Lang Resources & Evaluation (2012) 46:523–541
DOI 10.1007/s10579-011-9145-0

ORIGINAL PAPER

The TORGO database of acoustic and articulatory speech from speakers with dysarthria

Frank Rudzicz · Aravind Kumar Namasivayam · Talya Wolff

Published online: 26 March 2011

Abstract This paper describes the acquisition of a new database of dysarthric speech in terms of aligned acoustics and articulatory data. This database currently includes data from seven individuals with speech impediments caused by cerebral palsy or amyotrophic lateral sclerosis and age- and gender-matched control subjects. Each of the individuals with speech impediments are given standardized assessments of speech-motor function by a speech-language pathologist. Acoustic data is obtained by one head-mounted and one directional microphone. Articulatory data is obtained by electromagnetic articulography, which allows the measurement of the tongue and other articulators during speech, and by 3D reconstruction from binocular video sequences. The stimuli are obtained from a variety of sources including the TIMIT database, lists of identified phonetic contrasts, and assessments of speech intelligibility. This paper also includes some analysis as to how dysarthric speech differs from non-dysarthric speech according to features such as length of phonemes, and pronunciation errors.

Keywords Speech · Articulation · Dysarthria

F. Rudzicz (✉)
Department of Computer Science, University of Toronto, Toronto, ON, Canada
e-mail: frank@cs.toronto.edu

A. K. Namasivayam
The Speech and Stuttering Institute, Toronto, ON, Canada

A. K. Namasivayam
Oral Dynamics Laboratory, Department of Speech-Language Pathology, University of Toronto, Toronto, ON, Canada
e-mail: a.namasivayam@utoronto.ca

T. Wolff
Holland Bloorview Kids Rehabilitation Hospital, Toronto, ON, Canada
e-mail: twolff@bloorview.ca

1 Introduction

This paper describes the collection of a new English speech database of aligned acoustics and measured 2D and 3D articulatory features from speakers with either cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), which are two of the most prevalent causes of speech disability (Kent and Rosen 2004). This database, called TORGO, is the result of a collaboration between the departments of Computer Science and Speech-Language Pathology at the University of Toronto and the Holland-Bloorview Kids Rehab hospital in Toronto.

Both CP and ALS are examples of a set of speech impairments, collectively called dysarthria, which are caused by disruptions in the neuro-motor interface. These disruptions do not affect the comprehension and cognitive aspects of the production of natural language, but distort motor commands to the articulators of speech, resulting in atypical and relatively unintelligible speech in most cases (Kent 2000). This unintelligibility adversely affects the use of traditional automatic speech recognition (ASR) software to the point where severely dysarthric subjects may have a word-error rate of 97.5% on modern systems against 15.5% for the general population (Rudzicz 2007). The inability of modern ASR to effectively understand dysarthric speech is a major problem, since the more general physical disabilities often associated with the causative neuro-motor disruptions can make other forms of computer input, such as keyboards or touch screens, especially difficult (Hosom et al. 2003). Since speakers with dysarthria differ from the general population in the manner of their articulation, measuring that articulation empirically is essential.

The TORGO database is primarily a resource for developing ASR models more suited to the needs of people with atypical speech production, although it is equally useful to the more general ASR community. A primary reason for collecting detailed physiological information is to be able to explicitly learn 'hidden' articulatory parameters automatically in computer speech models via statistical pattern recognition. Recent research has shown that modelling conditional relationships between articulation and acoustics in Bayesian networks can reduce error by about 28% (Markov et al. 2006; Rudzicz 2009) relative to acoustic-only models for regular speakers. Research in this area has been relatively preliminary, however there has been a marked increase in the use of articulatory models in speech recognition recently (Livescu et al. 2007).

This database is also useful in the clinical domain, especially by linguists and pathologists interested in studying atypical speech production. The stimuli uttered by the participants in this study have been carefully chosen to emphasize articulatory contrasts, which in turn can be compared against thorough assessments of speech-motor function, as described below.

1.1 Existing databases

To date, no database combining the acoustics and endogenous articulation (e.g., tongue movement) of dysarthric speech is publicly available. Data collection with dysarthric speakers has usually involved fewer than 5 participants

(Hasegawa-Johnson et al. 2006), frequently producing only about 25 utterances each (Jayaram and Abdelhamied 1995). One exception is the A. I. duPont Institute's Nemours database, which is composed of 11 male speakers with varying degrees of dysarthria, each of whom speaks 74 nonsense sentences and two sensible paragraphs (Menendez-Pidal et al. 1996). Each nonsense sentence has the form *The* N_0 *is Ving the* N_1, where N_0 and N_1 are unique monosyllabic nouns and V is a monosyllabic verb. The target words, N_0, V, and N_1, were randomly selected without replacement in order to provide closed-set phonetic contrasts (e.g., place, manner, voicing). Additionally, one non-dysarthric speaker repeated each sentence in the database. This database costs $100USD to license. In our work on building ASR systems for people with dysarthria, we found that the phonemic distribution in the Nemours database was relatively sparse (Rudzicz 2007, 2009). Naturally, the absense of articulatory measurements in this database prohibits its use in the construction of more complex models in speech recognition, and in the study of the underlying effects of dysarthria.

Public databases of endogenous articulation exist, but only for non-dysarthric speakers. The University of Edinburgh's free MOCHA database consists of 460 sentences derived from the TIMIT database (Zue et al. 1989) and consists of articulatory measurements from electromagnetic articulography (EMA) (500 Hz sample rate), laryngography (16 kHz sample rate) and electropalatography (EPG, 200 Hz sample rate) (Wrench 1999). Each of the 460 phrases in this database were uttered by both a male and a female British speaker without dysarthria. The EMA system measured the 2D midsagittal co-ordinates of 8 points of interest in the vocal tract, namely the upper lip, lower lip, upper incisor, lower incisor, tongue tip, tongue blade (1 cm from the tongue tip), tongue dorsum (1 cm from the tongue blade), and velum.

Recently, Yunusova et al. (2008) have collected X-ray microbeam data with 7 individuals with Parkinson's disease and 8 with ALS. This data includes point-data in similar positions to the MOCHA database and generally follows the protocol and methodology of the Wisconsin X-ray microbeam database for non-dysarthric speakers (Westbury 1994). This database only includes 10 stimuli per speaker, however, which is not enough to train ASR systems. This database may be freely procured from the authors for academic use. The University of Illinois also provides data from 10 dysarthric individuals with cerebral palsy (Kim et al. 2008), although this data does not include measurements of the tongue, for instance.

The following sections describe our study population, their speech-motor assessment, and the data collection process.

