JSLHR

Article

Vocal Tract Representation in the Recognition of Cerebral Palsied Speech

Frank Rudzicz,[a] Graeme Hirst,[a] and Pascal van Lieshout[a,b,c]

Purpose: In this study, the authors explored articulatory information as a means of improving the recognition of dysarthric speech by machine.

Method: Data were derived chiefly from the TORGO database of dysarthric articulation (Rudzicz, Namasivayam, & Wolff, 2011) in which motions of various points in the vocal tract are measured during speech. In the 1st experiment, the authors provided a baseline model indicating a relatively low performance with traditional automatic speech recognition (ASR) using only acoustic data from dysarthric individuals. In the 2nd experiment, the authors used various measures of entropy (statistical disorder) to determine whether characteristics of dysarthric articulation can reduce uncertainty in features of dysarthric acoustics. These findings led to the 3rd experiment, in which recorded dysarthric articulation was directly encoded into the speech recognition process.

Results: The authors found that 18.3% of the statistical disorder in the acoustics of speakers with dysarthria can be removed if articulatory parameters are known. Using articulatory models reduces phoneme recognition errors relatively by up to 6% for speakers with dysarthria in speaker-dependent systems.

Conclusions: Articulatory knowledge is useful in reducing rates of error in ASR for speakers with dysarthria and in reducing statistical uncertainty of their acoustic signals. These findings may help to guide clinical decisions related to the use of ASR in the future.

Key Words: dysarthria, articulation, speech recognition

Dysarthria classifies a group of motor speech disorders that can be associated with "weakness, slowness, and/or lack of coordination of the speech musculature as the result of damage to the central or peripheral nervous system. Phonation, respiration, resonance, articulation, and prosody are affected. Movements may be impaired in force, timing, endurance, direction, and range of motion. Symptoms may include slurred speech, weak or imprecise articulatory contacts, weak respiratory support, low volume, incoordination of the respiratory stream, hypernasality, and reduced intelligibility" (American Speech-Language-Hearing Association, 2011, p. 38).

Dysarthria can be caused by genetic or traumatic factors and may occur at any age (including prenatal stages). Dysarthria differs from another motor speech disorder, known as *verbal apraxia,* in that the latter reduces a speaker's ability to plan speech movements, whereas dysarthria strictly reflects an inability to execute those movements and does not typically affect the regular comprehension or production of meaningful, syntactically correct language. Usually, dysarthrias are classified according to perceived speech symptoms and the assumed affected neurological systems causing these symptoms (Duffy, 1995). For example, damage to the recurrent laryngeal nerve introduces a weakness or paralysis of the intrinsic laryngeal musculature, which will affect the quality and intensity of voice signals. Inadequate control of velar movement caused by a lesion that would affect the pharyngeal branches of the vagus cranial nerve may lead to a disproportionate amount of air being released through the nose during speech (i.e., *hypernasality*). Lesions that more generally affect the ability to adequately control articulator movements will often produce heavily slurred speech and a more diffuse and less differentiable vowel target space (Kent & Rosen, 2004). Moreover, the lack of articulatory control often leads to various involuntary sounds caused by

[a]University of Toronto, Ontario, Canada
[b]Institute of Biomaterials and Biomedical Engineering, Toronto
[c]Toronto Rehabilitation Institute

Correspondence to Frank Rudzicz, who is affiliated with both the University of Toronto and Toronto Rehabilitation Institute: frank@cs.toronto.edu

Editor: Anne Smith
Associate Editor: Wolfram Ziegler

Received August 13, 2011
Revision received November 3, 2011
Accepted December 6, 2011
DOI: 10.1044/1092-4388(2011/11-0223)

velopharyngeal or glottal noise or by noisy swallowing problems (Rosen & Yampolsky, 2000).

Later-onset causes of dysarthria are typically traumatic, including cerebrovascular stroke, with the severity of impairment varying with the amount of cerebral damage (Rudzicz, 2011b). Other sources of dysarthria include multiple sclerosis, Parkinson's disease, myasthenia gravis (i.e., blocked acetylcholine receptors), and amyotrophic lateral sclerosis (Kent & Rosen, 2004). Childhood manifestations of dysarthric speech are often the result of trauma (e.g., stroke, asphyxia, infections) or congenital malformations of the brain, inhibiting normal development in the speech motor areas. *Cerebral palsy* (CP) is the most common physical impairment in pediatrics, affecting approximately 2.5 per 1,000 children in Western countries (Shevell, Miller, Scherer, Yager, & Fehlings, 2011), 88% of whom are dysarthric throughout adulthood (Augmentative Communication Inc., 2007). It is a nonprogressive symptom complex with a movement disorder as a central defining characteristic that is quite heterogeneous in terms of etiology, presentation (including the presence of learning disabilities and both cognitive and linguistic problems in many of these children), severity, and many other factors (Shevell et al., 2011). This is also reflected in the speech symptoms that, if present in certain individuals suffering from CP, may show different degrees and combinations of types of dysarthria.

By definition, dysarthria can have dramatic consequences for speech intelligibility among human listeners (Hustad, Gorton, & Lee, 2010). In this article, we focus on the problems experienced by artificial listeners—that is, speech recognition systems.

Background: Automatic Speech Recognition

Current automatic speech recognition (ASR) has produced tools for the general public that are in widespread use, but a high error rate in recognizing and adapting to dysarthric speech has kept such software effectively inaccessible to individuals with severe speech disorders. Depending on the severity of the speech disorder, as few as 3% of the words uttered by speakers with dysarthria might be recognized by traditional software in contexts in which approximately 85% of the words spoken by an individual without dysarthria might be recognized correctly (Rudzicz, 2007).

The goal of ASR, in general, is to decide on the optimal phoneme sequence $\rho_c = p_1 p_2 \ldots p_n$ that describes an acoustic input speech signal X:

$$\rho_c = \arg\max_{\rho} \frac{P(\rho)P(X|\rho)}{P(X)}, \quad (1)$$

where $P(\rho)$ and $P(X|\rho)$ are the language model and the acoustic model, respectively. For example, a particular sequence of phonemes, ρ_1, might have a very high lexical probability, $P(\rho_1)$, but might have a very low acoustic probability, $P(X|\rho_1)$. Naturally, given a particular acoustic observation, X, there are many possible phoneme sequences, ρ, that one can use to evaluate $P(\rho)P(X|\rho)$ (some of them are very unlikely), and it can be computationally expensive to consider them all. However, one can use various techniques to determine which phoneme sequences are the most likely—a process called *decoding*. The decoding process essentially classifies an observed input by assigning a particular sequence of phonemes to it.

The process of deciding the optimal phoneme sequence depends on probability distributions $P(\rho)$ and $P(X|\rho)$ that are representative of realistic data. Therefore, a crucial component of speech recognition is the automatic adjustment of those probabilities given large amounts of sample data, a process called *machine learning* or *automatic training*. In this work, as in most work with ASR, the observable sequences, X, are sequences of vectors of mel-frequency cepstral coefficients, each representing subsequent windows of approximately 20 ms of speech. To compute these representative features of speech, we use the first 13 cepstral dimensions, the total log energy, and their first-order (δ) and second-order ($\delta\delta$) derivatives, giving sequences of T 42-dimensional real-valued observations, $O = o_1 o_2 \ldots o_T$, for variable T.

