The Effect of Visual Information on Word Initial Consonant Perception of Dysarthric Speech

Richard P. Schumeyer
schumeye@asel.udel.edu

Kenneth E. Barner
barner@asel.udel.edu

Applied Science and Engineering Laboratories
University of Delaware/A.I. duPont Institute
P.O. Box 269
1600 Rockland Road
Wilmington, DE 19899

ABSTRACT

Disabled individuals will realize many benefits from automatic speech recognition. To date, most automatic speech recognition research has focused on normal speech. However, many individuals with physical disabilities also exhibit speech disorders. While limited research has been conducted focusing on dysarthric speech recognition, the preliminary results indicate that additional study is necessary. Recently, increasing attention has been given to multimodal speech recognition schemes that utilize multiple input sources - most commonly audio and video. This multimodal approach has been applied to normal speech with demonstrated effectiveness. Through studying the effect of audio and visual information in a human perception experiment, this study attempts to discover whether such an approach would be useful for dysarthric speech recognition. Results of a closed vocabulary perception test are presented. In this test, 15 normal hearing viewers were presented with videotapes of three dysarthric speakers speaking a series of one syllable nonsense words. These words differed only in the initial consonant. The words were presented in both audio-only and audio-visual modes. Perception rates in both modes were measured. The results are analyzed and compared to other studies of visual speech perception and dysarthric speech articulation.

1. INTRODUCTION

Automatic speech recognition promises to dramatically improve the quality of life for individuals with physical disabilities [9]. It will allow operation of devices such as telephones, lamps, and doors without requiring the user to touch a device. Current speech recognition technology requires consistent speech to be effective. Unfortunately, many disabled individuals also exhibit speech disorders. While some studies have investigated dysarthric speech recognition[4, 6, 7], further research in this area is required.

Most current speech recognizers are audio-only. That is, the input to the system consists only of the audio speech signal. Several researchers have recently investigated multimodal speech recognition schemes. In their research, the audio signal was supplemented by a video signal of the speaker's face. Features from both the audio and video signals were used by the recognizer. The basis of this research was the knowledge that lipreading aids human perception of speech, even for normal hearing listeners [1, 5]. Researchers have found that the addition of visual information to an automatic speech recognition system improves the recognition rates of normal speakers, especially when the audio signal is degraded by noise [2, 11, 12].

It is unknown whether lipreading information is reliable with dysarthric speech. The purpose of this study is to determine whether the addition of visual information improves perception of Cerebral Palsy (CP) dysarthric speech. The results will help determine whether visual information would also be helpful for an automatic speech recognizer. We performed an audio-visual perception experiment involving dysarthric speakers and normal hearing viewers. Dysarthric speakers were recorded while speaking words which differed only in the initial consonant. For each word, the viewers indicated which consonant they perceived in the initial position.

This study focused solely on differentiation of initial consonants. This has precedent in many previous lipreading studies which examined only one class of sound. While data on final consonants or vowels would also be useful, it would have required collecting substantially more samples from the speaking subjects. Although collecting the data took only about 15 minutes for each speaker, each appeared to be fatigued from the effort. Even though vowels and final consonants were not studied, conclusions about lipreading can be drawn from the initial consonant test.

The remainder of this paper is organized as follows: Section 2 describes both speaking and listening subjects, and recording and testing procedures. Perception rates and confusion matrices are presented in section 3, and are further discussed in section 4. Conclusions and suggestions for future research are in section 5.

2. METHOD

2.1. Speaking Subjects

The speaking subjects were three clean-shaven adult male subjects with Cerebral Palsy. The type of CP and seriousness of speech disorder for each subject, as reported by the subject's speech therapist, are given in Table 1.

Subject	Age	CP type	Severity of Dysarthria
1	54	spastic quadriplegic	moderate to severe
2	34	athetoid	moderate
3	42	spastic dyplegic	moderate

Table 1: Type of CP, age, and severity of dysarthria for the three speaking subjects.

