Towards a Comparative Database of Dysarthric Articulation

Frank Rudzicz[1], Pascal van Lieshout[2], Graeme Hirst[1],
Gerald Penn[1], Fraser Shein[3], and Talya Wolff[3]

University of Toronto, [1] *Department of Computer Science*,
[2] Department of Speech-Language Pathology. [3] Bloorview Kids Rehab Hospital
E-mail: frank@cs.toronto.edu

Abstract

This work describes ongoing data collection of synchronized acoustic and kinematic features of dysarthric speech. Recordings consist of electromagnetic articulographic measurements and 3D reconstructions from video over a variety prompts read by individuals with cerebral palsy and matching non-dysarthric subjects. Preliminary results show a preponderance of mispronounced plosives in dysarthric speech, and greater articulatory variability. An experiment with a standard hidden Markov model for speech recognition resulted in 84% triphone recognition accuracy for non-dysarthric speech, but only 6% accuracy for individuals with cerebral palsy.

1 Introduction

Dysarthria is a speech disorder that can severely limit intelligibility, affecting millions worldwide. The impairment of the facial muscles and other articulators in dysarthria is normally symptomatic of more general neuro-motor disabilities that can profoundly restrict individuals in daily life and make communication nearly impossible. Despite these difficulties, dysarthric speakers tend to prefer the naturalness and speed of spoken expression over other physical modes [4]. Properly engineered speech recognition systems would potentially improve the quality of life for these individuals, but current software is profoundly inadequate. Our prior experiments with traditional models, for example, show word-level accuracy of less than 4.5% on severely dysarthric speech against 84.8% on non-disabled speech on small-vocabulary sentences [8].

There is increasing evidence that the use of articulatory parameters improves speech recognition for non-dysarthric speakers [3]. For example, replacing typical Gaussian mixture output densities in hidden Markov models with Bayes nets representing the position and velocity of speech articulators has improved accuracy over acoustic-only models [6]. Training such acoustic-articulatory relationships can be performed with articulographic databases such as MOCHA [9], but there is currently no corresponding public database of dysarthric speech production. The Nemours database is currently the most extensive dysarthric speech database, consisting of 11 dysarthric males each uttering 74 syntactically invariant sentences and two additional paragraphs [7]. Despite its popularity, the Nemours database is limited in scope, and lacks physiological information.

2 Data Acquisition

The Torgo database of dysarthric speech is an ongoing project that will consist of aligned acoustic and articulatory recordings for the purpose of learning statistical relationships between dysarthric and non-dysarthric speech production. This database will consist of 12 to 15 subjects with dysarthria resulting primarily from cerebral palsy (spastic, athetoid, or ataxic) or amyotrophic lateral sclerosis, and gender-matched controls and is currently approximately half complete. Each participant records 3 hours of data split across multiple sessions in two 3D measurement environments. The first environment uses electromagnetic articulography (EMA) to measure the kinematics of the jaw, lips, and midsagittal plane of the tongue. Figure 1 shows an analysis window for a segment of data in which the waveform and spectrogram are aligned with the (x, y, z) co-ordinates of three points on the tongue. This system uses alternating electromagnetic fields generated by transmitter coils attached to a cube that surrounds the speaker's head. The second environment uses the Ariel Performance Analysis System (APAS) to reconstruct 3D motion parameters from video recordings of facial markers on the face [1].

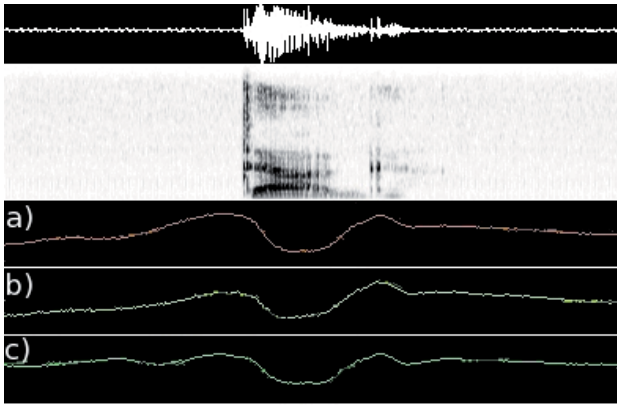


Figure 1: Aligned waveform, spectrogram, and tongue a) tip, b) body, and c) dorsum displacement during non-dysarthric */d ah g/*. Displacement data is measured in millimeters in three dimensions, and only the *y* dimension is shown, for clarity.

