INTERSPEECH 2009 BRIGHTON

Universal Access: Speech Recognition for Talkers with Spastic Dysarthria

Harsh Vardhan Sharma[1,2], *Mark Hasegawa–Johnson*[1,2]

[1]Beckman Institute for Advanced Science and Technology, Urbana, USA
[2]Department of Electrical and Computer Engineering, University of Illinois, Urbana, USA
{hsharma,jhasegaw}@illinois.edu

10.21437/Interspeech.2009-444

Abstract

This paper describes the results of our experiments in small and medium vocabulary dysarthric speech recognition, using the database being recorded by our group under the Universal Access initiative. We develop and test speaker-dependent, word- and phone-level speech recognizers utilizing the hidden Markov Model architecture; the models are trained exclusively on dysarthric speech produced by individuals diagnosed with cerebral palsy. The experiments indicate that (a) different system configurations (being word vs. phone based, number of states per HMM, number of Gaussian components per state specific observation probability density etc.) give useful performance (in terms of recognition accuracy) for different speakers and different task-vocabularies, and (b) for very low intelligibility subjects, speech recognition outperforms human listeners on recognizing dysarthric speech.

Index Terms: speech recognition, dysarthria, cerebral palsy, human-computer interface, assistive technology, augmentative communication

1. Introduction

Automatic speech recognition (ASR) software with high word recognition accuracy is now widely available to the general public; accuracy of the newest generation of large vocabulary speech recognizers, after adaptation to a user without speech pathology, typically exceeds 95% (Dragon claims a 99% accuracy for Dragon Naturally Speaking version 9 [1]). It has been reasonably successful at providing a useful human-computer interface especially for people who find it difficult to type with a keyboard (e.g., patients with carpal tunnel syndrome).

However, many individuals with gross motor impairment, including some people with cerebral palsy and closed head injuries, have not enjoyed the benefit of these advances in speech technology, mainly because their general motor impairment includes a component of *dysarthria*: reduced speech intelligibility caused by neuromotor impairment. Such people find their participation in society limited by their inability to use a personal computer, and it is these aforementioned motor impairments that 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 ASR system, with code words representing letters and formatting commands, and with acoustic speech recognition carefully adapted to the speech of the individual user (e.g., see [2],[3]).

We are currently engaged in acquiring and experimenting with a database of dysarthric speech, with an aim to develop dysarthric ASR systems and a corresponding human-computer interface (see [4]) for use by students with dysarthria, at the University of Illinois. The work described in this paper has investigated the performance (in terms of the recognition accuracy) on a part of this database, of word– and phone-based audio speech recognition models employing the hidden Markov Model (HMM) architecture, for both small– and medium-size vocabularies, designed to be used for unrestricted text entry on a personal computer.

2. Background & Motivation

2.1. Prevalence of dysarthria

Speech and language disorders are caused by various types of congenital or traumatic disorders of the brain, nerves and/or muscles [5]. Dysarthria is a collective term used for referring to a group of motor speech disorders resulting from disturbed muscular control of the speech mechanism due to damage of the peripheral or central nervous system. People suffering with one of the "dysarthrias" exhibit oral communication problems due to weakness, incoordination or paralysis of speech musculature. The physiologic characteristics of dysarthria include abnormal/disturbed strength, speed, range, steadiness, tone and/or accuracy of muscle movements. The communication characteristics include disturbed pitch, loudness, voice quality, resonance, respiratory support for speech, and articulation. For details regarding etiology, assessment and treatment of dysarthria, refer to [6].

2.2. ASR for dysarthria: Motivation and Present state of art

Although dysarthria can differ notably from normal speech due to imprecise articulation, the articulation errors are generally neither random (unlike, for example, in the case of apraxia) nor unpredictable. In fact, previous studies show that most articulation errors in dysarthria can be described in terms of a small number of substitution error types [7],[8]. Kent et al. [7], for example, suggest that most articulation errors in dysarthric speech are primarily errors in the production of one distinctive feature. When articulation errors occur in a consistent manner and, as a result, they are predictable, *there exists the possibility of using ASR*, even for speech that is highly unintelligible for human listeners. Several studies have repeatedly demonstrated that adults with dysarthria are capable of using ASR, and that in some cases, human-computer interaction using speech recognition is faster and less tiring than interaction using a keyboard ([2],[3]).

