HMM-BASED AND SVM-BASED RECOGNITION OF THE SPEECH OF TALKERS WITH SPASTIC DYSARTHRIA

Mark Hasegawa-Johnson, Jonathan Gunderson, Adrienne Perlman, Thomas Huang

{jhasegaw,jongund,aperlman,t-huang1}@uiuc.edu

ABSTRACT

This paper studies the speech of three talkers with spastic dysarthria caused by cerebral palsy. All three subjects share the symptom of low intelligibility, but causes differ. First, all subjects tend to reduce or delete word-initial consonants; one subject deletes all consonants. Second, one subject exhibits a painstaking stutter. Two algorithms were used to develop automatic isolated digit recognition systems for these subjects. HMM-based recognition was successful for two subjects, but failed for the subject who deletes all consonants. Conversely, digit recognition experiments assuming a fixed word length (using SVMs) were successful for two subjects, but failed for the subject with the stutter.

1. MOTIVATION AND BACKGROUND

Speech and language disorders result from many types of congenital or traumatic disorders of the brain, nerves, and muscles [1]. 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 [1]. The most common category of dysarthria among children and young adults is spastic dysarthria [2], typically characterized by strained phonation, imprecise placement of the articulators, incomplete consonant closure, 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 [2]. Spastic dysarthria is associated with a variety of disabilities such as, but not limited to, cerebral palsy and traumatic brain injury [1]. Adults with cerebral palsy are able to perform most of the tasks required of a college student, including reading, listening, 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.

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 ASR is faster than interaction using a keyboard [3, 4, 5]. With few exceptions, the technology used in these studies is speaker - dependent or speaker-adaptive commercial off-the-shelf speech recognition technology. Raghavendra et al. [6] 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 a speaker-dependent ASR, however, 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. [7] 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.

This paper studies the speech of three talkers with spastic dysarthria caused by cerebral palsy, and one control subject, all recorded using an array of seven microphones. All three subjects share the symptom of low intelligibility, but the causes of their low intelligibility are subject-dependent. Based on phonological analysis, we conclude that the speech samples of these three subjects differ in primarily two respects. First, all subjects tend to reduce or delete word-initial consonants, resulting in low intelligibility. This tendency is especially strong in one subject, who tends to delete all consonants in the word; this tendency reduces her intelligibility relative to her peers. Second, in addition to his tendency to reduce word-initial consonants, one subject also exhibits a slow stutter, perceived by most listeners as an indeterminate number of syllables per word.

Two algorithms were used to develop automatic isolated digit recognition systems for these subjects. Digit recognition using hidden Markov models (HMMs) was successful for two subjects, but failed for the subject with the most pronounced tendency to reduce or delete consonants. Conversely, digit recognition experiments assuming a fixed word length (using support vector machine or SVM classifiers) were successful for two subjects, but failed for the subject with the stutter. From these results, we tentatively conclude that the dynamic time warping features of the HMM give it some degree of robustness against large-scale word-length fluctuations, while the regularized discriminative error metric used to train the SVM gives it some degree of robustness against the reduction and deletion of consonants.

Table 1. 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%

2. DATA ACQUISITION AND CHARACTERIZATION

Data were recorded from three subjects with spastic dysarthria: two male (M01, M03), and one female (F01). Data were also recorded from one control subject with no perceptible speech pathology (M02). Subjects were recorded using an array of eight microphones and four cameras [8]; of the recorded signals, seven microphone signals were used in the experiments reported here. Cameras and microphones were mounted on top of a computer monitor. One-word prompts were displayed on the monitor using PowerPoint. Subjects with spastic dysarthria were unable to easily control a keyboard or mouse, therefore an experimenter sat next to the monitor, advancing the PowerPoint slides after each word spoken. 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). Each subject recorded a total of 541 words, including 395 distinct words. All recordings are available upon request.

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 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 first author of this paper. Other listeners (L2 and L3) are students of speech and language technology. 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 1. Several findings are apparent. First, the control subject (M02) is much more intelligible than the other talkers. 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 intelligibility 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 2. 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 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 2 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, because they lacked any discernable consonants. 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, frequently deleted word-initial consonants. Across all speakers, word-initial and word-final consonant errors were more frequent than word-medial consonant and vowel errors.

Table 2. 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.		
	DEL	INS	SUB	DEL	INS	SUB
All	37	6	83	16	7	63
F01	8	2	25	2	2	20
M01	3	0	40	11	4	20
M02	2	2	3	0	0	1
M03	24	2	15	3	1	22

	Final Cons.			Vowel		Word
	DEL	INS	SUB	SUB	NS	WD
All	45	32	64	74	29	87
F01	15	5	15	22	5	46
M01	18	17	19	34	14	27
M02	0	2	0	1	0	0
M03	12	8	30	17	10	14

3. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition experiments were conducted using two recognition paradigms. In the first two experiments, recognition was performed using speaker-dependent phone-based hidden Markov models (HMM). In the third and fourth experiments, isolated word classification was performed using speaker-dependent support vector machines (SVM).