2 Data collection

Data collection began in 2008 through collaboration between the departments of Computer Science and Speech-Language Pathology at the University of Toronto, Holland-Bloorview Kids Rehab hospital in Toronto, and the Ontario Federation for Cerebral Palsy. The following section describes various aspects of the data collection process.

2.1 Subjects

Seven dysarthric subjects (4 male, 3 female) have so far been assessed in this study, covering a wide range of intelligibility. Dysarthric subjects were recruited by a speech-language pathologist at the Bloorview Research Institute in Toronto. The subjects were between the ages of 16 and 50 years old and have dysarthria resulting from cerebral palsy (e.g., spastic, athetoid, or ataxic). In addition, one subject with a confirmed diagnosis of amyotrophic lateral sclerosis (ALS) was recruited. These individuals were matched according to age and gender with non-dysarthric subjects from the general population. Having an equal number of dysarthric and control speakers is useful for comparing acoustic and articulatory differences, and for analyzing these relationships mathematically and functionally (Hosom et al. 2003; Kain et al. 2007). Data has since been collected from two additional subjects with cerebral palsy, although that data has not yet been analyzed.

Each subject began the data collection process with a short questionnaire that covers general demographic data and health-related questions that can impact speech and language function including various types of motor problems, both gross (e.g., standing, balancing) and fine (e.g., writing, swallowing). All participants were required to have a negative history of severe hearing or visual problems and of substance abuse, and to be able to read at a 6th grade elementary level. This was further quantified by requiring that their cognitive function lie above or at level VIII (i.e., Purposeful-Appropriate) on the Rancho scale (Herndon 1997), which is determined during a pre-visit questionnaire.

2.2 Assessment

The motor functions of each experimental subject were assessed according to the standardized Frenchay Dysarthria Assessment (FDA; Enderby 1983) by a speech-language pathologist. This assessment is designed to categorize and diagnose individuals with dysarthria while being easily applicable to therapy, sensitive to changes in speech, simple and quick to administer, and easily communicable within professional teams. There exist other assessment measures of oral motor ability, such as the Assessment of Intelligibility of Dysarthric Speech (AIDS; Yorkston and Beukelman 1981), which quantifies the intelligibility of single words, sentences, and speaking rates of adults and adolescents with dysarthria. However, these tend to focus only on speech production, whereas the FDA also includes analysis of the movement of the articulators in non-linguistic contexts.

The Frenchay assessment measures 28 relevant perceptual dimensions of speech grouped into 8 categories, namely reflex, respiration, lips, jaw, soft palate, laryngeal, tongue, and intelligibility as described in Table 1. Influencing factors such as rate and sensation are also recorded. To measure most of these dimensions, the administering clinician either engages the subject in communication or has the subject perform a simple task (e.g., drinking from a cup of water) while observing their oral movements. The subject's oral behaviour is rated on a 9-point scale and plotted with a simple bar graph. The assessment provides characterizations of

Table 1 Frenchay dysarthria assessment dimensions (Enderby 1983), each on a scale of 0 (no function) to 8 (normal function)

Category	Test	Observation	Males $\mu(\sigma)$	Females $\mu(\sigma)$	All $\mu(\sigma)$
Reflex	Cough	Presence of cough during eating and drinking.	6(2.45)	7.3(0.96)	6.6(1.85)
	Swallow	Speed and ease of swallowing liquid	7(2.0)	8(0.0)	7.5(1.41)
	Dribble	Presence of drool generally	6.5(2.38)	7.5(1.0)	7(1.77)
Respiration	At rest	Ability to control breathing during rest	4(2.71)	8(0.0)	6(2.78)
	In speech	Breaks in fluency caused by poor respiratory control	4(2)	6.5(3)	5.3(2.71)
Lips	At rest	Asymetry of lips during rest	6.3(2.36)	8(0.0)	7.1(1.81)
	Spread	Distortion during smile	6(2.31)	8(0.0)	7(1.85)
	Seal	Ability to maintain pressure at lips over time	3.3(3.4)	7(2)	5.1(3.27)
	Alternate	Variability in repetitions of “*oo ee*”	3.8(2.87)	7(2)	5.4(2.88)
	In speech	Excessive briskness or weakness during regular speech	4.3(1.89)	6.5(1.91)	5.4(2.13)
Jaw	At rest	Hanging open of jaw at rest	7(1.15)	8(0.0)	7.5(0.93)
	In speech	Fixed position or sudden jerks of jaw during speech	5.8(2.63)	6.3(2.36)	6.1(2.42)
Velum	Fluids	Liquid passing through velum while eating	7(2.0)	8(0.0)	7.5(1.41)
	Maintenance	Elevation of palate in repetitions of “*ah ah ah*”	5.8(2.06)	7.5(1.0)	6.6(1.77)
	In speech	Hypernasality or imbalanced nasal resonance in speech	6.3(2.36)	6(2.83)	6.1(2.42)
Laryngeal	Time	Sustainability of vowels in time	5.3(2.5)	7.5(1.0)	6.4(2.13)
	Pitch	Ability to sing a scale of distinct notes	2(2.16)	5.3(2.5)	3.6(2.77)
	Volume	Ability to control volume of voice	3.5(3.11)	4.8(3.2)	4.1(3.0)
	In speech	Phonation, volume, and pitch in conversational speech	3.3(2.87)	6(2.83)	4.6(3.02)
Tongue	At rest	Deviation of tongue to one side, or involuntary movement	5.5(2.08)	5.5(1.73)	5.5(1.77)
	Protrusion	Variability, irregularity, or tremor during repeated tongue protrusion and retraction	3.8(3.1)	5.3(1.5)	4.5(2.39)
	Elevation	Laboriousness and speed of repeated motion of tongue tip towards nose and chin	3.3(3.2)	4.3(1.71)	3.7(2.43)
	Lateral	Laboriousness and speed of repeated motion of tongue tip from side to side	3.8(3.1)	3.5(1.91)	3.6(2.39)
	Alternate	Deterioration or variability in repetitions of phrase “*ka la*”	4(2.71)	5.3(1.91)	4.9(2.9)
	In speech	Correctness of articulation points and laboriousness of tongue motion during speech generally	4(2.71)	6(2.83)	5(2.78)

Table 1 continued

Category	Test	Observation	Males $\mu(\sigma)$	Females $\mu(\sigma)$	All $\mu(\sigma)$
Intelligibility	Words	Interpretability of 10 isolated spoken words from a closed set	4(2.94)	4.5(2.52)	4.3(2.55)
	Sentences	Interpretability of 10 spoken sentences from a closed set	3.5(3.32)	5.3(3.4)	4.4(3.25)
	Conversation	General distortion or decipherability of speech in casual conversation	4.5(2.38)	6.5(1.91)	5.5(2.27)

behaviours across this 9-point scale. For example, for the cough reflex dimension, a subject would receive a grade of ‘a’(8) for no difficulty, ‘b’(6) for occasional choking, ‘c’(4) if the patient requires particular care in breathing, ‘d’(2) if the patient chokes frequently on food or drink, and ‘e’(0) if they are unable to have a cough reflex. The resulting graph provides a high-level overview to the clinician to quickly identify problematic aspects of speech or non-speech (e.g., swallowing).