Speaking for long periods of time is tiring, especially for a person with dysarthria; therefore it is difficult for a person with dysarthria to train a speaker-dependent ASR. Unfortunately, speaker-independent and speaker-adaptive recognizers, of the kind used by talkers with no pathology, are of less use to talkers with dysarthria, because the substitution errors characteristic of

dysarthria dramatically increase word error rates. Raghavendra et al. [9], for example, 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 talkers 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.

The technology used in these studies is commercial off–the–shelf speech recognition technology. Further, these studies have focused on small-vocabulary applications, with vocabulary sizes ranging from ten to seventy words. 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. The first publicly available database suitable for training medium-vocabulary automatic dysarthric speech recognition for talkers with high, moderate, low, or very low intelligibility is the UA-Speech database, reported in [10].

3. Experiments

3.1. Data used

The experiments described in this paper utilized speech of 7 subjects from the UA-Speech database [10]. This corpus was constructed with the aim of developing large-vocabulary dysarthric ASR systems which would allow users to enter unlimited text into a computer. All subjects exhibited symptoms of spastic dysarthria, according to an informal evaluation by a certified speech-language pathologist. Each subject recorded 765 isolated words in 3 blocks of 255 words each; (a) common to all blocks: 10 digits (D), 19 computer commands (C), 26 radio alphabet letters (L), and 100 common words (CW) selected from the Brown corpus of written English; and (b) unique to each block: 100 uncommon words (UW) selected from children's novels digitized by Project Gutenberg. Vocabularies D and CW were primarily composed of monosyllables, C and L of bisyllables, and UW of polysyllabic words. The subjects' speech was affected by dysarthria associated with cerebral palsy. Kim et al. [10] describe in detail, the acquisition of and intelligibility assessment on this database. Two hundred distinct words were selected from the recording of the second block: 10 digits, 25 radio alphabet letters, 19 computer commands and, 73 words randomly selected from each of the CW and UW categories. Five naive listeners were recruited for each speaker and were instructed to provide orthographic transcriptions of each word that they thought the speaker said. The percentage of correct responses was then averaged across five listeners to obtain each speakers intelligibility. Table 1 lists the subjects whose speech materials from the UA database were used, along with their human listener intelligibility ratings. The first letter of the subject code ('M' or 'F') indicates their gender.

3.2. ASR tasks and task–vocabularies

Ten recognition tasks were set up using the recorded data, as summarized in Table 2.

The measure used for assessing the performance of the developed recognizers is the fraction of task–vocabulary words correctly recognized (in percent), defined in Equation 1.

$$PWC = \frac{\#\ words\ correctly\ recognized}{vocabulary\ size(\#\ words)} \times 100 \quad (1)$$

Table 1: *Summary of Speaker Information (in decreasing order of human listener intelligibility rating).*

Speaker	Age	Speech Intelligibility (%)
M09	18	high (86%)
M05	21	mid (58%)
M06	18	low (39%)
F02	30	low (29%)
M07	58	low (28%)
F03	51	very low (6%)
M04	>18	very low (2%)

Table 2: *ASR Recognition Tasks and corresponding Vocabulary Sizes*

Task	Vocabulary	Vocabulary Size
T01	D	10
T02	C	19
T03	L	26
T04	D+L+C	55
T05	CW	100
T06	UW	100
T07	L+C+CW	145
T08	D+L+C+CW	155
T09	L+C+CW+UW	245
T10	D+L+C+CW+UW	255

3.3. Architecture

HMM-based speech recognizers employing three configurations were developed and tested: whole-word, monophone and triphone (word-internal, context-dependent). Blocks 1 and 3 were used for training and block 2 for testing, for each speaker–task combination. Referring to Table 2, word-level recognizers were built for tasks T01-T05, T07 and T08 (the training vocabulary was thus exactly twice the size of test vocabulary). The other tasks had uncommon words as part of their vocabularies, and since each block's uncommon words were unique to it, therefore could not be modeled at the word level. For tasks T06-T10, phone-level (both monophone and triphone) recognizers were built. Hence, tasks T07 and T08 are the ones for which all three configurations were tested.