The purpose of this study was to understand to what extent relevant articulatory data can instruct us on how to manage the acoustic behaviors of speakers with dysarthria, particularly in specialized ASR. The standard ASR configuration is explored in Experiment 1, in which baseline recognition results were obtained with systems that use only acoustic input. Experiment 2 was an attempt to characterize data from speakers with dysarthria in terms of alternative representations—for example, distortions of underlying task dynamics (Saltzman & Munhall, 1989)—and to analyze the entropy (statistical disorder) in dysarthric acoustics and articulation. If measurements of the vocal tract of speakers with dysarthria are unavailable or sparse, then characterizing (or synthesizing) such data as a distortion of other representations for which data are available would be useful in training or initializing ASR systems. A reduction in entropy in the acoustics given articulatory information provides a theoretical argument for incorporating vocal tract data in ASR. In Experiment 3, we provided an empirical argument of such incorporation by describing models that are initialized with explicit articulatory measurements. In the experiments described in this article, we used a new database called TORGO (Rudzicz, Namasivayam, & Wolff, 2011), which consists of dysarthric speech including articulation data acquired using three-dimensional (3D) video and articulography.

Materials

Because speakers with dysarthria are especially susceptible to fatigue, collecting data from this population can be challenging. The Alfred I. duPont Institute's Nemours database (Menéndez-Pidal, Polikoff, Peters, Leonzjo, & Bunnell, 1996) consists of 11 male participants with dysarthria who have varying degrees of intelligibility and one male participant without dysarthria. Each participant uttered 74 syntactically invariant and semantically meaningless short sentences and two additional paragraphs designed to provide closed-set phonetic contrasts (e.g., place, manner, voicing; Menéndez-Pidal et al., 1996). The Nemours database includes assessments of each speaker's motor function by a speech-language pathologist (SLP), but it does not include any articulatory data.

The University of Edinburgh's MOCHA database (Wrench, 1999) consists of 460 sentences uttered by one male British speaker and one female British speaker, both of whom do not have dysarthria. All acoustic data were temporally aligned with electromagnetic articulography (EMA; recorded at 500 Hz), laryngography (at 16 kHz), and electropalatography (at 200 Hz). The EMA data consist of bivariate positional information from eight articulatory locations—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. Each parameter was measured in the two dimensions of the midsagittal plane.

The TORGO database of dysarthric articulation consists of aligned acoustic and articulatory measurements from eight individuals (five male, three female) with dysarthria (Rudzicz et al., 2011). The data used in the following experiments are from participants with CP (spastic, athetoid, or ataxic) who are between the ages of 16 and 50 years. These individuals were matched according to age and gender with control participants from the general population. Each participant recorded 3 hr of data. In that time, speakers with dysarthria recorded approximately 500 utterances, on average, and control speakers recorded approximately 1,000 utterances, on average. The motor functions of each participant in the TORGO database were assessed according to the standardized Frenchay Dysarthria Assessment (Enderby, 1983) by an SLP affiliated with Holland Bloorview Kids Rehabilitation Hospital and the University of Toronto.

Individual prompts were presented to participants in random order. They were derived from nonwords (e.g., to collect information on plosive consonants in the presence of high and low vowels; Bennett, van Lieshout, & Steele, 2007), short words (e.g., contrasting pairs from Kent, Weismer, Kent, & Rosenbek, 1989), and restricted sentences (e.g., the sentence intelligibility section of the Yorkston–Beukelman Assessment [Yorkston & Beukelman, 1981] and sentences used in the MOCHA database [Wrench, 1999]).

We collected the EMA data using the 3D Carstens Medizinelektronik AG500 system (Zierdt, Hoole, & Tillmann, 1999; van Lieshout, Merrick, & Goldstein, 2008). Sensors were attached to three points on the surface of the tongue—namely, the tongue tip (1 cm behind the anatomical tongue tip), the tongue middle (3 cm behind the tongue tip coil), and the tongue back (approximately 2 cm behind the tongue middle coil). A sensor for tracking jaw movements was attached to a custom mold that fits the surface of the lower incisors as described by van Lieshout and Moussa (2000). Four additional coils were placed on the upper and lower lips and the left and right corners of the mouth. Further coils were placed on the participant's forehead, nose bridge, and behind each ear above the mastoid bone for reference purposes. Except for the left and right mouth corners, all sensors that measure the vocal tract were generally on the midsagittal plane, on which much of the relevant motion of speech takes place.

Aspects of Dysarthric Speech in TORGO

A number of features differentiate dysarthric and nondysarthric speech in our recorded data. Table 1 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, 5% of corresponding plosives in total are mispronounced in nondysarthric 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 nondysarthric speech data.

Table 1. 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 (%)		
Source	**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

Figure 1 shows the durations of various vowels averaged across speakers from the dysarthria and control groups of TORGO. All vowels produced by speakers with dysarthria are significantly slower than those produced by speakers without dysarthria at the 95% confidence interval and can be up to twice as long, on average. This might partially be explained by an increase of brief gaps in exhalation during sonorants. The vowels produced by speakers with dysarthria have variances that are similar to those of their counterparts without dysarthria.

Experiment 1: Hidden Markov Model (HMM) Baselines

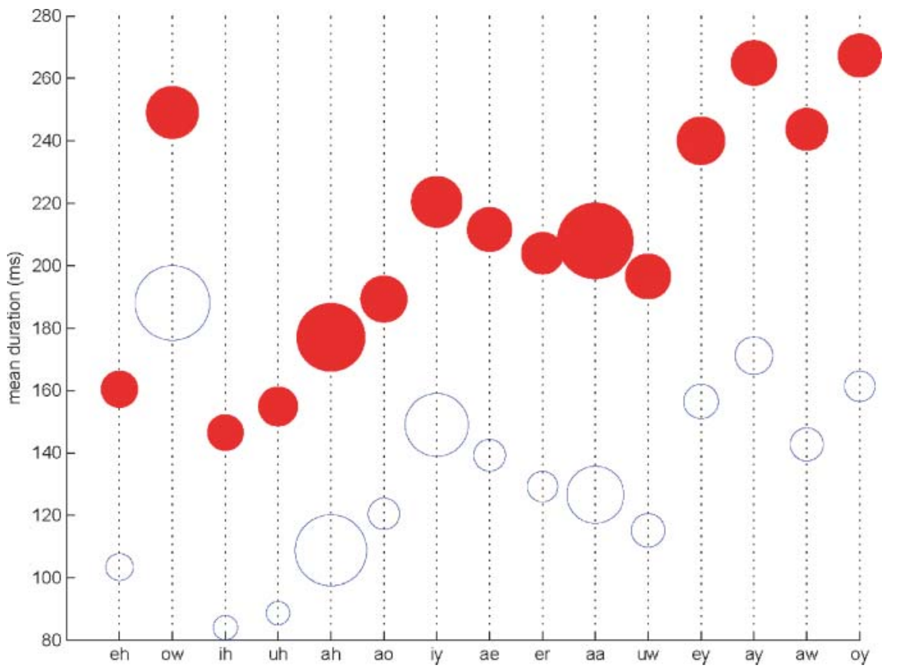
The most common model used in ASR is the *hidden Markov model* (HMM), which categorizes observable temporal data sequences according to hidden (i.e., unobservable) statistical parameters and an underlying connected-state structure. In speech recognition research, the unobservable state is typically an unspecified high-level abstraction of a relatively invariant section of speech whose distribution is described by a weighted mixture of Gaussian components, the number of which is a manually specified parameter that we can empirically alter. Mathematical details are provided in the Appendix.