The mildly dysarthric speakers were able to participate in all tasks required of them for the assessment. The more severely dysarthric speakers also engaged in all tasks but levels of fatigue and poor breath control inhibited them from completing some of these tasks. Assessment data of this type is useful in analyzing how modifications to ASR software affects achievable accuracy across the spectrum of intelligibility levels. For example, alterations to the process by which vowels are categorized by the machine may have greater impact for those individuals with more atypical tongue movement, as opposed to pronounced velum differences. Table 1 shows the mean (μ) and standard deviation (σ) of our participants, split by gender, across each of the 28 dimensions of the Frenchay assessment.

2.3 Speech stimuli

All subjects read English text from a 19-inch LCD screen placed 60 cm in front of them. One subject experienced some visual exhaustion near the end of one session, and therefore repeated a small section of verbal stimuli spoken by an experimenter rather than read these stimuli. No discernible effect of this approach was measured. The stimuli were presented to the participants in randomized order from within fixed-sized collections of stimuli in order to avoid priming or dependency effects. Dividing the stimuli into collections in this manner guaranteed overlap between subjects who speak at vastly different rates, which is the case when dealing with severely dysarthric speakers, especially when the time allowable for each session is limited. There is no dependency relation between the sessions and the presented stimuli. The collected speech data covers a wide range of articulatory contrasts, is phonetically balanced, and simulates simple command vocabularies typical of assistive ASR technology. The following types of stimuli are included across all collections of data:

Non-words These are used to control for the baseline abilities of the dysarthric speakers, especially to gauge their articulatory control in the presence of plosives and prosody. Speakers are asked to perform the following:

- 5–10 repetitions of */iy-p-ah/*, */ah-p-iy/*, and */p-ah-t-ah-k-ah/*, respectively. These sequences allow us to observe phonetic contrasts around plosive consonants in the presence of high and low vowels, and have been used in other studies (Bennett et al. 2007).
- High-pitch and low-pitch vowels maintained over 5 s (e.g., *"Say 'eee' in a high pitch for 5 s"*). This allows us to explore the use of prosody in assistive technology, as many dysarthric speakers who have difficulty with articulation can control pitch to some degree (Patel 2002).

Short words These are useful for studying speech acoustics without the need for word-boundary detection. These stimuli include formant transitions between consonants and vowels, the formant frequencies of vowels, and acoustic energy during plosive phonemes, as explored by Roy et al. (2001). This category includes the following:

- Repetitions of the English digits 1–10, *yes*, *no*, *up*, *down*, *left*, *right*, *forward*, *back*, *select*, *menu*, and the international radio alphabet (i.e., *alpha*, *bravo*, *charlie*, etc.). These words are useful for hypothetical command-and-control software for accessibility.
- 50 words from the the word intelligibility section of the Frenchay Dysarthria Assessment (Enderby 1983).
- 360 words from the word intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman 1981). These are grouped into phonetically similar words, as was presented in the Nemours database (Menendez-Pidal et al. 1996) (e.g., *hit*, *hat*, and *hut* are a trio of monosyllabic words differing only in their vowel).
- The 10 most common words in the British National Corpus (Clear 1993).
- All phonetically contrasting pairs of words from Kent et al. (1989). These are grouped into 18 articulation-relevant categories that affect intelligibility, including glottal/null, voiced/voiceless, alveolar/palatal fricatives and stops/nasals; these are shown in Table 3 in Appendix.

Restricted sentences In order to utilize lexical, syntactic, and semantic processing in ASR, full and syntactically correct sentences are recorded. These include the following:

- Preselected phoneme-rich sentences such as "*The quick brown fox jumps over the lazy dog*", "*She had your dark suit in greasy wash water all year*", and "*Don't ask me to carry an oily rag like that*".
- The Grandfather passage from the Nemours database (Menendez-Pidal et al. 1996).
- 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman 1981).
- The 460 TIMIT-derived sentences used as prompts in the MOCHA database (Wrench 1999; Zue et al. 1989).

Unrestricted sentences Since a long-term goal is to develop applications capable of accepting unrestricted and novel sentences, we elicited natural descriptive text by asking participants to spontaneously describe 30 images of interesting situations taken randomly from among the cards in the Webber Photo Cards: Story Starters collection (Webber 2005). These are similar in nature to images used in other standardized tests of linguistic proficiency (Campbell et al. 2001). This data complements restricted sentences in that they more accurately represent naturally spoken speech, including disfluencies and syntactic variation.

2.4 Instrumentation

In each of three sessions, subjects are prepared for either of two instrumental studies. The first involves the use of EMA and the other involves video recordings of facial markers using specialized software to extract their positions over time. For EMA, the preparation takes approximately 30 min in which sensors are placed on the relevant locations of the speech articulators as described below in Sect. 2.4.1. In the video-based setup, preparation takes about 20 min and involves the placement of phosphorescent markers on relevant landmark positions of the face, as described in Sect. 2.4.2. The actual data collection process takes no more than 1 h thereafter in either the EMA or video configurations. Of the three recording sessions, two are within the EMA environment since we are interested in the motion parameters of the tongue, which are unavailable in the video setup. We perform three sessions for each participant in order to check the reliability and variability of our data over time. Moreover, the literature suggests that EMA can provide a reliable estimate of speaker variability of speech parameters over time (van Lieshout et al. 1997).

2.4.1 Electromagnetic articulograph kinematics

The collection of movement data and time-aligned acoustic data is carried out using the three-dimensional AG500 electro-magnetic articulograph (EMA) system (Carstens Medizinelektronik GmbH, Lenglern, Germany) with fully-automated calibration. The 3D-EMA system is considered state-of-the-art technology for studying speech movements and its principles have been elaborated elsewhere (Hoole et al. 2003; van Lieshout et al. 2008; Yunusova et al. 2009; Zierdt et al. 2000). This system allows for 3D recordings of articulatory movements inside and outside the vocal tract, thus providing a detailed window on the nature and direction of speech related activity.

