utterances of each digit, and one utterance of each of the other 35 words. All other data were used to train monophone HMMs: other data included TIMIT sentences, the Grandfather passage, and one other utterance of each digit. The third experiment used the same training data, but test data were restricted to include only the digits; the recognizer was restricted to select the

best option from a 10-word vocabulary. Results are reported in Table 5. With a 45-word vocabulary, the "HV" scheme is nearly acceptable for subject M02, but not for any of the dysarthric subjects. With a 10-word vocabulary, the "HV" scheme is acceptable for the subjects M01, M02, and M03, but unacceptable for subject F01.

Our research on landmark-based speech recognition [26, 27, 4] has demonstrated that support vector machines (SVMs) are capable of extracting discriminative information from a sequence of acoustic spectra. SVMs were therefore tested for the task of dysarthric speech recognition. In all experiments, SVMs were tested using both linear and nonlinear (radial basis function) classifiers; best results were usually obtained with a nonlinear classifier. The start and end times of each word were first detected using a single-channel two-Gaussian voice activity detector (VAD) followed by multi-channel voting. Accuracy was verified by manually endpointing 20 multi-channel waveforms; single-channel VAD often failed, but multi-channel VAD was found to be accurate within 10ms in all 20 labeled files. SVM observations were then constructed by concatenating consecutive PLP frames, in order to construct a "cepstrogram observation." Two types of SVM were trained: 10-ary Word-SVMs, and binary Word-Feature-SVMs (WF-SVMs; Table 5). Word-SVMs were trained using two examples of each digit, while the third example was used for testing; the observation for a Word-SVM was always a cepstrogram of length 640ms or 1280ms, beginning at the beginning of the word. Word-Feature-SVMs (WF-SVMs) were a bank of 170 different binary-output SVMs, trained and tested with (10 different input cepstrograms)$\times$(17 different binary target functions). Among the 17 target functions, 7 were trained to classify distinctive features of the word-initial consonant (sonorant, fricated, strident), of the vowel (round, high, diphthong), or of the word-final consonant (nasal vs. non-nasal). The remaining 10 target functions were binary one-vs-all targets, i.e., each SVM was trained to distinguish a particular digit from all other digits. Recognizer output was computed by adding together the real-valued discriminant outputs of the SVMs, with sign permutations dependent on the distinctive features of the words being recognized, e.g. "one" is [+sonorant,-fricated,-strident,+round,-high,-diphthong,+nasal]; the word with the highest resulting score was taken as the recognizer output. Results are reported in Table 5 in columns "WF" (all microphones scored separately) and "WFV" (microphones vote to determine final system output).

Different recognition architectures succeed for different speakers. Speakers M02 and M03 produce speech with nearly canonical phoneme content, therefore the phone-based HMM architecture succeeds for these speakers. By contrast, the SVM-based architecture succeeds well for subjects F01 and M03, possibly because these two subjects do not stutter; the SVM observation vector is not robust to variation in the timing of phonetic events in a word. In the sections that follow, we discuss alternative SVM observation vectors that may be more robust to the timing variability associated with stuttering. We expect that the proposed research will continue to show that different dysarthric speakers will obtain best results from radically different speech recognition architectures.

A 10-word vocabulary is not sufficient for meaningful human-computer interface. In the sections that follow, we propose methods for extending the success of SVM and HMM recognition to larger vocabularies.

4 Database Acquisition and Speech Recognizer Development

4.1 Subjects

During the first two years of the proposed research, we will record subjects with Spastic Cerebral Palsy (the most common CP diagnosis). We propose to record 50 subjects, including, if possible, 25 female subjects. Subjects will be recruited from several locations in the state of Illinois, based primarily on word-of-mouth contacts established with the help of interested organizations in Urbana-Champaign (Division of Rehabilitation - Education Services, PACE), and with the help of United Cerebral Palsy of Illinois. See letters of support from these organizations, attached as supplementary documents to this proposal, in which they offer their assistance in publicizing the study to potential participants. Subjects will be free to terminate participation at any time and they will be compensated for their time ($25/hour) and travel expenses (actual costs or university travel reimbursement

rates).

4.2 Data Acquisition Methods

Subjects will be asked to read word lists including (1) computer command words (19 words), (2) letters of the international radio alphabet (26 words), (3) isolated digits (ten words), (4) a list of the 150 most frequent words in English text (primarily function words), and (5) phonetically balanced word lists: instead of reading the Grandfather passage and TIMIT sentences, as in our preliminary research, subjects will read through 12 phonetically balanced word lists with each list containing 50 words. (1)-(4) are designed to be useful for training whole-word speech recognizers; phonetically balanced word lists will help us to train phone-based and distinctive-feature-based recognizers. Including all word lists, there will be a total of 795 unique words. The subjects will read these lists during 2 separate sessions that will last up to 2 hours each. During each session the subject will read all of the 795 words once. Each subject will receive the same word lists, but the word ordering will be randomized for each subject to control any sequential effects of the word lists.