In our study, each phoneme was represented by its own HMM. Three states indicated whether the phoneme was in its initial phase, its relatively steady central phase, or its final phase. This tri-state model allowed for variation that can occur at the beginning and end of phonemes, given their context. These states were hidden because one cannot directly observe whether the speaker is producing these segments in their initial, central, or final phase—this is an approximation of the underlying dynamics.

Finding the state sequence having the highest probability given an observation sequence is tantamount to finding the most likely phoneme sequence, assuming that phoneme HMMs are concatenated to form words. The Viterbi algorithm is used to determine the most likely state sequence that represents an observation, given a set of an HMM's parameters (Huang, Acero, & Hon, 2001).

Alternatives to HMMs for speech recognition have been explored in previous work (Rudzicz, 2011a). Some of these alternatives, including neural networks and support vector machines, are *discriminative* models in that their parameters are adjusted to minimize the expected rates of error on unseen test material. The HMM, by contrast, is *generative* in that its parameters are adjusted so that unseen test material obtains the highest expected likelihood. One practical advantage that the HMM has over other mechanisms is that it explicitly encodes the notion of temporal dynamics into its parameters, which is useful for

Figure 1. Duration of vowels among dysarthric (filled circles) and control (unfilled circles) speakers. Circle positions correspond to the average duration (in ms) of the associated vowel, and the radii of the circles represent 1 *SD* of the data.



speech data. It is also still the most common mechanism used for commercial ASR and for ASR research (Lamere et al., 2003; Vertanen, 2006). We studied one generative alternative to the HMM, namely, the dynamic Bayes network (DBN), in Experiment 3.

Method: Maximum a Posteriori and Maximum Liklihood Linear Regression Adaptation

When the statistics of HMMs are trained solely with data from an individual speaker, they constitute *speaker-dependent* models, which are theoretically resilient to any loss of precision caused by interspeaker variation. However, when these models are trained using data from several (usually many) speakers, they constitute *speaker-independent* models, which are less sensitive to individual characteristics (due to interspeaker variation) but can be trained with more data and can therefore capture the most common aspects of speech.

Adaptation of model parameters is used when the conditions in which those parameters were trained no longer reflect the conditions in which new data will be observed. For example, if a model is trained in a quiet environment but will be used in a noisy one, model parameters must be adjusted to reflect this new environment using a small amount of calibration data, which typically is much smaller in scope than the original training data set. HMM adaptation is employed here using a combination of two standard techniques—namely, *maximum a posteriori (MAP) estimation* and *maximum likelihood linear regression* (*MLLR*). These techniques are described in detail in the second section of the Appendix.

This process of MAP followed by MLLR constitutes a *speaker-adaptive* system. It is iterative and can be considered as an interpolation between speaker-dependent and speaker-independent models, compromising between the advantages and drawbacks of each approach (Huang et al., 2001). Dysarthric speech is atypical, so it is important to study the benefits and limitations of both dependent and adaptive models. For instance, speakers with a limited but very uncommon range of acoustic characteristics may, in theory, be more accurately represented by dependent models, whereas those with more typical vocal characteristics but limited available data (or diffuse data) may be more accurately represented by adaptive models in which data from other speakers fill in the gaps. The relative merits of adaptive versus dependent models for dysarthric speech is still a matter of ongoing research (Raghavendra, Rosengren, & Hunnicutt, 2001; Sharma & Hasegawa-Johnson, 2010). Although speaker adaptation typically involves more data than speaker dependence, the additional data are derived from other speakers whose patterns of speech may vary significantly from those of the target speaker and are not used during the adaptation phase.

Comparing speaker-independent, speaker-adaptive, and speaker-dependent models with dysarthric speech indicates to what extent ASR will be useful for these speakers using traditional acoustic-only models. This also provides the baseline against which we can compare results in Experiment 3. Splitting analysis according to severity is relevant to clinicians who work with a spectrum of vocal capabilities.

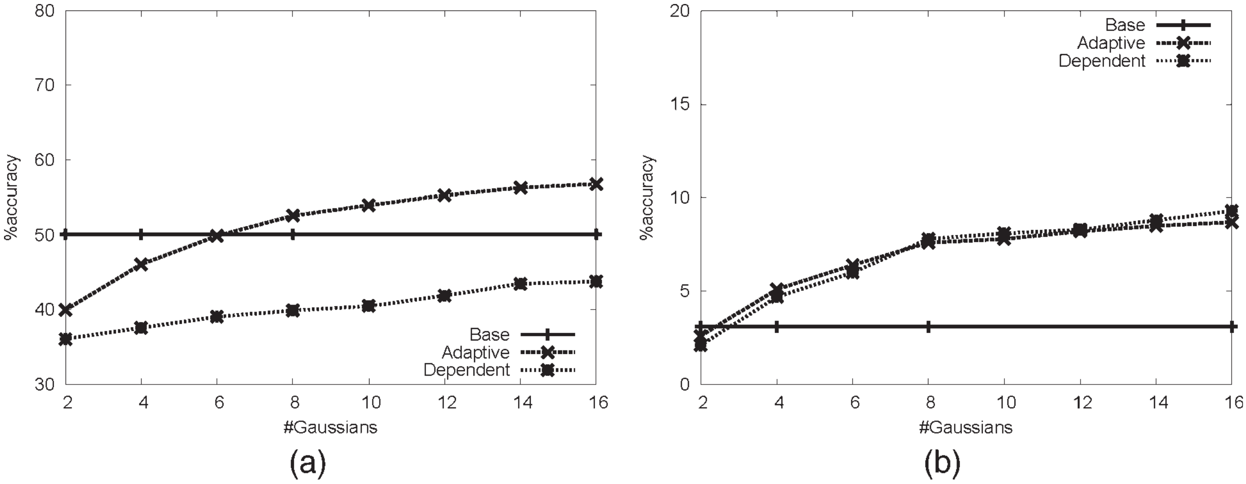
Results

We categorized each of the 10 speakers with dysarthria in the Nemours database according to the speakers' recognition rate using a baseline HMM with acoustic input only. This acoustic model is trained with approximately 211 hr of spoken transcripts of *The Wall Street Journal (WSJ)* from more than 100 individuals with typically developing speech (Lamere et al., 2003; Vertanen, 2006). Because this model encapsulates the statistics of a large amount of speech from the general population, it is a speaker-independent model that is not necessarily representative of dysarthric speech. The four speakers having word-level recognition rates below 10% with the baseline model were grouped as *severe,* the four speakers with rates between 11% and 30% were grouped as *moderate,* and the three speakers with rates between 31% and 60% were grouped as *mild.* The word-level recognition rate with the control speaker was 84.8%.