In the AG500 system, six transmitters attached to a clear cube-shaped acrylic plastic structure (dimensions L 58.4 × W 53.3 × H 49.5 cm) generate alternating electromagnetic fields as shown in Fig. 1a. Each transmitter coil has a characteristic oscillating frequency ranging from 7.5 to 13.75 kHz (Yunusova et al. 2009). When sensors (also called *transducers*) are brought into the field, induction generates a weak current oscillating with the same frequencies. The energy in each frequency of the induced complex signal depends on the distance of the sensor from the transmitters and its orientation. The spatial position of the sensor coil in the field is then determined by identifying the strength of the contribution of each transmitter coil via a process of demodulation of the complex signal induced in the sensor

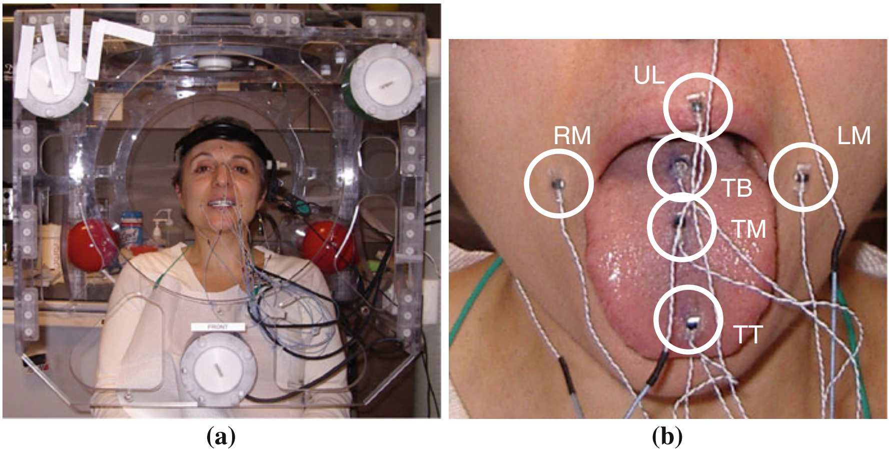

Fig. 1 The AG500 electromagnetic articulography system. **a** shows a participant seated in the center of the EMA cube. **b** shows the placement coils on the right mouth (*RM*), left mouth (*LM*), upper lip (*UL*), tongue tip (*TT*), tongue middle (*TM*), and tongue back (*TB*)

(Yunusova et al. 2009). The induced voltage values in the sensors are compared to expected values based on a known field model (Zierdt et al. 1999) and the difference is expressed as root-mean-square (RMS) error. The system translates these voltages into 3D coordinates of sensor positions over time. As will be discussed later, the RMS error is used to position the subject within the recording field and in part to measure the recording accuracy of the system.

As recommended by the manufacturer, the AG500 system is calibrated prior to each session subsequent to a minimum of a 3 h warm-up time. It is reported that, at or close to the cube's centre, positional errors are significantly smaller (Yunusova et al. 2009) compared to the peripheral regions of the recording field within the cube. For our system, the stable volume around the center was roughly 0.008 m^3 (approximately the size of a basketball). Thus, care was taken to ensure that all participants were as close to the cube centre as possible, as shown in Fig. 1a. The subject positioning within the cube was aided visually by the `Cs5view` real-time position display program (Carstens Medizinelektronik GmbH, Lenglern, Germany). This allowed the experimenter to continuously monitor the subject's position within the cube (repositioning the subject if required) and thereby maintain low RMS error values[1] to ensure good tracking of the sensor coils.

Sensor coils were attached to three points on the surface of the tongue, namely tongue tip (TT—1 cm behind the anatomical tongue tip), the tongue middle (TM—3 cm behind the tongue tip coil), and tongue back (approximately 2 cm behind the tongue middle coil). A sensor for tracking jaw movements (JA) is attached to a custom mould made from polymer thermoplastic that fits the surface of the lower incisors and which is necessary for a more accurate and reproducible recording.

[1] The `Cs5view` real-time position display flags a coil in red if the RMS error exceeds 30 units; however, the RMS during rarely exceeded 8 units across all coils, which is suitable for minimizing position tracking errors (Kroos 2008; Yunusova et al. 2009).

Four additional coils are placed on the upper and lower lips (UL and LL) and the left and right corners of the mouth (LM and RM). The placement of some of these coils is shown in Fig. 1b. Further coils are placed on the subject's forehead, nose bridge, and behind each ear above the mastoid bone for reference purposes and to record head motion. Except for the left and right mouth corners, all sensors that measure the vocal tract lie generally on the midsagittal plane on which most of the relevant motion of speech takes place. Sensors are attached by thin and light-weight cables to recording equipment but do not impede free motion of the head within the EMA cube. Many cerebrally palsied individuals require metal wheelchairs for transportation, but these individuals were easily moved to a wooden chair that does not interfere with the electromagnetic field for the purposes of recording.

2.4.2 Video-based articulatory kinematics

Although EMA provides detailed recordings of the tongue, which is not normally visible, typical use of speech recognition software will not likely involve such measurements. Therefore, we implement a second recording environment whose purpose is to derive more varied surface-level facial information using digital cameras. Here, recorded positions are meant to mimic the type of information that can be extracted from webcam-based face-recognition software.

Here, two digital video cameras are placed equidistant from the subject, at approximately 45 degree angles to their midsagittal plane, to the front-left and front-right of the subject. Video is captured at 60 frames per second and audio at 16,000 Hz on both cameras. This audio is used for synchronizing the frames from both cameras and for separate acoustic measurements.

Two 250 W black lights are used to illuminate small (2 mm radii) glow-in-the-dark markers placed on the surface of the subject's face at selected points around the lips and over the orbicularis oris, depressor anguli oris, and depressor labii inferioris muscles as in previous studies on speech production (Craig et al. 2007) and as shown in Fig. 2.

Facial markers are tracked by specialized vision software based on strong contrasts between the reflection of the markers and the relatively darker background. These positions are converted into 3-dimensional co-ordinates using pairs of aligned video images and an estimated inter-camera calibration (Tsai 1987). Calibration between cameras is performed by first filming a reference object with a known geometry, namely a cube with 30 cm sides.