Each subject will be seated in a quiet room, and a headset microphone will be positioned on his or her head. For each subject, we will record 4 channels of video and 8 channels of audio. One channel of audio will be recorded from the headset microphone; seven channels of audio and four channels of video will be recorded from the array of microphones and cameras depicted in Fig. 1. The position of the headset will be controlled between subjects, so that all subjects have the same microphone position. Prompts will be presented to subjects using PowerPoint; audio will be recorded using an 8-channel ADAT, and video will be recorded using a DV tape recorder.

4.3 Intelligibility and Phonetic Analysis

Speech samples from two of the word lists from each subject will reviewed by two speech and language pathology graduate students who have completed the dysarthria course (SPSHS 385) and at least one course in acoustic phonetics at the University of Illinois. These students will listen to the samples and record the word they thought the subject pronounced. These listener transcriptions will be compared to the prompt list, as read by the subject, in order to calculate a percentage of intelligibility for each talker-listener pair, and in order to calculate the average intelligibility of each talker. Previous studies have shown that intelligibility of dysarthric speech is a function of listener experience, thus we expect that intelligibility scores computed as described here (with graduate students in Speech and Hearing Science as the listeners) will be somewhat higher and less variable than intelligibility scores computed with the naïve listeners (engineering students) who participated in our preliminary experiments.

Average intelligibility will be used to classify each talker into one of four categories: very low (0-25%), low (25-50%), medium (50-75%) or high (75-100%) intelligibility.

Data in each of these four groups will be subjected to interactive phonetic analysis, in order to test the hypothesis that the probability of a phoneme substitution error in these dysarthric productions may be factored into independent probabilities of distinctive feature errors. Two graduate students will interactively segment and phonetically transcribe at least 400 phonetically balanced words from at least 8 subjects (one male subject and one female subject per intelligibility category), for a total of 3200 transcribed words. The phonetic transcription will seek to label the manner, place of articulation, and voicing with which every consonant is physically articulated, and the height, fronting, tenseness, and rounding of every vowel, regardless of whether or not the articulated segment is a valid phoneme in the English language. Often, in dysarthric speech, secondary articulations (especially voicing and nasality) are mis-timed relative to the primary articulation of a consonant (the oral closure); events of this type will be noted, and the start and end times of each secondary articulation will be labeled.

Words will be transcribed in blocks of 100, including 50 words transcribed using audio only, and 50 words transcribed using audio and video. The two graduate students working on this project will work separately to transcribe each block of 100 words, and will then confer in order to resolve their differences and agree on a common transcription. The result of this effort will be a set of three transcriptions: one transcription produced in isolation by each graduate student, and

one transcription resulting from the conference of the two students. Inter-transcriber agreement will be calculated by comparing the transcriptions of the two students working in isolation, and by comparing each of their transcriptions to the conference transcription; results will be reported using the kappa statistic [10]. Inter-transcriber agreement will be computed separately for words transcribed using audio information versus audiovisual information, in order to determine whether or not visual information helps human listeners to more precisely decide the phonetic content of an utterance. Inter-transcriber agreement will also be computed separately for each block of words, in order to determine whether the process of working out a conference transcription for one block of words helps to improve inter-transcriber agreement on future blocks. If inter-transcriber reliability statistics change significantly as a function of block number, or as a function of observation type (audio vs. audiovisual), transcribers will revisit all data blocks with low inter-transcriber reliability, in order to bring reliability up to the same level for all transcribed blocks. Once reliability is at the same level for all blocks, data from all blocks will be pooled for each talker, resulting in a total of 400 words of data for each talker.

This research will conclude with an evaluation of the hypothesis that the probability of a phoneme production error may be factored into the product of independent distinctive feature error probabilities (Eq. 1). Phoneme pronunciation errors will be compiled in the form of confusion matrices [51]. Separate confusion matrices will be constructed for each talker. The proposed hypothesis will be compared to two alternative hypotheses, with, respectively, more and fewer degrees of freedom than the proposed hypothesis. The first alternative hypothesis proposes that all phoneme errors are equally likely. The second alternative hypothesis proposes that phoneme error probabilities are constrained only by symmetry, i.e., the probability of producing "b" in place of "p" is the same as the probability of producing "p" in place of "b." These three models will be compared using a Bayesian information criterion [64], and, if bilinearly transformed frequencies show a Gaussian distribution [33], using an F-ratio test [29].