Both the dependent and adaptive models for each speaker were tri-state left–right HMMs. For each speaker, we initialized the HMM acoustic parameters of the dependent model randomly and initialized the adaptive model with the common *WSJ*-trained baseline. As an experimental parameter, we independently varied the number of Gaussians (M in Equation 7 of the Appendix) and the amount of training utterances in order to measure how precision and data coverage accommodate the variability of dysarthric speech; we applied the iterative Baum-Welch training algorithm on both models for each speaker. Language models consisted of combined bigram (i.e., two-word) and unigram (i.e., one-word) statistics gathered with maximum likelihood estimation over the textual representation of the *WSJ* corpus and were not adjusted iteratively during HMM training. Evaluation and training materials consisted of 74 syntactically invariant sentences with the template<<*The (NOUN) is (GERUND VERB) the (NOUN).*>>, where nouns and verbs were selected without replacement.

Figure 2 shows word-level accuracy increasing monotonically with the number of Gaussians for the

Figure 2. Automatic speech recognition accuracy measured against acoustic model precision (i.e., number of Gaussians). Baselines represent models trained on the *Wall Street Journal* corpus. We did not alter the parameters of the baseline model, hence the constant accuracy; we provide their performance here for comparison.



speakers with mild and severe dysarthria. In all cases, word accuracy is determined as the proportion of words correctly recognized over the total number of words in the true orthography, as determined by the Levenshtein alignment algorithm. Reported results are averages over accuracies within the specified group, as determined above, and across a tenfold cross-validation. The horizontal bars denote average baseline recognition accuracy obtained with the speaker-independent model averaged across the speakers in the respective groups. In all cases but the most severe, the adaptive models outperformed their dependent counterparts and reduced relative error by up to 23.1% for the mild group, by 4.9% for the moderate group, and by 30.7% for the control speaker. This suggests that taking advantage of preexisting models on the basis of speech samples from individuals who do not have dysarthria may best suit speakers with dysarthria who have higher intelligibility scores. By contrast, there are only slight gains over the baseline as the number of Gaussians increases for speaker-dependent models, as seen for the speakers with severe dysarthria.

Carnegie Mellon's Sphinx system allows the recognition process to decompose its output into a string of phonemes, effectively ignoring word order in the generation of this hypothesis (Lamere et al., 2003). Using this method, we can analyze the more low-level sources of confusion that may confound the higher-level classification—namely, in terms of phoneme insertion, deletion, and substitution errors made by the ASR system. Of the 485 insertion errors made by the model with dysarthric speech, /*ih*/ and /*d*/ are the most common, with 63 and 51 errors, respectively. The most commonly dropped phonemes by these speakers are /*b*/ (118), /*s*/ (111), /*w*/ (60), /*f*/ (55), and /*l*/ (48), among 649 deletion errors in total. The most common substitutions are /*ng*/ for /*n*/ (125) and, surprisingly, /*t*/ for /*uw*/ (87), /*ey*/ for /*ih*/ (84), and /*t*/ for /*n*/ (77). These observations suggest that ASR software might be made more accessible to speakers with dysarthria by increasing robustness against consonant variations in general.

Experiment 2: A Noisy-Channel Model of Dysarthria

Dysarthria is sometimes characterized as a distortion of parallel biological pathways that corrupt motor signals before execution (Freund, Jeannerod, Hallett, & Leiguarda, 2005). This type of degradation is characteristic of the *noisy-channel model,* in which signals are distorted according to known statistics. Therefore, in this experiment, the speech–motor interface was cast within the framework of the noisy-channel model. First, we asked whether the incorporation of articulatory data was theoretically useful in reducing uncertainty in dysarthric acoustics.

Second, we asked which of two alternative noisy-channel models best described the observed variations in dysarthric speech. This work was based on Rudzicz (2010c). In this experiment, the author used TORGO data from three speakers with dysarthria who have CP (male participants M01 and M04, and female participant F03) as well as their age- and gender-matched counterparts from the general population (male participants

MC01 and MC03, and female participant FC02). This subset of the data from speakers with dysarthria was used because, at the time of these experiments, it contained the most phonemic annotation, which is a necessary precondition to some of our analyses. Experiments were restricted to 100 phrases uttered in common by all six speakers.

Method: Entropy

We wanted to measure the degree of statistical disorder in both acoustic and articulatory data for speakers with and without dysarthria as well as the posterior disorder of one type of data given the other. These measurements were made on the data directly, rather than on the models trained in Experiment 1, in order to obtain a representative characterization of the observation spaces themselves. This quantification would inform us as to the relative merits of incorporating knowledge of articulatory behavior into ASR systems for speakers with dysarthria. *Entropy* (also called *Shannon entropy*), $H(X)$, is a measure from information theory of the degree of uncertainty in a random variable, X. When observed values are continuous—as they are in our acoustic and articulatory database—we must use *differential entropy,* defined by

$$H(X) = -\int_X f(X) \log\, f(X) dX, \qquad (2)$$

where $f(X)$ is the probability density function of X. For a number of distributions, $f(X)$, the differential entropy has known forms (Lazo & Rathie, 1978). However, because both acoustic and articulatory data follow non-Gaussian distributions, these spaces must be represented by mixtures of Gaussians. Huber, Bailey, Durrant-Whyte, and Hanebeck (2008) developed an accurate algorithm for estimating differential entropy of Gaussian mixtures on the basis of iteratively merging Gaussians and the approximate upper bound of the entropy,

$$\tilde{H}(X) = \sum_{i=1}^{L} \omega_i \Big(-\log \omega_i + \frac{1}{2}\log((2\pi e)^N |\textstyle\sum_i|) \Big), \qquad (3)$$

where ω_i is the weight of the ith ($1 \le i \le L$) Gaussian, and $\sum_i$ is that Gaussian's covariance matrix. This method was used to approximate entropies here, with $L = 32$ being determined empirically. Although differential entropies can be negative and not invariant under change of variables, other properties of entropy are retained (Huber et al., 2008), such as the chain rule for conditional entropy, $H(Y|X) = H(Y, X) - H(X)$, which describes the uncertainty in Y given knowledge of X, and the chain rule for mutual information, $I(Y; X) = H(X) + H(Y) - H(X, Y)$, which describes the mutual dependence between X and Y. Here, we quantized entropy with the *nat,* which is the natural logarithmic unit (≈ 1.44 bits).

Our purpose in measuring the statistical disorder in our data was to provide a theoretical justification for the use of the data in Experiment 3. Specifically, if articulatory measurements sufficiently reduce the entropy in the acoustics, ASR systems that encode the former when recognizing words in the latter should be more accurate.

Results: Entropy

We measured the differential entropy of acoustics [$H(Ac)$], of articulation [$H(Ar)$], and of acoustics given knowledge of the vocal tract [$H(Ac\,|\,Ar)$] in order to obtain theoretical estimates of the utility of articulatory data. Table 2 shows these quantities across the six speakers in this study. As expected, the acoustics of speakers with dysarthria were much more statistically disordered than for speakers without dysarthria. One unexpected finding is that there was very little difference between speakers in terms of their entropy of articulation. Although speakers with dysarthria clearly lack articulatory dexterity, the comparable statistical disorder implies that they nonetheless articulate with a level of consistency similar to that of their counterparts without dysarthria.[1] However, the equivocation $H(Ac\,|\,Ar)$ is 1 order-of-magnitude lower for speakers without dysarthria. This implies that there is very little ambiguity left in the acoustics of speakers without dysarthria, given simultaneous knowledge of the vocal tract, but that quite a bit of ambiguity remains for our speakers with dysarthria, despite significant reductions. Further investigation should confirm the causes of this remnant ambiguity. Potential sources include unmeasured interaction between articulators and unmeasured lateral asymmetry in the tongue.