The features extracted from the speech waveform comprised of 12 Perceptual Linear Prediction coefficients [11] for 25 ms Hamming-windowed segments obtained every 10 ms, plus the energy of the windowed segment. 'Velocity' and 'Acceleration' components were also calculated for this 13-dimensional feature, which finally resulted in a 39-dimensional acoustic feature vector.

For each configuration-task combination, the number of Gaussian components in the state-specific observation probability densities was increased (in an iterative manner) in powers of 2, starting from 1 and stopping when either (a) the number of components had risen to 32, or (b) the PWC score had decreased on two consecutive iterations. The number of states per HMM was fixed at 3 for monophone and triphone systems, but was varied from 3 through 9 for the whole-word systems. The scores reported are for a particular HMM configuration (in terms of number of states per HMM and number of Gaussian probability density components) because the stopping stage of the above-mentioned iterative process is likely to be different for different

task and subject combinations. Standard methods for choosing HMM comfiguration (using development test data) could not be employed on account of insufficient data. The results reported in the next section should therefore be interpreted as development test results. In order to avoid over-tuning, the HMM configuration was constrained to be the same across all speakers (and if possible across all tasks, especially for whole-word systems where the number of states per HMM was also varied). For the whole-word systems, results are for HMMs with 2 Gaussian components per probability density and 6 states per HMM. The decision to report only 6-state results is a heuristic attempt to improve the generalizability of these results. For the monophone and triphone systems, results are for HMMs with 16 and 2 Gaussian components per probability density, respectively.

4. Results

Tables 3-6 list the PWC scores for whole-word, monophone and triphone systems respectively. The subjects are listed in decreasing order of intelligibility rating.

Table 3: *PWC for whole-word systems: tasks T01-T04. Intelligibility of each talker is given in the second column.*

		ASR Task			
Talker	Intel.	T01	T02	T03	T04
M09	86	84.29	100	97.25	89.87
M05	58	90	78.95	77.47	63.12
M06	39	92.86	81.95	77.47	72.21
F02	29	94.29	83.46	69.78	72.99
M07	28	100	86.47	85.71	80.78
F03	6	74.29	63.91	41.76	40.26
M04	2	46	23.16	19.23	14.18

Table 4: *PWC for whole-word systems: tasks T05,T07,T08. Intelligibility of each talker is given in the second column.*

		ASR Task		
Talker	Intel.	T05	T07	T08
M09	86	63.29	69.26	65.9
M05	58	56.14	54.68	52.9
M06	39	52.86	53	51.24
F02	29	64.43	58.72	57.24
M07	28	56.14	58.92	58.89
F03	6	49.43	36.35	33.82
M04	2	6.4	7.03	6.84

whole-word ASR: For all subjects, recognition accuracy deteriorates with increase in vocabulary size. However, for speakers with low and very low intelligibility (all except M09 and M05), *recognition accuracy is higher than their respective intelligibility ratings* (the magnitude of difference is larger for small vocabularies than for medium sized ones). For M09 and M05, the recognition accuracy is higher than their intelligibility ratings for small sized vocabularies (tasks T01-T04) but not the medium sized ones (tasks T05, T07, T08).

monophone ASR: For all subjects, ASR accuracy on task T06 (uncommon words only) was always worse than their respective intelligibility ratings. ASR was less accurate on T06 (100 polysyllabic uncommon words) than on any task containing monosyllables (including those with twice the vocabulary size). For all subjects, the recognition scores for tasks T07 and T08 were respectively higher than those on T09 and T10 (which are T07 and T08 with the uncommon words added). Comparing T09 and T10 scores, it appears that incorporating digits into the vocabulary improves the PWC score by 4-7% absolute, for all speakers except F03 and M04 (very low intelligibility). For these two speakers, there is only a slight improvement. Finally, for low and very low intelligibility speakers, the monophone system recognizes their speech as well as the human listeners on all tasks but T06 (all speakers) and possibly T09 (M06 and F02).