2.4.3 Acoustics and microphones

All acoustic data is recorded simultaneously through two microphones. The first is an Acoustic Magic Voice Tracker array microphone with 8 recording elements generally arranged horizontally along a span of 45.7 cm. The device uses amplitude information at each of these microphones to pinpoint the physical location of the speaker within its 60-degree range and to reduce acoustic noise by spatial filtering and typical amplitude filtering in firmware. This microphone records audio at

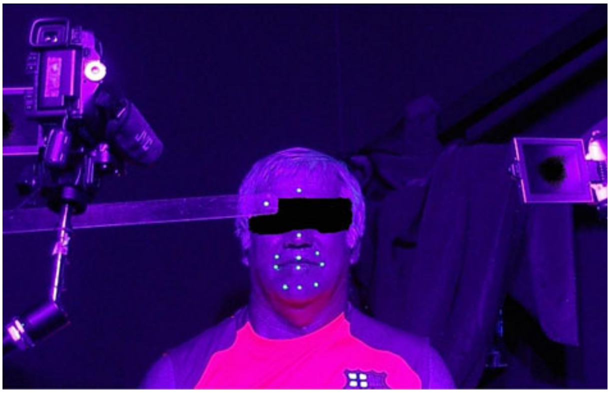
[redacted]

Fig. 2 The binocular video recording setup showing the placement of *phosphorescent dots* on the subject's face

44.1 kHz and is placed facing the participant at a distance of 61 cm. The second microphone is a head-mounted electret microphone which records audio at 16 kHz. The electromagnetic field produced by this microphone does not demonstrably affect the field of the EMA system, and so it can be worn during all recordings.

Signals from the two microphones are temporally aligned using simple cross-correlation. Namely, given the two discrete signals f and g, we compute the complex conjugate of the first, giving signal f^* consisting of real and phase values, and compute the cross-correlation by

$$(f\bigstar g)[n] = \sum_{m} = -N^N f^*[m]g[n+m]$$

where N is the length of the longer of the two sequences. The maximum value of this cross-correlation signal is the time delay between the jointly stationary signals, which is the speech signal recorded by both microphones. An example of this alignment is shown in Fig. 3.

Finally, acoustic noise reduction is performed using the minimum mean squared error estimate of the spectral amplitude (Ephraim and Malah 1985; Martin 2001). Furthermore, the use of multiple microphones permits the use of various noise-reduction algorithms not suitable for single sources (Aarabi and Shi 2004; Shi et al. 2007).

3 Data post-processing

The AG500 EMA system has an expected error specification of up to 0.5 mm in each dimension (X, Y, and Z) and an angular error (θ) of less than half of a degree. However, in reality it is possible that accuracy may vary slightly across different AG500 systems due to set-up and environmental conditions such as ambient room temperature, type of sensor coils used, and existing electromagnetic fields in the room. These conditions may also vary across time (Kaburagi et al. 2005; Yunusova et al. 2009). To estimate more realistic values, we carried out a series of static and

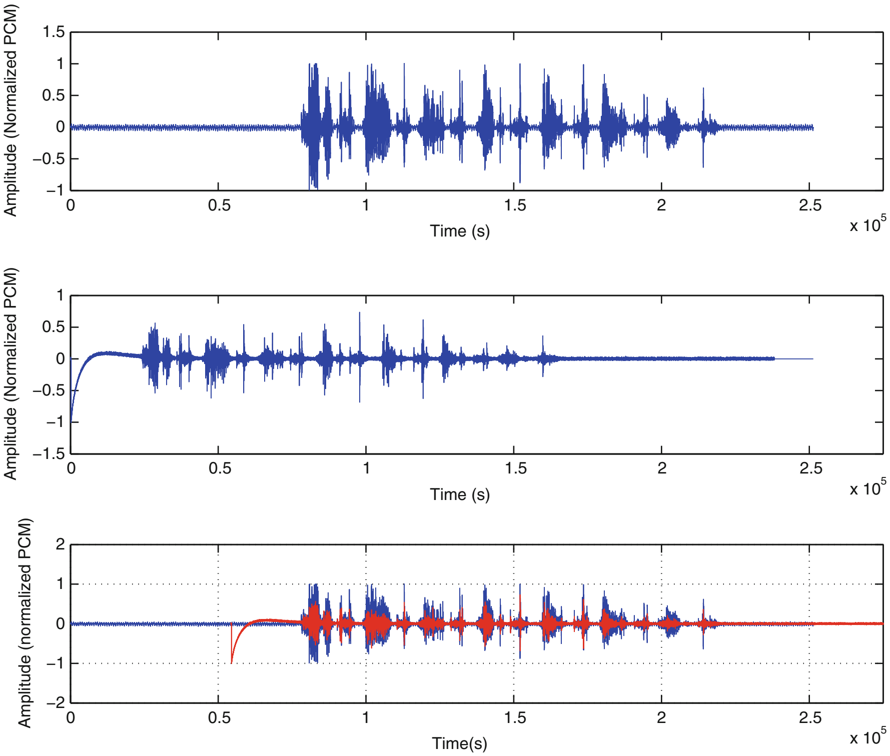

Fig. 3 Alignment of two acoustic sources with the cross-correlation method. The *top waveform* is the signal recorded by the head-mounted microphone. The *middle waveform* is the signal recorded by the directional microphone. The *bottom superposition of waveforms* is produced by cross-correlation

dynamic accuracy measurements for the AG500 system. For static measurement, 3-dimensional Euclidean distances between pairs of sensor coils were calculated. The sensors used here were those located on relatively rigid surfaces, namely the forehead, nose bridge, and behind the ear on the skin covering the right and left mastoid bone. Under ideal conditions, the distance between the pairs of sensors should remain constant throughout all trials for a given session. In other words, smaller average standard deviations for the 3D Euclidean distances between pairs of reference coils would imply lower static system noise or relative error. Similar methods have been applied to a camera-based marker tracking system (Craig et al. 2007) and in other 3D EMA systems (Hoole and Zierdt 2010; Yunusova et al. 2009). This Euclidean RMS method provides a real and accurate measure of intrinsic system noise and relative error for each recording session. The average value was 0.2 mm across all pairs. These numbers may be taken roughly as the lower limit of the system's resolution (Kroos 2008).