Table 3 shows the average mutual information between acoustics and articulation for each type of speaker, given knowledge of the manner of articulation. Table 1 shows a prevalence of pronunciation errors among speakers with dysarthria for plosives, but Table 3 shows no particularly low congruity between acoustics and articulation for this type of phoneme. Those pronunciation errors tended to be voicing errors, which would involve glottal activity—something that was not measured in this study.

In Table 3, it appears that there is little mutual information between acoustics and articulation in vowels across all speakers. However, this is almost certainly the result of our exclusion of tongue-blade and tongue-dorsum measurements[2] in order to standardize across

[1] This is borne out in the literature (Kent & Rosen, 2004).

[2] We retained the tongue-tip, jaw, and four lip measurements.

Table 2. Differential entropy in nats across speakers in the dysarthria group and the control group for acoustics (*Ac*), articulation (*Ar*), and acoustics given articulation (Ac|Ar).

Variable	*H*(*Ac*)	*H*(*Ar*)	*H*(*Ac*/*Ar*)
Dysarthria group			
M01	66.37	17.16	50.30
M04	33.36	11.31	26.25
F03	42.28	19.33	39.47
Average	*47.34*	*15.93*	*38.68*
Control group			
MC01	24.40	21.49	1.14
MC03	18.63	18.34	3.93
FC02	16.12	15.97	3.11
Average	*19.72*	*18.60*	*2.73*

Note. $H(x)$ = entropy; $H(x|x)$ = conditional entropy.

the speakers who could not accept these sensors. Indeed, the configuration of the entire tongue is known to be useful in discriminating among the vowels (O'Shaughnessy, 2000). An ad hoc analysis including all three tongue sensors for Speakers F03, MC01, MC03, and FC02 revealed mutual information between acoustics and articulation of 16.81 nats for F03 and 18.73 nats for the control speakers, for vowels. We compared this with mutual information of 11.82 nats for F03 and 13.88 nats for the control speakers across all other manners of articulation. The trend is that acoustics are better predicted given more tongue measurements, as expected (Mefferd & Green, 2010).

Method: The Noisy Channel

In the noisy-channel model, a signal $x \in X$ is distorted by a medium that transmits signal $y \in Y$ according to some distribution $P(Y|X)$. Given some probability that the received signal y is corrupted, the message produced by the decoder may differ from the original message (Shannon, 1949). To what extent can the effects of dysarthria be described within an information-theoretic noisy-channel model? We pursued two competing hypotheses. The first hypothesis modeled the assumption that dysarthric speech is a distorted version of typical speech. Here, signals X and Y represent the vocal characteristics of the general population and the population with dysarthria, respectively, and $P(Y|X)$ models the distortion between them. The second hypothesis modeled the assumption that both dysarthric and typical speech are different distorted versions of some common abstraction. Here, Y_d and Y_c represent the vocal characteristics of speakers with and without dysarthria, respectively, and X represents a common, underlying representation and that $P(Yd|X)$ and $P(Yc|X)$ model distortions from that representation. These two hypotheses are visualized in Figure 3. In each of these cases, signals can be acoustic, articulatory, or some combination thereof.

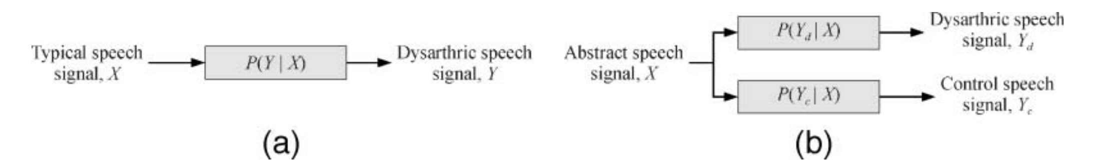
To test the hypothesis that both dysarthric and control speech are (different types of) distortions of a common abstraction of the vocal tract, we incorporated the theory of *task dynamics,* in which the dynamic patterns of speech are represented as the result of overlapping *gestures,* which are high-level reconfigurations of the vocal tract such as bilabial closure or velar opening (Saltzman, 1986). The open-source TADA system (Nam & Goldstein, 2006) estimates the positions of various articulators during speech according to parameters that have been carefully tuned by the authors of TADA according to a generic, speaker-independent representation of the vocal tract (Saltzman & Munhall, 1989). Given a word sequence and a syllable-to-gesture dictionary, TADA produces the continuous-tract variable paths that are necessary to produce that sequence. This takes into account various physiological aspects of human speech production, such as interarticulator coordination and timing (Nam & Saltzman, 2003).

We used TADA to produce estimates of a global, high-level representation of speech common to speakers both with and without dysarthria. Given a word sequence uttered by both groups, TADA produced five continuous curves prescribed by that word sequence in order to match our available EMA data. Those curves were lip aperture and protrusion, tongue-tip constriction location and degree (representing front–back and top-down positions of the tongue

Table 3. Mutual information *I*(*Ac*; *Ar*) of *Ac* and *Ar* for speakers with dysarthria and for control participants, across phonological manner of articulation.

Manner of articulation	Plosives	Affricates	Fricatives	Nasals	Glides	Vowels
I(*Ac*; *Ar*)						
Speakers with dysarthria	10.92	8.71	9.30	13.29	11.92	6.76
Control participants	16.47	9.23	10.94	15.10	12.68	7.15

Figure 3. Sections of alternative noisy channel models for the neuromotor interface in speakers with dysarthria.


Typical speech signal, X
P(Y | X)
Dysarthric speech signal, Y
(a)
Abstract speech signal, X
P(Y_d | X)
Dysarthric speech signal, Y_d
P(Y_c | X)
Control speech signal, Y_c
(b)

tip, respectively), and lower incisor height. These curves are then compared against observed EMA data, as described below.

Results: The Noisy Channel

To compare the scenarios in Figure 3, we designed a transformation system that produced the most likely observation in one data space given its counterpart in another. This transformation in effect implements the noisy channel itself. To accomplish this, probability distributions were automatically computed given our EMA data. First, we pooled all dysarthric data together, and we did the same for nondysarthric data. We then considered the acoustic (*Ac*) and articulatory (*Ar*) subsets of these data. In each case, Gaussian mixtures, each with 60 components, were trained over 90% of the data in both dysarthric and nondysarthric speech. Here, each of the 60 phonemes in the data is represented by one Gaussian component, with the weight of that component determined by the relative proportion of 10-ms frames for that phoneme. Similarly, all training word sequences were passed to TADA, and a mixture of Gaussians was trained on its articulatory output. As with our experiments with entropy, these measurements were made on the data directly rather than on the HMMs trained in Experiment 1.