Table 5: *PWC for monophone systems. Intelligibility of each talker is given in the second column.*

		ASR Task				
Talker	Intel.	T06	T07	T08	T09	T10
M09	86	31.14	47.49	46.82	46.53	50.36
M05	58	29.43	48.57	50.05	32.89	39.1
M06	39	14.57	37.54	36.31	26.82	30.48
F02	29	17.43	42.76	42.76	26.82	31.2
M07	28	15.57	44.83	43.78	28.98	34.01
F03	6	2.14	25.22	22.4	6.94	8.8
M04	2	1.2	6.07	5.94	2.37	2.59

Table 6: *PWC for triphone systems. Intelligibility of each talker is given in the second column.*

		ASR Task				
Talker	Intel.	T06	T07	T08	T09	T10
M09	86	25.29	63.65	63.32	46.47	52.04
M05	58	13.57	54.48	53.55	30.44	35.52
M06	39	5.43	58.82	52.53	29.8	34.01
F02	29	3.43	54.48	56.68	27.81	35.06
M07	28	7.86	57.14	60.65	32.48	43.87
F03	6	1	40	38.89	9.97	12.61
M04	2	1.8	4.83	5.81	1.96	2.82

triphone ASR: For all subjects, the variation in scores is similar to the one in the monophone case: performance on T06 worse than intelligibility rating; and scores for T07 and T08 respectively higher than those on T09 and T10, indicating performance deterioration on adding uncommon words. As with the monophone recognizers, incorporating digits into the vocabulary improves the PWC score (comparing T09 and T10 scores) by 5-12% absolute, for all subjects except F03 and M04 (for these two, there is again, only a slight improvement). On task T10, the triphone architecture gives a higher score than the intelligibility rating for all subjects with low and very low intelligibility, except M06. In fact, for these subjects, triphone ASR has also been able to achieve a performance at par with or better than the human listener rating on task T09.

monophone vs. triphone ASR: Refer to Figure 1. For vocabularies not containing the uncommon words (T07, T08), triphone systems outperform monophone systems by 3-17% absolute, for all subjects except M04 (2% intelligibility). For task T06 (uncommon words only), the monophone systems have better PWC scores (6-16% absolute) except for subjects F03 and M04. For these very–low intelligibility subjects, the performance of the two architectures is comparable. Finally, for tasks T09 and T10, monophone and triphone systems have similar

performance in terms of the PWC score.

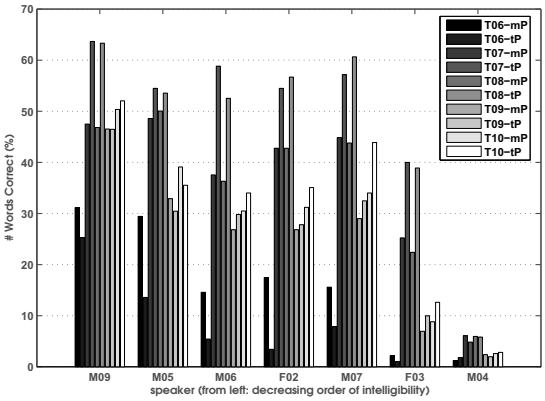

Figure 1: *Comparison of monophone ('mP') and triphone ('tP') PWC scores.*

whole-word vs. monophone vs. triphone ASR: Refer to Figure 2. For both tasks T07 and T08 and all subjects except M04, the monophone systems have the worst performance among the three architectures. For M04, the whole-word systems give the best performance on both tasks; for all other subjects, either whole-word (M09, M05, F02) or triphone (M06, F03) system performs best on both tasks (except for M07: whole-word score higher for task T07 and vice-versa for task T08).