4.4 Software Development

Based on our preliminary experiments, we propose the hypothesis that it will not be possible to use the same speech recognition architecture for all subjects. Dysarthria induces variability, and different subjects exhibit dramatically different symptoms. We propose therefore to test, for every subject, a large number of audio and audiovisual speech recognition architectures, including HMMs, whole-word and distinctive-feature based SVMs, and a hybrid SVM-HMM architecture [4, 26]. All of these architectures will be the subject of intensive experimentation during the course of the proposed research, in order to find the best possible configuration or configurations of each algorithm for dysarthric subjects with a variety of different symptom sets. We expect that one graduate research assistant will focus on optimizing audio and audiovisual HMM and DBN architectures, while the other will focus on optimizing audio and audiovisual SVMs. The result of this research will be a set of algorithm training and test scripts capable of automatically optimizing an algorithm for a new subject (using, as training data, one utterance of each word list by a given subject), and of automatically evaluating its performance (using, as test data, the other utterance of each word list).

We will test each of the proposed recognizers with three closed-set vocabularies (10-word, 45-word, and 195-word, where the 195-word vocabulary includes the list of 150 common English words), and three or more open-set vocabularies (2000-word, 6000-word, and 20000-word; larger vocabularies will be tested only when the smaller vocabulary has been proven successful). Open-set vocabularies include words not present in the training data; closed-set vocabularies do not. The 10-word vocabulary is designed to be a fail-safe option: digit recognition is not an extremely useful human computer interface, but if nothing else works, some subjects may find a 10-word user interface to be better than none. The 45-word vocabulary is designed to allow our least intelligible subjects to spell out written documents, letter by letter. The 195-word, 2000-word, and larger vocabularies are designed for experiments in whole-word dictation-based human-computer interface. In all four cases, subjects will be able to enter common words using a whole-word recognition mode, but less common words will need to be spelled.

Video features for speech recognition will be extracted using methods adapted from our AVICAR

audiovisual speech recognition research. A facial feature tracking algorithm will first identify the eyes, nostrils, and corners of the mouth. Based on these feature points, a rectangle of pixels including the lip region will be extracted. The vector of pixel values will then be compressed using speaker-independent linear discriminant analysis [52]. In our AVICAR project, we are experimenting with the extraction of geometric (lip-contour-based) features; if current experiments are successful by the start of the proposed research, geometric lip features will also be used.

Speaker-dependent phone-based HMMs will be trained and tested as in our preliminary research. For small-vocabulary tasks, we will also experiment with whole-word HMMs. We propose to test a relatively large number of multi-channel audio and audiovisual integration algorithms, including a beamforming and post-processing algorithm based on our current AVICAR research [40], a voting scheme (in which eight audio-only recognizers and four video-only recognizers each get one vote), a multi-stream HMM (in which all eight audio channels and four video channels are treated as independent observations generated by the same HMM state sequence), and a coupled hidden Markov model (CHMM) [14, 15].

SVM research will focus primarily on the development of acoustic features (SVM inputs) and phonologic features (target SVM outputs) that are robust to the symptoms of dysarthria. For example, acoustic feature representations more robust to stuttering might include a linear discriminant projection of the cepstrogram, a sequence of spectra synchronized with detected landmarks [27], or a dynamic time warping of the cepstrogram [32]. SVMs will be trained based on two different types of target labels. For closed-set recognition tasks, it will be possible to use Word-SVMs (Table 5), and Word-Feature-SVMs trained using a one-vs-all binary word recognition paradigm. For open-set recognizers, it will only be possible to use SVMs trained to detect the phonologic features of the word. We will experiment with a variety of different phonologic word classifications, including distinctive features (sonorant, continuant, ...), prosodic features (number of syllables, stress pattern, ...), features that encode the presence vs. absence of various types of landmarks at various positions relative to the start or end of the word (stop releases, fricative closures, ...;[27]), and features that encode the presence vs. absence of particular articulatory gestures (tongue-tip-closure, glottal-opening, velar-opening,...; [5]). Multi-channel audio and video integration methods will include post-SVM combination (voting, boosting, or multistage SVM), and pre-SVM combination (concatenation of features from multiple audio and video channels into a single SVM observation).

Hybrid SVM-HMM recognizers will be developed by training HMMs to observe the real-valued discriminant outputs of the phonologic-feature SVMs [4]. Rather than being applied once per word, phonologic-feature SVMs will be applied once per 10ms frame of speech. The real-valued discriminant outputs of the SVMs will be concatenated to form an observation vector once per 10ms, which will be observed by an HMM, multi-stream HMM, or CHMM.

In addition to the recognition algorithms described above, we will experiment with dictionary adaptation methods, for use with the phone-based HMM and distinctive-feature-based SVM recognizers. Two types of dictionary adaptation will be tested. Phoneme-based lexical adaptation will assume that each phoneme in the dictionary may be pronounced as any other phoneme according to a probability p_{ij} given by element a_{ij} of the talker-dependent confusion matrix. Distinctive-feature-based lexical adaptation will assume that distinctive feature errors are independent events; the probability of each distinctive feature error will be estimated using the confusion matrices resulting from manual phonetic labeling. These two types of lexical adaptation will be compared to a default system that uses no lexical adaptation; instead, the default system will assume that every word must be pronounced exactly as specified in a standard pronunciation dictionary.