Across all Gaussian mixtures, five types of Gaussians were tuned to various aspects of each phoneme p: its dysarthric acoustics and articulation, $\mathbf{N}_p^{Ac}(\mathrm{Y}_d)$ and $\mathbf{N}_p^{Ar}(\mathrm{Y}_d)$; its nondysarthric acoustics and articulation, $\mathbf{N}_p^{Ac}(\mathrm{Y}_c)$ and $\mathbf{N}_p^{Ar}(\mathrm{Y}_c)$; and its prescribed articulation from TADA, $\mathbf{N}_p^{Ar}(\mathrm{X})$. Each Gaussian, $\mathbf{N}_p^{A}(B)$, is represented by its mean $\mu_p^{(A,B)}$ and its covariance, $\sum_p^{(A,B)}$. Furthermore, the cross-covariance matrix is computed between Gaussians for a given phoneme; for example, $\sum_p^{(Ac,Y_c)(Ac,Y_d)}$ is the cross-covariance matrix of the acoustics of the control (Y_c) and dysarthric (Y_d) speech for phoneme p. Associating each phoneme with its own Gaussian allowed us to build in some useful prior categorical knowledge, but this technique was used mainly to reliably compute cross-covariance matrices $\sum_p^{(Ac,Y_c)(Ac,Y_d)}$, which required pairing each component in the Gaussian mixture that models all control acoustics with a component in the Gaussian mixture that models all dysarthric acoustics. Given these parameters, the most likely frame in one domain is estimated given its counterpart in another. For example, given a frame of acoustics from a control speaker y_c, this approach can synthesize the most likely frame of acoustics for a speaker with dysarthria, y_d, given an application of the noisy channel proposed by Hosom et al. (2003) using

$$\begin{aligned} f_{\mathrm{Ac}}(y_c) &= \boldsymbol{E}(y_d|y_c) \\ &= \sum_{i=1}^{P} h_i(y_c)\left[\mu_i^{(A_c,Y_d)} + \textstyle\sum_i^{(Ac,Y_c)(Ac,Y_d)} \times \left(\textstyle\sum_i^{(Ac,Y_c)}\right)^{-1} \times \left(y_c - \mu_i^{(Ac,Y_c)}\right)\right], \end{aligned} \tag{4}$$

where

$$h_i(y_c) = \frac{\alpha_i N\left(y_c; \mu_i^{(Ac,Y_c)}, \sum_i^{(Ac,Y_c)}\right)}{\sum_{j=1}^{P} \alpha_j N\left(y_c; \mu_j^{(Ac,Y_c)}, \sum_j^{(Ac,Y_c)}\right)}, \tag{5}$$

and α_p is the proportion of the frames of phoneme p in the data. Transforming between different sources of data was accomplished merely by substituting with the appropriate Gaussians.

We then measured how closely the transformed data spaces matched their true target spaces. In each case, test utterances (recorded or synthesized with TADA) were transformed according to functions learned in training (i.e., the remaining 10% of the data were used for each speaker type). We then compared these transformed spaces against their target space in our data. Table 4

Table 4. Average phoneme-level Kullback–Leibler (KL) divergences of acoustic and articulatory spaces given transformed and original control and dysarthric models, weighted by the relative proportions of the phoneme.

		KL divergence (10–2 nats)	
Type 1	Type 2	Acoustic	Articulatory
Control	Dysarthric	25.36	3.23
Control → Dysarthric	Dysarthric	17.78	2.11
TADA → Control	Control	N/A	1.69
TADA → Dysarthric	Dysarthric	N/A	1.84

Note. See Goldberger et al. (2003).

shows the Gaussian mixture phoneme-level Kullback–Leibler divergences given various types of source and target data, weighted by the relative proportions of the phonemes. It is not generally tractable to compute Kullback–Leibler directly for arbitrary pairs of Gaussian mixtures; however, several methods approximate this measure. We implemented the technique proposed by Goldberger, Gordon, and Greenspan (2003), which takes advantage of a known matching function between com ponents across mixtures. Each pair of N-dimensional Gaussians (N_i with mean μ_i and covariance $\sum_i$) for a given phoneme and data type was compared with

$$D_{KL}(\mathbf{N}_0||\mathbf{N}_1) = \frac{1}{2}\left(\ln\left(\frac{|\sum_1|}{|\sum_0|}\right) + \mathrm{trace}\left(\sum_1^{-1}\sum_0\right) + (\mu_1 - \mu_0)^T \sum_1^{-1}(\mu_1 - \mu_0) - N\right), \quad (6)$$

where ln () is the natural logarithm, trace () is the sum of the elements on the main diagonal of the supplied square matrix, $\sum_1^{-1}$ is the inverse of the covariance matrix $\sum_1$, and $(\mu_1 - \mu_0)^T$ is the transpose of the difference between means μ_1 and μ_0. The baseline showed that the speech from the control participants and the speakers with dysarthria was far more similar in articulation than in acoustics, according to this measure. When our control data (both acoustic and articulatory) were transformed to match the dysarthric data, the result was predictably more similar to the latter than if the conversion had not taken place. This corresponds to the noisy-channel model seen in Figure 3a, in which dysarthric speech is modeled as a distortion of nondysarthric speech. However, when dysarthric and control speech are modeled as distortions of a common abstraction (i.e., task dynamics) as shown in Figure 3b, the resulting synthesized articulatory spaces are more similar to their respective observed data than the articulation predicted by the first noisy-channel model. Dysarthric articulation predicted by transformations from task-dynamics space differ significantly from those predicted by transformations from control EMA data at the 95% confidence interval. From a purely statistical perspective, this demonstrates that an abstract continuous representation of speech is a more effective basis for the analysis of dysarthric speech than one derived from nondysarthric data. This may be due to constraints in task dynamics that help restrict analysis in dysarthric data. That is, because EMA data derived from controls contain more noise than the abstract representations in TADA, it is simpler to train the latter noisy-channel model. In practice, if an accurate synthetic representation of dysarthric speech was required, then this would not mitigate the utility of the abstract model: The end result is all that is required. From a more clinical perspective, this observation implies that the effects of dysarthria may be more appropriately mitigated by considering them as a distortion of gestural goals rather than as a distortion of preferred acoustics. However, given the possible statistical effects associated with training, more work is required to answer this question.

Discussion: Entropy in Dysarthric Speech

We have considered the amount of statistical disorder (i.e., entropy) in both acoustic and articulatory data in speakers with and without dysarthria. The use of articulatory knowledge reduces the degree of this disorder significantly for speakers with dysarthria (relatively 18.3%), although far less so than for speakers without dysarthria (relatively 86.2%). In real-world applications, we are not likely to have access to measurements of the vocal tract; however, there are many approaches that estimate the configuration of the vocal tract given only acoustic data (Richmond, King, & Taylor, 2003; Toda, Black, & Tokuda, 2008), often to an average error of less than 1 mm. The generalizability of such work to new speakers (particularly those with dysarthria) without training is an open research question.

We have argued for noisy-channel models of the neuromotor interface, assuming that the pathway of motor command to motor activity is a linear sequence of dynamics. The biological reality is much more complicated. In particular, the pathway of verbal motor commands includes several sources of sensory feedback (Seikel, King, & Drumright, 2005) that modulate control parameters during speech (Gracco, 1995). These senses include exteroceptive stimuli (auditory and tactile) and interoceptive stimuli (in particular, proprioception and its kinesthetic sense; Seikel et al., 2005), the disruption of which can lead to a number of production changes. For instance, Abbs, Folkins, and Sivarajan (1976) showed that when conduction in the mandibular branches of the trigeminal nerve is blocked, the resulting speech has considerably more pronunciation errors, although it is generally intelligible. Barlow (1989) argued that the redundancy of sensory messages provides the necessary input to the motor planning stage, which relates abstract goals to motor activity in the cerebellum (Guenther & Perkell, 2004, described a speech motor production model based on similar principles). However, despite the fact that the noisy-channel models described here do not incorporate feedback, they do provide some clear insight into the possible underlying mechanisms of control in speakers both with and without dysarthria. In particular, the data seem to suggest that articulation in both groups can be derived from a common abstract gestural representation with a high degree of consistency, even though the nature of the distortion to these abstract representations differs for the two groups.