Using the HTK toolkit, speaker-dependent speech recognizers were trained and tested. All systems used a relatively standard HMM architecture: monophone or clustered triphone HMMs, three states per phone, mixture Gaussian observation PDFs, PLP+energy+d+dd spectral observations. 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 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 second 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 3. In Table 3, 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 simple delay-and-sum beamformer; more advanced beamformers were not tested. 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.

Table 3. 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	45 Words		10 Words (Digits)				
Algorithm	H	HV	H	HV	Word	WF	WFV
F01	44	55	71	80	97	86	90
M01	42	49	86	95	70	69	70
M02	87	89	99	100	90	90	90
M03	77	80	99	100	97	100	100

The third and fourth experiments tested support vector machines (SVMs) for fixed-length isolated word recognition. The start and end times of each word were first detected using seven independent single-channel Gaussian voice activity detectors (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 64 consecutive 10ms PLP frames, beginning at the detected word onset time, in order to construct a superframe observation.

Two types of SVM were trained: 10-ary Word-SVMs, and binary Word-Feature-SVMs (WF-SVMs; Table 3). Word-SVMs were trained using two examples of each digit, while the third example was used for testing. Word-Feature-SVMs (WF-SVMs) were a bank of 170 different binary-output SVMs, trained and tested with 17 different binary target functions, and with 10 different types of superframe observation (different lengths, anchored with respect to the word onset, word offset, or energy peak of the word). 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 3 in columns “WF” (all microphones scored separately) and “WFV” (microphones vote to determine final system output).

4. CONCLUSIONS

This paper has demonstrated automatic isolated digit recognition for talkers with very low intelligibility, caused by a variety of symptoms related to spastic dysarthria. HMM-based digit recognition was successful for two subjects, but failed for the subject with the most pronounced tendency to reduce or delete all of the consonants in a word. Conversely, digit recognition experiments assuming a fixed word length (using SVM classifiers) were successful for two subjects, but failed for the subject with a slow, deliberate stutter. From these results, we tentatively conclude that the dynamic time warping features of the HMM give it some degree of robustness against large-scale word-length fluctuations, while the regularized discriminative error metric used to train the SVM gives it some degree of robustness against the reduction and deletion of consonants.

A 10-word vocabulary is not sufficient for meaningful human-computer interface. In future work, we hope to extend this research to develop working speech recognizers with vocabulary sizes of several dozen words, including computer control commands and letters of the International Radio Alphabet. We intend to pursue at least two methods to achieve this goal. First, we intend to ask subjects to record three or more examples of each word in a larger vocabulary, so that we can develop whole-word isolated-word speech recognition models (HMM, SVM, hybrid SVM-HMM, or other models). Second, we intend to pursue methods that are theoretically capable of generalizing from a training vocabulary to a novel test vocabulary, including phone-based HMM recognizers and distinctive-feature-based SVM recognizers.

5. REFERENCES

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

[2] RJ Love, *Childhood Motor Speech Disability*, Allyn and Bacon, Boston, 1992.

[3] H-P. Chang, “Speech input for dysarthric users,” in *Meeting of the Acoustical Society of America*, Denver, CO, 1993, p. 2aSP7.

[4] N Thomas-Stonell, A-L Kotler, HA Leeper, and PC Doyle, “Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition,” *AAC: Augmentative and Alternative Communication*, vol. 14, no. 1, pp. 51–56, 1998.

[5] K Hux, J Rankin-Erickson, N Manasse, and E Lauritzen, “Accuracy of three speech recognition systems: Case study of dysarthric speech,” *AAC: Augmentative and Alternative Communication*, vol. 16, no. 3, pp. 186–196, 2000.

[6] P Raghavendra, E Rosengren, and S Hunnicutt, “An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems,” *AAC: Augmentative and Alternative Communication*, vol. 17, no. 4, pp. 265–275, 2001.

[7] PC Doyle, HA Leeper, A-L Kotler, N Thomas-Stonell, C O’Neill, M-C Dylke, and K Rolls, “Dysarthric speech: a comparison of computerized speechrecognition and listener intelligibility,” *J. Rehabilitation Research and Development*, vol. 34, pp. 309–316, 1997.

[8] B Lee, M Hasegawa-Johnson, C Goudeseune, S Kamdar, S Borys, M Liu, and T Huang, “Avicar: Audio-visual speech corpus in a car environment,” in *Proc. Internat. Conf. Spoken Language Processing*, 2004.