5 Evaluation

Evaluation of the models will determine how effective the audio and audiovisual speech recognizers are in recognizing the speech of persons with dysarthria. Evaluation will be designed to test the main effects listed in Table 6, and the interactions listed in Table 7. Main effects will include the gender of the speaker, the algorithm used for recognition, modality (audio vs. audiovisual), the severity of the dysarthric speech, vocabulary size, and word-list category (open vs. closed). Closed word

Table 6: Main effects tested during recognizer evaluations. Dependent variables (columns) will be tested for dependence on independent variables (rows) if marked with an X. Independent variables are either within-subject or between-subjects. Dependent variables: WER1=Word Error Rate of the recognizer's first-choice output. WER10=Word Error Rate after subject chooses from a list of the recognizer's 10 highest-scoring words. Time=Time required for a subject to finish the word list.

	WER1	WER10	Time
Typing vs. Automatic speech recognition (Within)			X
Recognition Algorithm (Within)	X	X	X
Modality (Audio vs. Audiovisual; Within)	X	X	X
Vocabulary Size (Within)	X	X	X
Open vs. Closed Word List (Within)	X	X	X
Subject Intelligibility (Between)	X	X	X
Subject Gender (Between)	X	X	X
Session Number (Within)	X	X	X

Table 7: Interaction effects that will be tested during recognizer evaluations in year 3 of the proposed research. Interactions will be tested only for the independent-variable pairs marked "X" in this table, except as noted in Sec. 5.4

	Modality	Vocabulary Size	Open vs. Closed	Subject Intelligibility
Algorithm	X	X	X	X
Modality		X	X	X
Vocabulary Size				X
Open vs. Closed List				X

lists contain words used during the initial training of the system. Open word lists contain words not present in the training data (only possible using a recognizer with an open-set vocabulary). We anticipate that some talkers with mild dysarthria may be able to use large-vocabulary recognition, but that as the severity of dysarthria increases, word error rate of both large-vocabulary and small-vocabulary recognition will also increase, rendering large-vocabulary recognition impractical for moderately and severely dysarthric talkers. The importance of the evaluation will be to determine the magnitude of these differences and provide an understanding of how speech recognition can be used as an input device for computer input, as well as for other uses including environmental control and augmentative communication devices.

5.1 Subjects

Subjects for this evaluation will be a subset of those described in Section 4.1. Sixteen subjects will be selected based on their expressed willingness to participate in the four sessions of the proposed user evaluation; if possible, subjects will be selected so that there are two males and two females from each of the four intelligibility categories. Subjects will be allowed to discontinue participation at any time for any reason and they will be compensated for their time ($25/hour) and travel expenses.

5.2 Software Development

Software for user evaluations will be written by the two engineering graduate students during the second year of the proposed research. Early in the second year of the grant, students working on this research will choose at least four audio-only and four audiovisual speech recognition algorithms that should be rewritten into real-time C functions, based on a comparison of recognition accuracies achieved. The graduate students will be capable of writing these programs because (1) examples of similar programs are available in our laboratory, and (2) both of them will have written isolated-word HMMs as part of their participation in the class ECE 594, "Mathematical Models of Spoken Language." The following restrictions will limit code complexity and computational complexity: (1) the search graph (2000 words, bigram grammar) is sufficiently small to be statically compiled,

therefore recognition code may be very efficient [3], and (2) the complexity of the recognizer will be limited by the requirement that subjects must insert a pause after each spoken word (so-called "discrete speech" dictation style).

5.3 Data Acquisition Methods

Subjects will participate in 4 sessions lasting approximately 90 minutes each. During each experimental session, the talker will read or type 9 word lists of 50 words each (a total of 450 words in each experimental session), resting after each word list. Each word list will be associated with one setting of the independent variables of the evaluation: (1) 45-word vocabulary, audio-only, algorithm A; (2) 45-word vocabulary, audiovisual, algorithm A; (3) 195-word vocabulary, audio-only, algorithm B (4) 195-word vocabulary, audiovisual, algorithm B; (5) 2000-word vocabulary, audio-only, closed word-list, algorithm A, (6) 2000-word vocabulary, audiovisual, closed word-list, algorithm A, (7) 2000-word vocabulary, audio-only, open word-list, algorithm B; (8) 2000-word vocabulary, audiovisual, open word-list, algorithm B; (7) Typed responses (closed word list).