Participants read each utterance from an LCD monitor from a set of over 2000 stimuli. These stimuli are grouped into smaller collections which are internally randomized at runtime to ensure direct comparability between speakers who complete data at different rates. Single-word stimuli include repetitions of the English digits, the international radio alphabet, the 20 most frequent words in the British National Corpus, and words selected by Kent et al. to demonstrate relevant phonetic contrasts (e.g., alveolar-palatal fricatives, front-back vowels, stop-nasals) [5]. Single word stimuli are useful to study variation in isolation without boundary detection. Sentence stimuli are derived from the Yorkston-Beukelman assessment of intelligibility [10] and the TIMIT database [11]. Additionally, each participant is shown a small number of photographs from standardized tests of linguistic ability and asked to describe their contents in their own words. Whereas the use of meaningful sentences in general is amenable to syntactic and semantic processing, naturally produced speech is more likely to contain disfluencies, and is more representative of dictation-style speech.

All data is phonetically annotated to the TIMIT phone set [11] by a trained speech language pathologist to allow supervised frame-level training of phone-dependent acoustic/kinematic models. Additionally, all dysarthric participants are diagnosed by a speech-language pathologist according to the standardized Frenchay Dysarthria Assessment [2], which evaluates the functions of the articulators (e.g., respiration, tongue, palate), and clinical intelligibility. This assessment will be used to search for correlations between observable accuracy in several speech classification models and particular speech deficits according to phonological features. For instance, the degree of tongue disablement may be an indicator of poorer discrimination between front-back vowels.

3 Preliminary Results

As of this writing, the Torgo database is approximately 50% complete in terms of the amount of raw recording. It currently consists of speech and assessment data from 6 dysarthric individuals and matched controls, and is now being phonetically annotated and cleaned of some environmental acoustic noise. Some early observations are discussed in the following subsections.

3.1 Analysis of speaker data

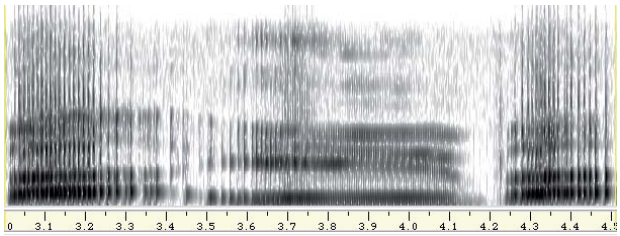
Table 1 shows the proportion of phonetic errors according to manner of articulation over the dysarthric data analyzed as of this writing. Notably, plosives are mispronounced 16%, 20%, and 19% of the time in word-initial, -medial, and -final positions, respectively, and substitutions in this class are exclusively from unvoiced to voiced, especially */t/* → */d/* and */p/* → */b/*. By comparison, only 5% of corresponding plosives are mispronounced, either dropped in the final position or incorrectly voiced in word-medial positions in regular speech. Also, our dysarthric data often includes many deleted affricates in word-final and fricatives in word-initial positions, almost all of which are static and alveolar. This does not occur in the corresponding non-dysarthric data.