Recent studies have indicated that position errors in dynamic measurements, as opposed to static measurements, may be larger in magnitude and may vary across the three spatial dimensions (Kroos 2008; Yunusova et al. 2009). We therefore ran a set of dynamic accuracy measurements for all coils using a specific tool recommended by the manufacturer. This allows us to estimate dynamic spatial errors as a function of sensor orientation. This accuracy checking tool is a mechanical device that is rigidly fixed in the centre of the cube's recording field and allows user defined manipulations of sets of coils in different orientations and directions. The device is constructed such that sets of coils placed on it can only travel a fixed distance (70 mm) in a particular direction. For the current study, we displaced 3 sets of 4 coils (i.e. $\langle(1..4, 5..8, 9..12\rangle)$ across the entire 70 mm distance six times in a row in each dimension (*X*, *Y*, and *Z*). A custom Matlab algorithm calculated the maximum 3D Euclidean displacement between points in that trial, as well as the average 3D Euclidean displacement. The algorithm automatically finds the coils that are being moved and the dimensions in which they are moving using maximum variance. Ideally, the maximum and average 3D Euclidean displacement values should be as close to 70 mm as possible. The amount of deviation from 70 mm provides an estimate of direction specific spatial accuracy of the system. We calculated the accuracy averaged across all 12 sensor coils per dimension. This was in the range of 0.54–0.60 mm in the *Z* (up/down) dimension, 0.34–0.59 mm in the *X* (front/back) dimension, and 0.84–1.07 mm in the *Y* (left/right) dimension.

3.1 Data normalization

Position normalizations and corrections for head movements were carried out using custom-made `NormPos` software from the manufacturer of the AG500. The `NormPos` program does a sample-by-sample head normalization by rotating and shifting the coordinate system such that all reference sensors remain in the same 3D location across all samples and trials. Computationally, this is carried out using algorithms similar to 3D pose estimation methods (Kroos 2008). Such algorithms calculate transformation parameters that can transform head position of a given sample to an experimenter chosen arbitrary reference position (that defines the orientation of the head and the origin of the coordinate system). The transformation parameters are derived by minimizing the sum of the squared distances between the reference sensor coils in the reference position and the actual position in other trials using linear least squares approaches such as (Kroos 2008). The `NormPos` program stores these transformational parameters as a normalization pattern file. This normalization pattern file is then used to rotate and translate all other (non-reference) sensor coils positions in the remaining trails of the experiment to yield articulation trajectories that are corrected for head movements and with a fixed head-orientation that is identical across trials (and across subjects).

Since the `NormPos` program uses a normalization pattern file that is based on a single trial, the quality of the head movement correction for the entire experiment depends on the quality of the data from the reference sensor coils in that trial. At times, the quality of data may not be equally good in all reference coils (as in the case of coil detachment and/or position tracking errors). For this reason, researchers have recommended the use of more than two reference sensor coils,[2] typically four,

[2] Two sensors, in principle, are sufficient to characterize the six degrees of freedom related to rigid-body motions (Hoole and Zierdt 2010).

Table 2 Proportion of phoneme 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

	SUB (%)			DEL (%)		
	i	*m*	*f*	*i*	*m*	*f*
Plosives	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

to allow for redundancy in the available reference sensor coils (Hoole and Zierdt 2010). For the present study, the two noise measures that were previously discussed were used to decide which two or three reference sensor coils (of the four available) were suitable to create the normalization pattern file (Hoole and Zierdt 2010). Generally, the nose bridge and the two sensor coils behind the ears had the least amount of noise and were chosen to create the normalization pattern.

4 Aspects of dysarthric speech in TORGO

There are a number of features which differentiate dysarthric and non-dysarthric speech in our recorded data. Table 2 shows the proportion of phonemes that were mispronounced according to manner of articulation for dysarthric speech. Plosives are mispronounced most often, with substitution errors exclusively caused by errant voicing (e.g. */d/* for */t/*). By comparison, only 5% of corresponding plosives in total are mispronounced in non-dysarthric speech. Furthermore, the prevalence of deleted affricates and plosives in word-final positions, almost all of which are alveolar, does not occur in the corresponding non-dysarthric speech data.

Figures 4 and 5 show the durations of various steady-state phonemes (i.e., vowels and consonants, respectively) averaged across the dysarthric and control groups of TORGO. All vowels produced by dysarthric speakers are significantly slower than their non-dysarthric counterparts at the 95% confidence interval and can be up to twice as long, on average. We note that the divergence of the nasal consonants are most severe, which may be indicative of poor control of the velum, but the degree of this divergence does not significantly outweigh those among the other consonants.

5 Ongoing work

All data in this paper, including noise-reduced audio and articulatory data, can be obtained by contacting the authors at the University of Toronto. This data is

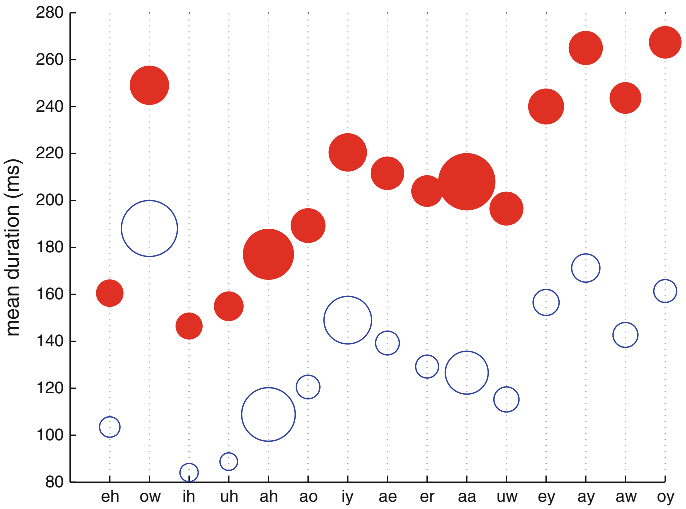

Fig. 4 Duration of vowels among dysarthric speakers (*filled circles*) and control speakers (*unfilled circles*). The heights of the circles correspond to the average duration, in milliseconds, of the associated vowel and the radii of the circles represent one standard deviation of the data. Vowels are sorted from left to right according to increasing divergence between groups

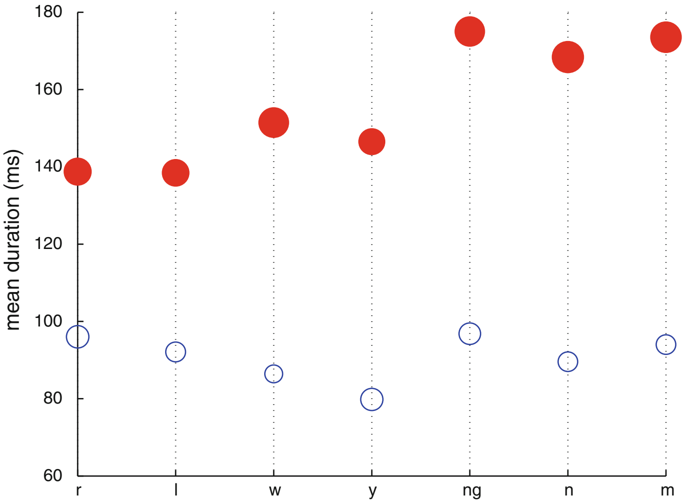

Fig. 5 Duration of selected consonants among dysarthric speakers (*filled circles*) and control speakers (*unfilled circles*). The heights of the circles correspond to the average duration, in milliseconds, of the associated consonant and the radii of the circles represent one standard deviation of the data. Consonants are sorted from left to right according to increasing divergence between groups

currently being used as input to automatic statistical pattern recognition software that identify relationships between the source features of speech and their intended linguistic meanings. Features such as frequency, energy, and pitch are automatically extracted from the raw audio signal using standard feature extraction techniques (e.g., Mel-frequency cepstral coefficients). These sorts of features are typical input to speech recognition programs that identify spoken phonemes through various multivariate Bayesian methods or by probabilistic regression. Articulatory models learned from kinematic data have been embedded directly within standard recognition systems, which has been shown to improve overall accuracy within the Hidden Markov Model (Markov et al. 2006; Rudzicz 2009).