Speech recognition algorithms for evaluation experiments will be selected based on the experiments performed in Sec. 4.4. For each talker, at each vocabulary level, in each modality (audio-only or audiovisual), we will select one HMM-based algorithm (including multi-stream HMM or CHMM) and one SVM-based algorithm. Under each of these conditions, the best HMM-based algorithm and the best SVM-based algorithm will be selected (as determined by the experiments in Sec. 4.4), so that, even though the details of the algorithms chosen will vary from speaker to speaker, it will be possible to make some general statements about the interactions among vocabulary size, modality, talker intelligibility, and algorithm. The HMM-based and SVM-based algorithms will be counterbalanced across talkers and sessions, i.e., if the HMM-based algorithm is "algorithm A" during the first evaluation session for one talker, then the SVM-based algorithm will be "algorithm A" during the second session for the same talker.

All subjects will read the same word lists, but the order of words will be randomized for each subject to control for any word list effects. Words for the open word lists will be selected from the 2000-word vocabulary. The order of the word list conditions will be counter balanced to control for any ordering effects during a session. The subject will not know which condition is being tested, they will only be instructed to read or type the words in the list.

The typed word list will determine the speed with which the user is able to type a word list comparable to the spoken word lists. Words in the typed list will be randomly selected from the 795-word closed-set vocabulary (the list of words recorded during training). User prompts will be identically formatted to those used during speech recognizer testing, but users will type their responses instead of speaking.

During spoken word lists, the user will be prompted to say each word with a visual prompt on a computer screen. When the subject says the word, a list of up to ten words will appear on the screen. The first word in the word list will be the most likely word estimated by the model, and the remaining ten will be the next likely candidates, and will be associated with the numbers 1 through 9. If the most likely word is correct the system will automatically move to the next word in the list after a pause of a few seconds. If the word is not correct the user will have the option of selecting from the list of 9 next most likely words by saying "Choose N" from the list, where N is the number of the word in the list. If the word is not in the list the system will automatically move on to the next word after a few seconds and the word will be recorded as not recognized. The word list technique is similar to techniques used in early speech recognition systems. The size and colors of the words can be adjusted to a comfortable reading size for the subject.

5.4 Analysis of Results

As shown in Table 6, the dependent variables for each word list include the number of words recognized, the number of words recognized by selecting from the recognizer's top-10 list, and the time to complete the list. The design of the experiment is split-plot factorial design with two between subject blocking factors and six within-subject blocking factors. The between blocking factors are gender of the subject and the level of speech intelligibility (very low, low, medium, or high). The

within subject blocking factors are whether words are spoken or typed, the speech recognition algorithm, audio vs. audiovisual inputs, vocabulary size, open vs. closed wordlists, and session number (learning effects). Learning and gender issues are considered nuisance effects. We do not plan on testing any interactions with those variables, unless significant main effects are observed. If significant main effects are observed, additional tests will then be conducted to identify the potential source of the apparent main effect, and to see if the speech recognition models can be adjusted to remove or minimize the effect.

The major purpose of the evaluation is to estimate the potential for use of each proposed algorithm in a functional speech recognition system for people with dysarthric speech. We estimate that a functional speech recognition system needs a word recognition accuracy of at least 90-95%; the goal of statistical analysis will be to determine, for subjects in each intelligibility class, the best way to achieve a 90-95% word recognition accuracy. Confidence intervals for accuracy of recognition and time to complete the tasks will be calculated (at a level of $p = 0.05$) for subjects in each grouping variable. Statistically significant main effects and interaction effects will be detected if confidence intervals do not overlap. By evaluating the confidence intervals, we expect that we will be able to make algorithm recommendations to subjects participating in the current study, and possibly also to other dysarthric users.

6 Summary

6.1 Scientific Merit

Few standard texts in acoustic phonetics, phonology, or psycholinguistics mention speech motor disorders (e.g., [68, 34, 13, 39], and [45] do not), and the only mention of dysarthria in the Handbook of Phonetic Sciences is in two chapters by professional speech pathologists [35, 75]. We do not know why standard linguistic and psycholinguistic theories ignore dysarthria; three possible reasons are a lack of understanding, a lack of interest (possibly caused by the lack of understanding), and a lack of data (the largest publicly available database of dysarthric speech, that we know of, contains 74 sentences produced by each of 10 talkers [50]). It is our belief that any complete theory of language must explain the way in which words are approximated by talkers who are unable to match, in production, the phoneme targets demanded of them by the language community. To put it succinctly, a theory of language must explain the failures of language. The experiments proposed in this study are a very small step toward the inclusion of dysarthria in theories of language. We propose, first, to test a particular theory of the relationship between articulation error and perceived phonology, and second, to provide transcribed speech data and confusion matrices that may be used by other researchers for the purpose of developing other such theories.