2.2. Viewing Subjects

The 15 viewing subjects consisted of 8 males and 7 females. They ranged in age from 25 to 55 years. Each described himself as having normal hearing and vision, although no tests were performed to verify this. They had varying degrees of familiarity with dysarthric speech among the viewers; however none of the viewers knew any of the speakers.

2.3. Stimulus Material

The stimulus material in this experiment was similar to that of [1], enabling a comparison of results between the two studies. It consists of one of sixteen consonants followed by /ɑd/. The sixteen consonants were /b/, /d/, /f/, /ɡ/, /k/, /m/, /n/, /p/, /s/, /ʃ/, /t/, /θ/, /ð/, /v/, /z/, and /ʒ/.

2.4. Recording Procedure

Each speaker was videotaped against a black background in a sound proof booth. The video camera's zoom lens was adjusted so that the speaker's entire head was visible. The camera's video output was connected to a Hi-8 VCR for recording. A separate microphone (ElectroVoice RE-55) was connected to the audio input of the VCR to record the sound.

Three lists of the 16 consonants were made, each in a different random order. Each speaker spoke the words on all three lists, resulting in 48 tokens for each. The collection of 48 tokens was called a *set*.

2.5. Editing Procedure

The test required two sets of the 48 tokens (one for audio, another for audio-visual). To create the first set, each token from the original tapes was copied one token at a time onto a VHS deck. The second set was created the same way, except with a different order. The final tape now contained two complete sets of tokens, each in a different order.

2.6. Test Procedure

Each viewer was presented all of the material. One set for each speaker was presented audio-only, and the other was presented audio-visual. Both portions of the test are on videotape; the only difference is that the monitor is on during the audio-visual portion, and off during the audio-only portion.

	Audio		AudVid		
Speaker	Correct	SD	Correct	SD	$AV - A$
1	27.4	8.3	33.5	6.3	6.1
2	23.3	8.3	19.2	5.3	-4.2
3	36.4	6.8	43.3	4.9	6.9
Total	29.0	5.5	32.0	3.8	3.0

Table 2: Percent correct and standard deviations for all speakers and modes. $AV - A$ is the percentage improvement for AudVid.

Several steps were taken to remove biases from the data:

- The order in which the speakers were presented was varied.
- The order in which the modes were presented was varied.
- The sets used for each mode were varied.
- The order of words within each set differed.

While viewing the tapes, each viewer indicated which of the sixteen consonants they perceived in the initial position. Each viewer was instructed not to leave any items blank, and to guess if unsure which consonant was spoken. Nevertheless, some items were left blank. Each viewer took the test in a quiet room, listening through Sony MDR 65 headphones. The audio-visual portion was shown on a 20 inch monitor.

3. RESULTS

The percentage of correct responses for each speaker in both modes is given in Table 2. Results of examining several effects with an analysis of variance are presented in Table 3. The data indicate that perception results are highly dependent on the speaker. The perception rates in the audio condition range from 23% to 36%, and in the audio-visual case from 19% to 43%. According to an analysis of variance, the difference in perception rates between speakers is significant. Although the overall increase in perception in the audio-visual mode is small (3%), the improvement is statistically significant. The reduction in standard deviation from 5.5% to 3.8% indicates that here was less variation in the scores between viewers in the audio-visual mode.

The analysis of variance also confirmed that there was interaction between the speakers and the different modes. This interaction was demonstrated by the fact that the perception accuracy for speaker 2 declined in the audio-visual mode, while it improved for speakers 1 and 3. Table 3 also indicates that the mode order was irrelevant.

The data were tabulated in confusion matrices, with phonemes grouped into homophenes (Table 4). The audio and audio-visual confusion matrices are in Tables 5 and 6, respectively. These matrices indicate the percentage that a group n phoneme was perceived as a group n phoneme; this does not imply that the phoneme was perceived correctly. The rows do not always sum to 100% because some viewers left some answers blank. The number of items perceived in the correct group was slightly improved in the audio-visual mode.