Acoustic signals, on the other hand, are less consistent in this respect; this finding highlights the nonlinear transformation of articulation to acoustics and the possible interference of other sources (e.g., vocalization, nasalization) on the acoustic spectrum.

Experiment 3: Incorporating EMA Data

Given the theoretical support in the previous experiment for the use of articulatory data in the reduction of acoustic uncertainty, Experiment 3 involved combining those sources into a single system for speech recognition and was based upon our previous work (see Rudzicz, 2009). Although it is impractical to perform articulography on each speaker to be modeled, kinematic models can be adapted to speakers for whom only acoustic data are available.

Traditional Bayes learning is a popular statistical framework that determines instantaneous and immutable conditional relationships between variables. *DBNs* are directed acyclic graphs connecting random variables that generalize stochastic Bayesian learning to time sequences. This temporal model generalizes the HMM, the Kalman filter, and many other statistical models (Murphy, 2002). In general, a DBN will have hidden variables; therefore, the likelihood of training data cannot be decomposed into a sum over individual variables, and one must use expectation-maximization to update DBN parameters, generalizing the approach used with HMMs. In this experiment, we compared DBNs augmented with articulatory data during training with HMMs.

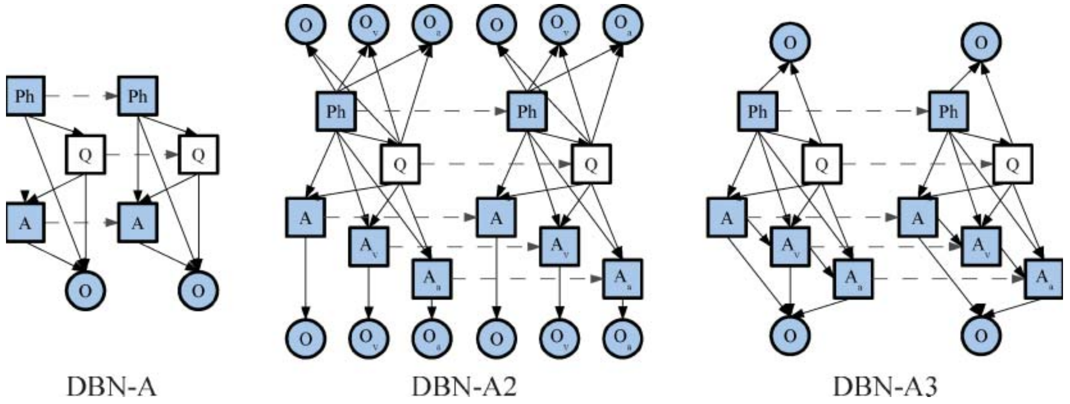
Results: DBNs

We conflated instantaneous EMA position data from the MOCHA and TORGO databases by first reducing their dimension from 16 to $N_p = 4$ or $Np = 8$ principal components by singular value decomposition specific to each phoneme in which $K = 4$, $K = 8$, or $K = 16$ mean vectors are computed according to the sum-of-squares error function. That is, we automatically found the N_p directions in the data along which variation is greatest (and therefore, most relevant), and found the K clearest clusters within that data. In a way, this allowed us to find the most characteristic discrete configurations of the data. MOCHA data were included here to ensure that the models we employed were not dependent on a single source of data. During training, the dynamic Bayes network variable A was the observed index of the mean vector nearest to the current frame of EMA data at time t. During inference, this variable was hidden, and we marginalized over all its values when computing the likelihood. In this way, DBN-A is essentially a DBN representation of an HMM with the hidden mixture index replaced by observed quantized articulation.We also applied this procedure to the velocities and accelerations of the articulators, producing indices A_v and A_a. We used these variables in alternative DBN topologies DBN-A2 and DBN-A3. In DBN-A2, the observation vector was trisected, with each 14-dimensional vector (i.e., mel-frequency cepstral coefficients, δ, and δδ) being conditioned on phoneme Ph, state Q, and one of A, A_v, and A_a. In a way, this modeled independence between position, velocity, and acceleration. The second alternative structure, DBN-A3, conditioned A_a on A_v, and A_v on A and conditioned the 42-dimensional observation vector on all variables. In general, this modeled dependence between position and velocity and between velocity and acceleration. The three kinematic DBN topologies are shown in Figure 4.

Recognition with nondysarthric speech. We compared the three DBN models on nondysarthric speech across the number of principal components, N_p, and the number of Gaussians K used in quantization. Previous research (Fukuda & Nitta, 2003; Wrench & Richmond, 2000) has shown that reducing dimensionality across heterogeneous acoustic/articulatory observations in this way preserves important features of both articulation and acoustics. Results of frame-level phoneme recognition are summarized in Table 5. Across all topologies, N_p = 16 is significantly more accurate than N_p = 8 at the 95% confidence level and N_p = 4 at the 99% confidence level. Results across MOCHA and TORGO and across the three topologies are statistically indistinguishable. However, both DBN-A2 and DBN-A3 are several times slower than DBN-A to train.

Retraining with dysarthric acoustics. We retrained models initialized on nondysarthric data given new dysarthric acoustics. That is, first we trained each kinematic DBN with nondysarthric acoustic/articulatory data (MOCHA and TORGO), then made indices A, A_v, and A_a and retrained on dysarthric acoustics (Nemours and TORGO). We trained all models with Expectation–Maximization and smoothed junction-tree inference, given their hidden variables. When retraining to dysarthric speech, we initialized new instantiations with the distributions learned on regular speech and retrained on speaker-specific acoustics until convergence. In all cases, training data included all phonemes observed during testing and were applied to the 46 phones that MOCHA, Nemours, and TORGO have in common. Data were randomly split into 90% training and 10% test data. Dysarthric TORGO and Nemours data were split by speaker into three categories according to the level of intelligibility as determined by the Frenchay assessment (Enderby, 1983). We designated individuals with intelligibility levels

Figure 4. Two-frame dynamic Bayes networks with electromagnetic articulography (EMA) measurements differing by their connectivity. Nodes *Ph, Q, O, A, A_v,* and A_a represent phoneme, state, mel-frequency cepstral coefficient observations, and EMA position, velocity, and acceleration, respectively. Interframe conditional links are dashed for clarity. DBN = dynamic Bayes network.



between 0% and 25% as "severe," between 25% and 62.5% as "moderate," and between 62.5% and 87.5% as "mild."

Table 6 shows the frame-level accuracy of unsegmented phoneme labeling on speaker-dependent and speaker-retrained distributions for each model, according to the severity of dysarthria. Here, DBN-A, DBN-A2, and DBN-A3 are trained to mixtures of 16 Gaussian clusters determined by unreduced (16-dimensional) articulatory data. In all cases, DBN models trained to the speech of the target speaker are more accurate than the equivalent speaker-dependent HMM, with significant relative error reductions of up to 2.68%, 4.85%, and 5.99% for severely, moderate, and mild speakers, respectively, relative to the HMM baseline. However, although retrained DBN models are typically more accurate than their speaker-dependent equivalents, in all cases, retrained HMM models are more accurate than all equivalent retrained DBN models. Data from Experiment 2 in Table 2 show that articulation removed far more uncertainty from nondysarthric speech than from dysarthric speech. There is a considerable amount of interspeaker variation in articulation, which may explain, in part, the success of retrained HMMs. In fact, the data in Table 6 confirm this hypothesis somewhat because there is a generally increasing benefit of retrained models over speaker-dependent models as intelligibility increases.