6.2 Broader Impacts

The research proposed in this document will develop and carefully test a suite of audio and audiovisual automatic speech recognition tools designed for users with spastic dysarthria, and a multimodal database designed to enable other researchers to experiment with automatic speech recognition for dysarthric subjects. Distribution of the database will follow the protocol established by our AVICAR project: after distributing a DVD sampler at a professional conference [41], we will encourage interested researchers to send us a hard disk. When we receive a disk from a known speech researcher, we will fill it up with data, and return it to the sender. As with our AVICAR corpus, distributed data will only include recordings of subjects who have explicitly given permission for the release of their data to other research institutions. Software developed in this research will be released open-source on our web page, like all other databases and software developed by this PI [12, 24, 28, 25].

References

[1] Jont B. Allen. Articulation and intelligibility. Unpublished manuscript, 2003.

[2] Arnold Aronson. *Dysarthria Differential Diagnosis*. Mentor Seminars S.L.P., Rochester, MN, 1999.

[3] Xavier L. Aubert. An overview of decoding techniques for large vocabulary continuous speech recognition. *Computer Speech and Language*, 16:89–114, 2002.

[4] Sarah Borys and Mark Hasegawa-Johnson. Distinctive feature based SVM discriminant features for improvements to phone recognition on telephone band speech. In Review, 2005.

[5] Catherine P. Browman and Louis Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.

[6] David Caplan. *Language: Structure, Processing and Disorders*. MIT Press, Cambridge, MA, 1992.

[7] Hwa-Ping Chang. Speech input for dysarthric users. In *Meeting of the Acoustical Society of America*, page 2aSP7, Denver, CO, 1993.

[8] Hwa-Ping Chang. Speech recognition for dysarthric computer users. In *International Clinical Phonetics and Linguistics Association*, New Orleans, LA, 1994.

[9] Hwa-Ping Chang. *Speech input for dysarthric computer users*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.

[10] Sandra Chavarria, Taejin Yoon, Jennifer Cole, and Mark Hasegawa-Johnson. Acoustic differentiation of ip and IP boundary levels: Comparison of L- and L-L% in the switchboard corpus. In *Proceedings of the International Conference on Speech Prosody*, Nara, Japan, 2004.

[11] Fangxin Chen and Aleksandar Kostov. Optimization of dysarthric speech recognition. In *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1436–1439, 1997.

[12] Ken Chen and Mark Hasegawa-Johnson. HDK: Extensions of HTK for explicit-duration hidden markov modeling. Software available at http://www.ifp.uiuc.edu/speech/software, 2003.

[13] N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper and Row, New York, NY, 1968.

[14] Stephen Chu and Thomas S. Huang. Bimodal speech recognition using coupled hidden Markov models. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2000.

[15] Stephen M. Chu and Thomas S. Huang. Multi-modal sensory fusion with application to audio-visual speech recognition. In *Proceedings of the European Speech Technology Conference (EUROSPEECH)*, 2001.

[16] F. Darley and A. Aronson. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12:246–269, 1969.

[17] F. Darley and A. Aronson. *Motor Speech Disorders*. W.B. Saunders Co., Philadelphia, PA, 1975.

[18] Ameya Deoras and Mark Hasegawa-Johnson. Recognition of digits in music background using factorial hidden Markov model. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2004.

[19] Ameya Nitin Deoras and Mark Hasegawa-Johnson. A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[20] Philip C. Doyle, Herber A. Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, Charlene O'Neill, Marie-Claire Dylke, and Katherine Rolls. Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility. *J. Rehabilitation Research and Development*, 34:309–316, 1997.

[21] J. Duffy. *Motor Speech Disorders*. Mosby, St. Louis, 1995.

[22] Mital A. Gandhi and Mark A. Hasegawa-Johnson. Source separation using particle filters. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2004.

[23] Mark Hasegawa-Johnson. Bayesian learning for models of human speech perception. In *IEEE Workshop on Statistical Signal Processing*, St. Louis, 2003.

[24] Mark Hasegawa-Johnson. PVTK: Periodic vectors extraction and nonlinear transformation toolkit. Software available at http://www.clsp.jhu.edu/ws2004/groups/ws04ldmk/PVTK.php, 2004.

[25] Mark Hasegawa-Johnson, Abeer Alwan, Jul Cha, Shamala Pizza, and Katherine Haker. Vowels MRI database. Retrieved August 7, 2001 from University of Illinois at Urbana-Champaign, Image Formation and Processing Group Web site: http://www.ifp.uiuc.edu/speech/mri/index.html, 2001.