Effect	DF	F	p
speaker	[2,28]	56.28	0.001
mode	[1,14]	5.77	0.032
speaker*mode	[2,28]	9.51	0.001
order	[1,13]	0.18	0.675
speaker*order	[2,26]	1.14	0.336
mode*order	[1,13]	1.24	0.286

Table 3: Analysis of variance results. "speaker" refers to the effect of different speakers; "mode" to difference between audio and audio-visual modes; "order" to the difference between presenting audio first or audio-visual first. "*" denotes interaction.

I	/p, b, m/
II	/f, v/
III	/θ, ð/
IV	/ʃ, ʒ/
V	/t, d, n, s, z, k, g/

Table 4: Phoneme groups used for confusion matrices.

		Perceived				
		I	II	III	IV	V
Intended	I	58.0	19.5	3.7	1.2	17.5
	II	25.9	40.0	13.3	1.9	18.9
	III	10.0	11.5	28.5	8.5	40.7
	IV	1.5	3.3	28.5	21.1	45.2
	V	7.9	4.8	11.6	9.0	66.2

Table 5: Audio confusion matrix, with visually similar phonemes grouped together.

		Perceived				
		I	II	III	IV	V
Intended	I	73.1	17.8	0.7	0.0	7.2
	II	25.6	50.0	13.3	0.7	10.4
	III	8.9	13.3	36.7	5.2	35.6
	IV	2.6	10.0	21.9	22.2	43.3
	V	8.8	5.7	11.4	6.7	66.5

Table 6: Audio-visual confusion matrix, with visually similar phonemes grouped together.

Phonemes were also grouped by manner of articulation, and the percentage of correct responses for each group was found (Table 7). In contrast to the previous tables, the entries in Table 7 indicate percentage of correct responses.

The perception consistency was also measured; the results are in Table 8. A token was perceived consistently if a viewer gave the same response, even if it was incorrect, to both the audio and audio-visual presentations of the token. Table 8 confirms that for speakers 1 and 3, when a viewer correctly perceived a token presented audio-only, the same token was usually correctly perceived when presented audio-visual. This implies that the improvement in the AV mode results from correctly perceiving the tokens which were correct in the A mode, and then using the visual information to correctly perceive additional tokens.

Manner	Audio	AudVid
Nasal	57.4	57.8
Fricative	18.1	19.4
Stop	34.1	40.1

Table 7: Percentage of correct responses when phonemes are grouped by manner of articulation.

Speaker	Pct.	$AV\|A$
1	42.9	76.6
2	30.8	46.0
3	52.6	73.9

Table 8: Percentage of consistent responses. $AV|A$ is the percentage of correct AudVid responses, given that the same token was perceived correctly in Audio.

4. DISCUSSION

Both audio and audio-visual perception rates in this study are lower than the rates reported by Binnie [1], who recorded one normal speaker speaking the same set of consonants. Binnie's viewers also attained greater improvement with the audio-visual mode when the audio was degraded by noise. In their noisiest test condition, the audio recognition rate was 6.1%, and the audio-visual rate was 47.7%, an improvement of 41.4%. This compares with an improvement of less than 7% for the most improved speaker in this study.

Lipreading effectiveness is highly speaker dependent for normal speakers, since many speakers achieve correct articulation with only slight mouth and lip movements, making lipreading less effective. This study shows that the successful lipreading of dysarthric speech is also speaker dependent. In fact, in some cases lipreading may hinder correct perception. For instance, 9 of the 15 viewers perceived speaker 2 *worse* in the audio-visual mode. Many viewers reported the video to be distracting for this speaker. This seems to confirm the anecdotal evidence that many listeners understand dysarthric speech better on the telephone than in person, or they do not look at the speaker while listening.

There are differences between lipreading a normal speaker and a dysarthric speaker which may explain the small gains from visual information. Both normal hearing and hearing impaired listeners employ lipreading to improve aural perception of normal speakers. In either case, the lip movements correspond to the speaker's clear articulation. With a dysarthric speaker, many phonemes are articulated inaccurately. The lip movements correspond to the imprecise articulation and are therefore misleading.