The data described in this paper reveals that a lack of articulatory control can often lead to observable acoustic consequences. For example, our dysarthric data contain considerable involuntary types of speech and non-speech noise such as velopharyngeal or glottal noise (often associated with respiration), noisy swallowing problems, hesitation (e.g., false starts), and repetition. We intend to work towards methods of explicitly identifying regions of non-speech noise in our speech recognition systems for dysarthric speakers. Since real-world applications of such technology are not likely to have access to measurements of the vocal tract, we are currently developing methods that estimate the configuration of the vocal tract given only acoustic data (Rudzicz 2010). Similar approaches have been shown to accurately estimate such positions within an average error of less than 1 mm (Richmond et al. 2003; Toda et al. 2008).

Data collection with individuals with dysarthria is ongoing according to the protocols described in this paper. All data described in this journal will be made available online in early 2011. The University of Toronto will not charge any fee for access to this data. Audio for each utterance is encoded in individual wave files encoded in the linear PCM format at 16 and 44.1 kHz. Both raw and normalized articulatory data are provided in EMA-format files at 200 Hz. Open-source programs will be provided to access the EMA file format; no proprietary EMA system or analysis software will be required to access this data. This database will occupy approximately 45 GB of disk space.

Acknowledgments The authors acknowledge the support of Toronto Rehabilitation Institute which receives funding under the Provincial Rehabilitation Research Program from the Ministry of Health and Long-Term Care in Ontario. The views expressed do not necessarily reflect those of the Ministry. Equipment and space have been funded with grants from the Canada Foundation for Innovation, Ontario Innovation Trust and the Ministry of Research and Innovation. This project is also funded by Bell University Labs, the University of Toronto, and the Natural Sciences and Engineering Research Council of Canada

Appendix: Articulatory contrasts

See Table 3.

Table 3 Articulatory contrasts, from Kent et al. (1989)

Front-back vowel	Knew/knee	Pat/pot	Him/hum	Shoot/sheet	Beet/boot
	Geese/goose	Feed/food	Air/are	Chop/chap	Fill/full
High-low vowel	Knew/know	Knew/gnaw	Him/hem	Him/ham	Shoot/shot
	Geese/gas	Geese/guess	Pit/pet	Pit/pat	Feet/fat
	Heat/hate	Had/hid			
Vowel duration	Beat/bit	Slip/sleep	Leak/lick	Knot/nut	Read/rid
	Ship/sheep	Feet/fit	Lip/leap	Ease/is	Reap/rip
Voicing, initial consonants	Pat/bat	Bad/pad	Pit/bit	Sip/zip	Coat/goat
	Dug/tug	Cash/gash	Tile/dial	Bunch/punch	
Voicing, final consonants	Feet/feed	Bad/bat	Leak/league	Knot/nod	Write/ride
	Side/sight	Coat/code	Dug/duck	Ate/aid	At/add
Alveolar-palatal	Sip/ship	Shoot/suit	Shy/sigh	Sell/shell	Sin/shin
	Sew/show	See/she	Sheet/seat		
Consonant place	Bug/dug	Tile/pile	Cake/take	Meat/neat	Bill/dill
	Bill/gill	Ache/ape	Ache/ate	Lip/lit	
Other fricative	Sheet/feet	Sigh/thigh	Hill/fill	Hand/sand	Sew/foe
	See/he	Nice/knife	Hat/fat	Sell/fell	Feet/heat
	Hat/that	Hold/fold	Hail/sail	Harm/farm	Seed/feed
Fricative-affricate	Chair/share	Wish/witch	Much/mush	Ship/chip	Chop/shop
	Cash/catch	Sheer/cheer	Hash/hatch	Harm/charm	
Stop-affricate	Chair/tear	Much/mut	Chop/top	Witch/wit	Much/muck
Stop-nasal	Beat/meat	Knot/dot	Side/sign	Nice/dice	Steak/snake
	Bill/mill	Dock/mock	Dock/knock	Bunch/munch	Tile/mile
Initial glottal-null	Air/hair	Ate/hate	At/hat	Hand/and	Hold/old
	Heat/eat	Hash/ash	Harm/arm	Had/add	Hail/ail
Initial consonant-null	Air/fair	Ate/fate	At/at	Sin/in	Sheet/eat
	Chair/air	Spit/it	Blend/end	Ease/peas	Ease/cheese
	Sink/ink	Cake/ache	Rise/eyes	Row/ow	
Final consonant-null	Feed/fee	Side/sigh	Blow/bloat	Fork/four	Rake/ray
	Leak/lee	Meat/me	Bunch/bun	Seed/see	
Initial cluster-initial singleton	Slip/sip	Slip/lip	Spit/pit	Spit/sit	Blend/bend
	Blend/lend	Sticks/six	Sticks/ticks	Steak/take	Steak/sake
	Blow/low	Blow/bow			
Final cluster-final singleton	Sticks/stick	Rock/rocks	Seed/seeds	Sink/sing	Cake/cakes
	Meat/meats	Fork/forks	Rake/rakes	Leak/leaks	Ache/aches
	Wax/wack	Docks/dock			
/r/-/l/	Read/lead	Write/light	Leak/reek	Rock/lock	Rake/lake
	Lip/rip	Reap/leap	Rise/lies	Row/low	Racks/lax
/r/-/w/	Read/weed	Write/white	Rich/witch	Rock/walk	Reap/weep
	Rise/wise	Row/woe	Racks/wax		

References

Aarabi, P., & Shi, G. (2004). Phase-based dual-microphone robust speech enhancement. *IEEE Transactions on Systems, Man, and Cybernetics Part B, 34*(4):1763–1773.