	SUB (%)			DEL (%)		
	i	m	f	i	m	f
stops	13.8	18.7	7.1	1.9	1.0	12.1
affricates	0.0	8.3	0.0	0.0	0.0	23.2
fricatives	8.5	3.1	5.3	22.0	5.5	13.2
nasals	0.0	0.0	1.5	0.0	0.0	1.5
glides	0.0	0.7	0.4	11.4	2.5	0.9
vowels	0.9	0.9	0.0	0.0	0.2	0.0

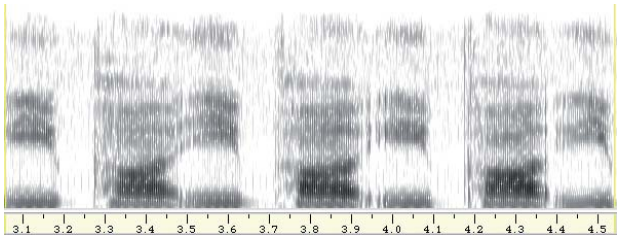
Table 1: Proportion of phone substitution (SUB) and deletion (DEL) errors in word-initial (i), word-medial (m), and word-final (f) positions across categories of manner for dysarthric data.

Figure 2 exemplifies some typical acoustic contrasts between dysarthric and non-dysarthric speech. In particular, dysarthric speech tends to be longer and more drawn out despite reduced breath support. On

average, dysarthric vowels are 116.7ms while control vowels are 45.5ms. This might partially be explained by an increase of brief staccato gaps in exhalation during sonorants. Dysarthric vowel acoustics are also slightly more variable, with an average variance across the first 7 mel-scaled frequency cepstral coefficients of 12.1, against 9.8 in control data. More severely dysarthric data tends to be more 'guttural', especially for velar phones.



(a)



(b)

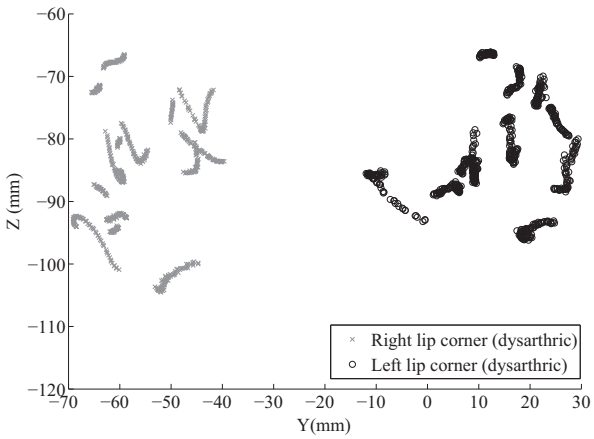
Figure 2: Repetitions of */iy pcl p ah/* over 1.5s by (a) a male speaker with athetoid CP, and (b) a female control. Dysarthric speech is notably several times slower and more strained than regular speech.

Figure 3 shows non-rotated head-relative positions of the left and right lip corners for a moderately dysarthric speaker and a control speaker pronouncing */uw/*. In general, tract variable targets for dysarthric data tend to be more variable.

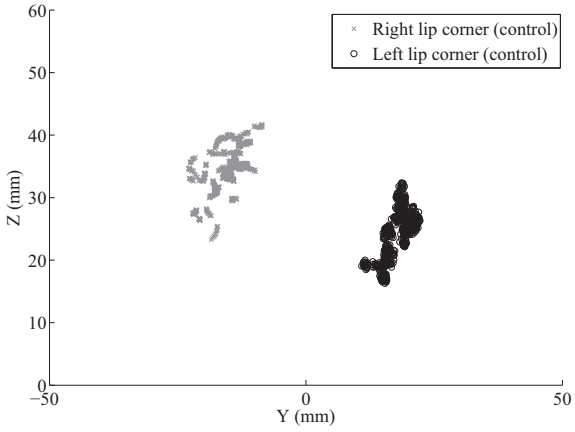
There is also a reduced range of motion as dysarthria becomes moderate or severe. During rounding of the lips, for example, the dysarthric speaker in Figure 3 could only reduce the spread of his lips to 69.8mm on average, against 35.1mm for the control speaker, despite similarly sized mouths.

3.2 Baseline speech recognition

We have compared our transcribed dysarthric and regular speech in a baseline triphone classifier consisting of standard tri-state left-right hidden Markov models (HMMs) with continuous 16-Gaussian mixture output densities decoded with the Viterbi algorithm and conditioned on the triphone label. When trained to regular speech, this model recognizes triphones with 6% and 84% accuracy for our dysarthric and control speakers, respectively.