[26] Mark Hasegawa-Johnson, James Baker, Sarah Borys, Ken Chen, Emily Coogan, Steven Greenberg, Amit Juneja, Katrin Kirchhoff, Karen Livescu, Srividya Mohan, Jennifer Muller, Kemal Sönmez, and Tianyu Wang. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[27] Mark Hasegawa-Johnson, James Baker, Steven Greenberg, Katrin Kirchhoff, Jennifer Muller, Kemal Sönmez, Sarah Borys, Ken Chen, Amit Juneja, , Karen Livescu, Srividya Mohan, Emily Coogan, and Tianyu Wang. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. Technical Report WS04, Johns Hopkins University Center for Language and Speech Processing, 2005. http://www.clsp.jhu.edu/ws2004/groups/ws04ldmk/ws04ldmk_final.pdf.

[28] Mark Hasegawa-Johnson, Jul Setsu Cha, and Katherine Haker. CTMRedit: a matlab-based tool for segmenting and interpolating MRI and CT images in three orthogonal planes. In *21st Annual International Conference of the IEEE/EMBS Society*, page 1170, Atlanta, GA, Oct. 1999.

[29] Mark Hasegawa-Johnson, Shamala Pizza, Abeer Alwan, Jul Setsu Cha, and Katherine Haker. Tongue height and formants show speaker-independent vowel categories, but oral area does not. *J. Speech, Language, and Hearing Research*, 46(3):738–753, 2003.

[30] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am*, 87(4):1738–1752, 1990.

[31] Karen Hux, Joan Rankin-Erickson, Nancy Manasse, and Elizabeth Lauritzen. Accuracy of three speech recognition systems: Case study of dysarthric speech. *AAC: Augmentative and Alternative Communication*, 16(3):186–196, 2000.

[32] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, 1975.

[33] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ, third edition, 1992.

[34] Michael Kenstowicz. *Phonology in Generative Grammar*. Blackwell, Cambridge, Massachusetts, 1994.

[35] Ray D. Kent and Kristin Tjaden. Brain functions underlying speech. In John Laver, editor, *The Handbook of Phonetic Sciences*. Blackwell Publishers, Ltd., Oxford, 1997.

[36] Ray D. Kent, Gary Weismer, Jane F. Kent, Houri K. Vorperian, and Joseph R. Duffy. Acoustic studies of dysarthric speech: Methods, progress, and potential. *J. Commun. Disord.*, 32:141–186, 1999.

[37] Paul Kingsbury, Stephanie Strassel, Cynthia McLemore, and Robert MacIntyre. *LDC97L20: CALLHOME American English Lexicon (PRONLEX)*. Linguistic Data Consortium, Philadelphia, 1997.

[38] Ava-Lee Kotler and Nancy Thomas-Stonell. Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment. *AAC: Augmentative and Alternative Communication*, 13(2):71–80, 1997.

[39] Peter Ladefoged and Ian Maddieson. *The Sounds of the World's Languages*. Blackwell Publishers, Oxford, 1996.

[40] Bowon Lee and Mark Hasegawa-Johnson. Voice activity detection based on source location information using a linear microphone array in automobile environments. In Review, 2005.

[41] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. AVICAR: Audio-visual speech corpus in a car environment. In *INTERSPEECH International Conference on Spoken Language Processing*, 2004.

[42] Chin-Hui Lee, Chih-Heng Lin, and Bing-Hwang Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 39(4):806–814, 1991.

[43] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang*, 9:171–185, 1995.

[44] M. Cristina Leske. Prevalence estimates of communicative disorders in the U.S.: Speech disorders. *ASHA Leader*, 23(3), 1981.

[45] William J. M. Levelt. *Speaking: from Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

[46] Karen Livescu and James Glass. Feature-based pronunciation modeling for speech recognition. In *Human Language Technology: Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, 2004.

[47] Karen Livescu and James Glass. Feature-based pronunciation modeling with trainable asynchrony probabilities. In *ICSLP*, 2004.

[48] R.J. Love. *Childhood Motor Speech Disability*. Allyn and Bacon, Boston, 1992.

[49] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[50] Xavier Menendez-Pidal, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzio, and H.T. Bunnell. Nemours database of dysarthric speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1962–1965, 1996.

[51] G. A. Miller and P. E. Nicely. Analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am*, 27:338–352, 1955.

[52] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Hervé Glotin, Dimitra Vergyri, June Sison, Azad Mashari, and Jie Zhou. Audio-visual speech recognition: Final report. Technical Report WS00, Johns Hopkins University Center for Language and Speech Processing, 2000.

[53] J. J. Odell, P. C. Woodland, and S. J. Young. Tree-based state clustering for large vocabulary speech recognition. In *Proc. Internat. Sympos. Speech, Image Process. and Neural Networks*, pages 690–693, Hong Kong, 1994.

[54] M. Kamal Omar and Mark Hasegawa-Johnson. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[55] M. Kamal Omar and Mark Hasegawa-Johnson. Approximately independent factors of speech using non-linear symplectic transformation. *IEEE Transactions on Speech and Audio Processing*, 11(6):660–671, 2003.