Speaker 2 had the lowest audio recognition rates. Speaker 2 was the only athetoid subject; the decreased perception of his speech corresponds to results found in [10], in which it was reported that intelligibility of individuals with spastic CP was superior to those with athetoid CP.

A comparison of the audio and audio-visual confusion matrices (Tables 5 and 6), in which homophenes are grouped together, reveals that the greatest improvement occurs in groups I, II, and III. For example, 58% of the group I phonemes were perceived as group I during the audio portion; this improves to 73% in the audio-visual segment. These three phoneme groups are labeled “visible” by Jeffers in [8], meaning that the movements associated with these phonemes are relatively easy to see. Groups IV and V are “obscure”, which explains the lesser degree of improvement in the audio-visual mode.

When the phonemes are grouped by manner of articulation, the nasals are most accurately perceived while the fricatives are the least accurately perceived. These results also agree with [10] which found greater articulatory accuracy for nasals than for stops and fricatives.

5. CONCLUSIONS

The goal of this study was to determine the usefulness of adding video to an audio speech recognizer for CP speech. The improvement in the audio-visual mode, while statistically significant, was small and highly variable, especially in comparison to normal speech. It does not appear great enough to warrant the additional expense of the video mode. On the other hand, studies have found that an automatic system can outperform human listeners if the speech is consistent and distinct [3]. If the visual information is consistent, perhaps an automatic system could make more use of the information than a human listener.

While this approach does not look promising for dysarthric speech, it may be useful for deaf speech. An individual with CP is physically incapable of moving their articulators clearly and accurately; this is not true of deaf individuals. While a deaf speaker obviously cannot hear how a word is pronounced, he can see how it is articulated. A study similar to the present one would reveal whether the visual content of deaf speech would yield perceptual improvement.

6. ACKNOWLEDGMENTS

Funding for this project was provided by the National Science Foundation, grant #: HRD-9450019. Additional support has been provided by the Nemours Research Programs.

I would like to thank Dr. H. Timothy Bunnell for his assistance with experiment design and data analysis.

7. REFERENCES

1. C. A. Binnie, A. A. Montgomery, and P. L. Jackson. Auditory and visual contributions to the perception of consonants. *J. Speech Hear. Res.*, 17:619–630, 1974.

2. C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *Proc. ICASSP 93*, volume 1, pages 557–560, 1993.

3. G. S. Carlson and J. Bernstein. Speech recognition of impaired speech. In *Proc. 10th Annl. Conf. Rehabil. Tech. 1987*, pages 103–105, San Jose, 1987.

4. J. Deller Jr., D. Hsu, and L. J. Ferrier. Encouraging results in the automated recognition of cerebral palsy speech. *IEEE Trans Biomed Eng*, 35(3):218–220, Mar. 1988.

5. N. P. Erber. Auditory-visual perception of speech. *J. Speech Hearing Dis.*, 40:481–492, 1975.

6. L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit. Dysarthric speaker's intelligibility and speech characteristics in relation to computer speech recognition. *Augment. Altern. Comm.*, 11:165–173, Sept. 1995.

7. M. Fried-Oken. Voice recognition device as a computer interface for motor and speech impaired people. *Arch Phys Med Rehabil*, 66:678–681, Oct. 1985.

8. J. Jeffers and M. Barley. *Speechreading (Lipreading)*. Charles C. Thomas, Springfield, Illinois, 1971.

9. J. M. Noyes and C. R. Frankish. Speech recognition technology for individuals with disabilities. *Augment. Altern. Comm.*, 8:297–303, Dec. 1992.

10. L. J. Platt, G. Andrews, M. Young, and P. T. Quinn. Dysarthria of adult cerebral palsy: I. intelligibility and articulatory impairment. *J. Speech Hear. Res.*, 23:28–40, Mar. 1980.

11. D. G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *Proc. IJCNN*, volume 2, pages 286–295, 1992.

12. B. P. Yuhas, M. H. Goldstein Jr., T. J. Sejnowski, and R. E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proc. IEEE*, 78(10):1658–1667, Oct. 1990.