(a)



(b)

Figure 3: Non-rotated left and right lip corner positions on coronal plane (*Z* is up-down, *Y* is left-right) for first 15 and 20 instances of */uw/* spoken by (a) a male dysarthric speaker, and (b) a female control, respectively. All points are relative to the centre of fixed positions on the speaker's head.

Our recent work comparing the effectiveness of neural network and kernel-based (i.e., support-vector machine) discriminative classifiers on phonological features of dysarthric speech from the Nemours database has shown a 10.9% relative error-rate reduction over traditional hidden Markov acoustic modeling, and an approximately linear relationship between performance accuracy and Frenchay-based intelligibility levels [2]. We will be applying these techniques to our own data in the near future.

4 Lessons learned and ongoing work

In addition to typical issues of speech data collection such as the need to suppress environmental noise, the development of the Torgo database has incurred some additional challenges specific to the

equipment and population. For instance, the induced magnetic field in the EMA cube is not completely uniform, so we take care to position the speaker within the center of the cube in order to minimize measurement error of recovered coil co-ordinates. The video cameras used for 3D APAS recordings are also somewhat sensitive to the position and facial characteristics of the speaker, so care must be taken that all markers are visible throughout each utterance. In general, APAS provides more facial motion data (e.g., the depressor anguli oris) but excluding all tongue motion.

There are also several practical challenges associated with recording individuals with cerebral palsy. For example, individuals in metal wheelchairs must be moved into a specially outfitted wooden chair prior to recording so as not to interfere with the EMA field. Decreased control of salivation and an increased risk of a severe gag reflex among cerebrally palsied participants can also make placing coils on the tongue very difficult. About 12% of EMA data from dysarthric individuals does not include all tongue coil positions. Involuntary movement such as shaking or extension of the neck also presents a problem for APAS recording, as the points on the face become less visible on video.

Data collection is ongoing and will continue until August 2009. We are currently developing mixed acoustic-articulatory models based on dynamic Bayesian networks that perform phone-based inference and classification. These temporal models allow arbitrary conditional probabilities to be learned between acoustics and articulatory motion and preliminary results indicate relative error reduction up to 15% over regular acoustic models. We have also begun to look at statistical relationships between acoustics and motor-function assessment, which includes correlation coefficients of over 0.95 between variables such as F2 variability and tongue protrusion. Work in the near future will include recovering articulation given acoustics, which will be compared against theoretical inference methods such as MIMICRI [3].

Acknowledgements

This research is made possible by Bell Canadas support through its Bell University Laboratories R&D program, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto.

References

[1] M. Craig, P. van Lieshout, and W. Wong. Suitability of a UV-based video recording system for the analysis of small facial motions during speech. *Speech Communication*, 49(9):679–686, September 2007.

[2] P. M. Enderby. *Frenchay Dysarthria Assessment*. College Hill Press, 1983.

[3] J. Hogden, P. Rubin, E. McDermott, S. Katagiri, and L. Goldstein. Inverting mappings from smooth paths through r^n to paths through r^m: A technique applied to recovering articulation from acoustics. *Speech Communication*, 49(5):331–436, 2007.

[4] J.-P. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely. Intelligibility of modifications to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 924–927, April 2003.

[5] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499, 1989.

[6] K. Markov, J. Dang, and S. Nakamura. Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. *Speech Communication*, 48(2):161–175, February 2006.

[7] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzjo, and H. Bunnell. The Nemours Database of Dysarthric Speech. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA, USA, October 1996.

[8] F. Rudzicz. Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. In *Proceedings of the Ninth International ACM SIGACCESS Conference on Computers and Accessibility*, Tempe, AZ, October 2007.

[9] A. Wrench. The MOCHA-TIMIT articulatory database, November 1999.

[10] K. M. Yorkston and D. R. Beukelman. *Assessment of Intelligibility of Dysarthric Speech*. C.C. Publications Inc., Tigard, Oregon, 1981.

[11] V. Zue, S. Seneff, and J. Glass. Speech database development: Timit and beyond. In *Proceedings of SIOA-1989*, volume 2, pages 35–40.