[56] M. Kamal Omar and Mark Hasegawa-Johnson. Maximum conditional mutual information projection for speech recognition. In *Proceedings of the European Speech Technology Conference (EUROSPEECH)*, 2003.

[57] M. Kamal Omar and Mark Hasegawa-Johnson. Non-linear independent component analysis for speech recognition. In *Proc. Internat. Cybernetics, Control, and Communications Technol. Conference (CCCT)*, 2003.

[58] M. Kamal Omar and Mark Hasegawa-Johnson. Nonlinear maximum likelihood feature transformation for speech recognition. In *Proceedings of the European Speech Technology Conference (EUROSPEECH)*, 2003.

[59] M. Kamal Omar and Mark Hasegawa-Johnson. Model enforcement: A unified feature transformation framework for classification and recognition. *IEEE Transactions on Signal Processing*, 52(10), 2004.

[60] M. Kamal Omar, Mark Hasegawa-Johnson, and Stephen E. Levinson. Gaussian mixture models of phonetic boundaries for speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2001.

[61] Parimala Raghavendra, Elisabet Rosengren, and Sheri Hunnicutt. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *AAC: Augmentative and Alternative Communication*, 17(4):265–275, 2001.

[62] Yuexi Ren, Sung-Suk Kim, Mark Hasegawa-Johnson, and Jennifer Cole. Speaker-independent automatic detection of pitch accent. In *Proceedings of the International Conference on Speech Prosody*, Nara, Japan, 2004.

[63] Eric Sanders, Marina Ruiter, Lilian Beijer, and Helmer Strik. Automatic recognition of Dutch dysarthric speech: A pilot study. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2002.

[64] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 5(2):461–464, 1978.

[65] Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, and Anand Venkataraman. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, in press.

[66] K. N. Stevens. Evidence for the role of acoustic boundaries in the perception of speech sounds. In Victoria A. Fromkin, editor, *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pages 243–255. Academic Press, Orlando, Florida, 1985.

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

[68] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1999.

[69] Kenneth N. Stevens. Acoustic landmarks in speech perception. Presentation delivered at WS04, slides available at http://www.clsp.jhu.edu/ws2004/groups/ws04ldmk, 2004.

[70] A. Stolcke, H. Franco, R. Gadde, M. Graciarena, K. Precoda, M. Venkataraman, D. Vergyri, W. Wang, J. Zheng, Y. Huang, B. Peskin, I. Bulyko, M. Ostendorf, and K. Kirchhoff. Speech-to-text research at SRI-ICSI-UW. In *Spring 2003 EARS Workshop*, Boston, MA, 2003.

[71] Andreas Stolcke, Victor Abrash, Horacio Franco, Ramana Rao Gadde, Elizabeth Shriberg, Kemal Sönmez, Anand Venkataraman, Dimitra Vergyri, and Jing Zheng. The SRI march 2001 hub-5 conversational speech transcription system. In *NIST Workshop on Speech Transcription*, 2001.

[72] Hai Tao and Thomas Huang. Explanation-based facial motion tracking using a piecewise Bezier volume deformation model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999.

[73] Nancy Thomas-Stonell, Ava-Lee Kotler, Herbert A. Leeper, and Philip C. Doyle. Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition. *AAC: Augmentative and Alternative Communication*, 14(1):51–56, 1998.

[74] Jan van Santen. Applying speech / language technologies to communication disorders: New challenges for basic research. Presentation delivered at WS04, abstract available at http://www.clsp.jhu.edu/ws2004/seminars/lecture_santen.php, 2004.

[75] Gary Weismer. Motor speech disorders. In John Laver, editor, *The Handbook of Phonetic Sciences*. Blackwell Publishers, Ltd., Oxford, 1997.

[76] Steve Young, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book*. Cambridge University Engineering Department, Cambridge, UK, 2002.

[77] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson. An empathic tutoring system using spoken language. In *Proc. Australian Conf. Computer-Human Interaction (OZCHI)*, 2003.

[78] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson. Mental state detection of dialogue system users via spoken language. In *ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Kyoto, Japan, 2003.

[79] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson. Automatic detection of contrast for speech understanding. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, 2004.

[80] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson. Children's emotion recognition in an intelligent tutoring scenario. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, 2004.

[81] Tong Zhang, Mark Hasegawa-Johnson, and Stephen E. Levinson. Semantic analysis for a speech user interface in an intelligent-tutoring system. In *Intl. Conf. on Intelligent User Interfaces*, Madeira, Portugal, 2004.

[82] Weimo Zhu, Mark Hasegawa-Johnson, and Mital Arun Gandhi. Accuracy of voice-recognition technology in collecting behavior diary data. In *Association of Test Publishers (ATP): Innovations in Testing*, 2005.

[83] V.W. Zue, S. Seneff, and J. Glass. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9:351–356, 1990.