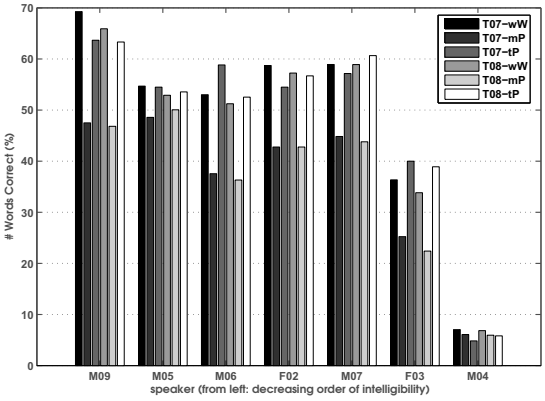

Figure 2: *Comparison of whole-word ('wW'), monophone ('mP') and triphone ('tP') PWC scores.*

5. Conclusions

We see that ASR systems trained specifically on a small amount of dysarthric speech (2 training tokens per utterance) have demonstrated recognition accuracies comparable to human listeners. Secondly, for some medium-sized vocabularies, the whole-word systems performed as well as triphone systems, indicating that simpler architectures are as capable as more complex ones for dysarthric speech recognition. These comparable performances permit the designer to choose from the two architectures: the triphone system is more flexible and scalable; on the other hand, the whole-word system is faster to train.

We believe that the most interesting outcome of these experiments is that, for subjects with low or very low intelligibility, ASR outperforms human listeners. This finding is not very surprising: (a) the ASR is speaker–dependent, therefore it has an advantage over unfamiliar human listeners, and (b) the ASR knows the task vocabulary, therefore it has an advantage over human listeners, who do not. Although the finding is easy to explain, it is significant because it demonstrates the feasibility of spoken language human–computer interaction for talkers with dysarthria. Most of the subjects described in this study are able to communicate reasonably well, in face-to-face interaction, with listeners who know them. Often, subjects will help their interlocutors to understand them by gesturing, repeating themselves, or, in other ways, providing situational context that helps a listener to guess what they might be trying to say. The results reported in this paper suggest that ASR with knowledge of the talker's voice and with knowledge of the task vocabulary outperforms human listeners without such knowledge, and that in many cases, the resulting PWC score approaches ranges that may be useful for human–computer interaction.

6. Acknowledgements

This work was supported by NSF grant 05-34106.

7. References

[1] D. Durham, "Key Steps to High Speech Recognition Accuracy," Sep. 2007. [Online]. Available: http://www.emicrophones.com.

[2] H. P. Chang, "Speech Input for Dysarthric Users," *Journal of the Acoustical Society of America*, vol. 94, no. 3, p. 1782, Sep. 1993.

[3] P. C. Doyle et al., "Dysarthric Speech: a comparison of Computerized Speech Recognition and Listener Intelligibility," *Journal of Rehabilitation Research and Development*, vol. 34, pp. 309–316, 1997.

[4] M. Hasegawa-Johnson, "Universal Access Automatic Speech Recognition Project," 2006. [Online]. Available: http://cita.disability.uiuc.edu/research/asr/overview.php.

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

[6] J. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. St. Louis: C. V. Mosby, 1995.

[7] R. D. Kent, G. Weismer, J. F. Kent, J. K. Vorperian, and J. R. Duffy, "Acoustic Studies of Dysarthric Speech: Methods, Progress and Potential," *Journal of Communication Disorders*, vol. 32, pp. 141–186, 1999.

[8] L. J. Platt, G. Andrews, and P. M. Howie, "Dysarthria of Adult Cerebral Palsy: II. Phonemic Analysis of Articulation Errors," *Journal of Speech and Hearing Research*, vol. 23, no. 1, pp. 41–55, 1980.

[9] 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.

[10] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric Speech Database for Universal Access Research," in *Proceedings of Interspeech*, 2008.

[11] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.