Bennett, J. W., van Lieshout, P., & Steele, C. M. (2007). Tongue control for speech and swallowing in healthy younger and older subjects. *International Journal of Orofacial Myology, 33*, 5–18.

Campbell, J. M., Bell, S. K., & Keith, L. K. (2001). Concurrent validity of the peabody picture vocabulary test-third edition as an intelligence and achievement screener for low SES African American children. *Assessment, 8*(1), 85–94.

Clear, J. H. (1993). The British national corpus. In: *The digital word: Text-based computing in the humanities* (pp. 163–187). Cambridge, MA: MIT Press.

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

Enderby, P. M. (1983). *Frenchay dysarthria assessment*. San Diego: College Hill Press.

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 33*(2), 443–445.

Hasegawa-Johnson, M., Gunderson, J., Perlman, A., & Huang, T. (2006). HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In: *Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP 2006)*, Vol. 3, pp. 1060–1063.

Herndon, R. M. (1997). *Handbook of neurologic rating scales* (1st ed.). New York: Demos Medical Publishing.

Hoole, P., & Zierdt, A. (2010). Five-dimensional articulography. In: B. Maassen & P. H. van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research* (Chap. 20, pp. 331–349). Oxford: Oxford University Press.

Hoole, P., Zierdt, A., & Geng, C. (2003) Beyond 2D in articulatory data acquisition and analysis. In: *Proceedings of the fifteenth international congress of phonetic sciences*, Barcelona, pp. 265–268.

Hosom, J. P., Kain, A. B., Mishra, T., van Santen, J. P. H., Fried-Oken, M., & Staehely, J. (2003). Intelligibility of modifications to dysarthric speech. In: *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP '03)*, Vol. 1, pp. 924–927.

Jayaram, G., & Abdelhamied, K. (1995) Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development, 32*(2), 162–169.

Kaburagi, T., Wakamiya, K., & Honda, M. (2005). Three-dimensional electromagnetic articulography: A measurement principle. *Journal of the Acoustical Society of America, 118*(1), 428–443.

Kain, A. B., Hosom, J. P., Niu, X., van Santen, J. P., Fried-Oken, M., & Staehely, J. (2007). Improving the intelligibility of dysarthric speech. *Speech Communication, 49*(9), 743–759.

Kent, R. D. (2000). Research on speech motor control and its disorders: A review and prospective. *Journal of Communication Disorders, 33*(5), 391–428.

Kent, R. D., & Rosen, K. (2004). Motor control perspectives on motor speech disorders. In: B. Maassen, R. Kent, H. Peters, P. V. Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (Chap. 12, pp 285–311). Oxford: Oxford University Press.

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

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., et al. (2008). Dysarthric speech database for universal access research. In: *Proceedings of the international conference on spoken language processing (Interspeech '08)*, Brisbane, Australia, pp. 1741–1744.

Kroos, C. (2008). Measurement accuracy in 3D electromagnetic articulography (Carstens AG500). In: *Proceedings of the 8th international seminar on speech production*, pp. 61–64.

Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., et al. (2007). Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop. In: *Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP 2007)*, Honolulu.

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

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing, 9*(5), 504–512.

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzjo, J. E., & Bunnell, H. (1996). The nemours database of dysarthric speech. In: *Proceedings of the fourth international conference on spoken language processing*, Philadelphia, PA, USA.

Patel, R. (2002). Prosodic control in severe dysarthria: Preserved ability to mark the question-statement contrast. *Journal of Speech, Language, and Hearing Research, 45*(5), 858–870.

Richmond, K., King, S., & Taylor, P. (2003). Modelling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language, 17*, 153–172.

Roy, N., Leeper, H. A., Blomgren, M., & Cameron, R. M. (2001). A description of phonetic, acoustic, and physiological changes associated with improved intelligibility in a speaker with spastic dysarthria. *American Journal of Speech-Language Pathology, 10*, 274–290.

Rudzicz, F. (2007). 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.

Rudzicz, F. (2009). Applying discretized articulatory knowledge to dysarthric speech. In: *Proceedings of the 2009 IEEE international conference on acoustics, speech, and signal processing (ICASSP 09)*, Taipei, Taiwan.

Rudzicz, F. (2010). Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics. In: *Proceedings of the 2010 IEEE international conference on acoustics, speech, and signal processing (ICASSP10)*, Dallas, TX.

Shi, G., Aarabi, P., & Jiang, H. (2007). Phase-based dual-microphone speech enhancement using a prior speech model. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(1), 109–118.

Toda, T., Black, A. W., & Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication, 50*(3), 215–227. doi: 10.1016/j.specom.2007.09.001.

Tsai, R. Y. (1987). A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation RA-3*(4), 323–344.

van Lieshout, P., Hulstijn, W., Alfonso, P. J., & Peters, H. F. (1997). Higher and lower order influences on the stability of the dynamic coupling between articulators. In: W. Hulstijn, H. F. Peters, & P. van Lieshout (Eds.), *Speech production: Motor control, brain research and fluency disorders* (pp. 161–170). Amsterdam: Elsevier Science Publishers.

van Lieshout, P., Merrick, G., & Goldstein, L. (2008). An articulatory phonology perspective on rhotic articulation problems: A descriptive case study. *Asia Pacific Journal of Speech, Language, and Hearing, 11*(4), 283–303.

Webber, S. G. (2005). Webber photo cards: Story starters.

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook*. Waisman Center on Mental Retardation & Human Development.

Wrench, A. (1999). The MOCHA-TIMIT articulatory database. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Yorkston, K. M., & Beukelman, D. R. (1981). *Assessment of intelligibility of dysarthric speech*. Tigard, OR: C.C. Publications Inc.

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research, 51*, 596–611.

Yunusova, Y., Green, J. R., & Mefferd, A. (2009) Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech, Language, and Hearing Research, 52*, 547–555.

Zierdt, A., Hoole, P., & Tillmann, H. G. (1999). Development of a system for three-dimensional fleshpoint measurement of speech movements. In: *Proceedings of the XIVth international congress of phonetic sciences*, p. 3.

Zierdt, A., Hoole, P., Honda, M., Kaburagi, T., & Tillmann, H. G. (2000). Extracting tongues from moving heads. In: *Proceedings of the 5th speech production seminar*, pp. 313–316.

Zue, V., Seneff, S., Glass, J. (1989). Speech database development: TIMIT and beyond. In: *Proceedings of ESCA tutorial and research workshop on speech input/output assessment and speech databases (SIOA-1989)*, Noordwijkerhout, The Netherlands, Vol. 2, pp. 35–40.