These results are generally consistent with similar work that retrained acoustic-only DBNs to Japanese kinematic data (Markov, Dang, & Nakamura, 2006) over one or two iterations of EM. That work showed an error reduction of between 0.7% and 3.8% on phoneme

Table 5. Accuracies of frame-level phoneme recognition across kinematic dynamic Bayes networks (DBNs) with varying quantities of principal components (*Np*) and Gaussians (*K*) for speaker-dependent, nondysarthric speech.

		DBN-A		DBN-A2		DBN-A3	
Np	*K*	MOCHA	TORGO	MOCHA	TORGO	MOCHA	TORGO
4	4	57.6	58.9	56.9	57.4	57.8	57.5
	8	66.8	67.2	66.5	67.2	66.8	67.1
	4	68.9	69.0	69.1	68.8	69.3	69.3
8	4	63.3	62.7	63.4	63.0	63.8	63.6
	8	71.0	70.8	71.1	71.3	71.3	71.6
	16	72.4	72.4	72.2	72.1	72.7	72.7
16	4	64.7	65.0	65.1	65.2	65.2	65.2
	8	72.5	72.6	72.4	72.4	72.7	72.5
	16	73.6	73.8	73.6	73.9	74.0	74.1

Table 6. Average accuracy (%) of correctly labeled phones of speaker-dependent and speaker-retrained EMA-initialized models, according to the severity of dysarthria.

Model	Severe	Moderate	Mild	Control
HMM				
Dependent	14.1	27.8	51.6	72.8
Retrained	16.8	32.1	58.9	
DBN-A				
Dependent	16.4	31.1	54.2	73.6
Retrained	16.2	31.7	58.3	
DBN-A2				
Dependent	16.3	31.1	54.3	73.6
Retrained	16.3	31.9	58.4	
DBN-A3				
Dependent	16.4	31.3	54.5	73.8
Retrained	16.5	32.0	58.7	

Note. EMA = electromagnetic articulography; HMM = hidden Markov model.

classification among a selection of alternative speaker-dependent DBNs relative to a baseline DBN.

The use of language models. Often, researchers use bigraphs to weigh the likelihood of transitioning from one phoneme or word to another. Because our data consist of many single-word utterances, we considered *phoneme bigraphs*, in which the conditional probability of one phoneme p_t following another p_{t-1} at time t is given by the total number of times in which p_t immediately follows p_{t-1} in the data over the total number of occurrences (i.e., whole sequences of frames) of p_{t-1}. We gathered these counts from the TIMIT database (Zue, Seneff, & Glass, 1989), which includes 2,472 unique bigraphs covering 172,460 adjacent pairs of phonemes, as determined by the included phonemic annotations. In a similar manner, we determined the unigraph probability of phoneme p_t from the same data by $P(p_t) = N_{(pt)} / \sum_{\rho} N_{(\rho)}$, where ρ was iterated over all 61 phonemes in the training data.

In order to implement systems that incorporate biographs or unigraphs, we trained individual HMM and DBN-A models for each phoneme, as before, where training data consisted of whole sequences of phonemes. The result was 61 HMMs and 61 DBN-A models, each consisting of three states with reflexive and left-to-right transitions. We connected the HMMs together and the DBN-As together by creating transitions from the last state of each phoneme model to the first state of all other phoneme models of the same type. The probabilities associated with these transitions were their bigraph probabilities. We then performed expectation-maximization for two iterations on each of the large connected HMM and DBN-A models in order to learn reflexive transition probabilities on the last state for each phoneme without overfitting. This is a common approach producing all-phoneme ergodic models (Miyazawa, 1993). We then repeated this process but with initial transition probabilities between phoneme models derived from their unigraph probabilities.

Given these connected models, we used the same data used to adapt to new acoustics to measure the average proportion of correctly labeled phones given phoneme models trained by the speaker-dependent method. Table 7 shows the frame-level phoneme recognition accuracies of each model across the same speaker intelligibility levels shown in Table 6. There are clear improvements in accuracy, but these improvements are still less than one would expect if full word-level bigrams were to be used, given more testing data.

Discussion: Incorporating EMA Data

In general, the results of Experiment 3 confirm the theoretical implications of Experiment 2 by demonstrating that statistical models that condition acoustic observations on their observed articulatory causes (i.e., the dynamic Bayes networks) are more accurate than acoustic-only models of speech production—but only when each of these models is trained to a particular speaker. When retrained acoustic–articulatory models were combined, the results were less impressive, possibly due to greater interspeaker variability at the articulatory level relative to the acoustic level. These results are expanded upon in work by Rudzicz (2011a), in which the articulatory DBN model was shown to be more accurate than several types of discriminative classifiers, including artificial neural networks that use only acoustic information. This indicates the utility of articulatory information, as discriminative classifiers often tend to be more accurate than generative models in categorizing sounds.

Table 7. Average frame-level accuracy (%) of unsegmented phoneme labeling given ergodic HMMs and DBN-As with unigraph and bigraph phoneme transition probabilities.

Level of severity	HMM		DBN-A	
	Unigraph	Biograph	Unigraph	Biograph
Severe	17.2	20.8	17.4	21.0
Moderate	33.4	37.3	34.1	37.9
Mild	60.1	63.5	60.5	63.7
Control	74.0	74.2	74.2	74.6

General Discussion and Conclusion

The purpose of this work was to determine whether an understanding of the articulatory sources of disorder in CP speech can instruct and improve automatic models of speech recognition. This work is based mainly on measurements in the new TORGO database of dysarthric articulation (Rudzicz et al., 2011). In Experiment 1, we explored adaptive and dependent speaker modeling techniques in HMMs given only acoustics, providing the baseline for what is possible with traditional technology. Here, speakers with severe dysarthria could expect no more than 10% of their words to be recognized by machine in a scenario in which a speaker with mild dysarthria might expect up to 60% of his or her words to be recognized. In Experiment 2, we examined whether articulatory data could theoretically be useful in reducing the ambiguities in the acoustics that can confound modern speech recognition systems. In fact, 18.3% of the statistical disorder in the acoustics can be removed if articulatory parameters are known, which is a significant reduction in entropy. This reduction in confusion was evident in Experiment 3, in which the rate of phoneme errors was relatively reduced by up to 6% for speakers with dysarthria in speaker-dependent systems that incorporated measured kinematics of the vocal tract. However, there is an increasing benefit of using retrained models rather than dependent models as dysarthric intelligibility increases. These results may guide clinicians and SLPs in ascertaining how to introduce alternative and augmentative communication systems on the basis of speech recognition to their clients.

This study represents an initial step toward a new type of speech recognition that incorporates long-term dynamics. Representing speech as a sequence of non-overlapping (though restricted) syllabic or phonemic units is the basis for ASR, and it has been useful in describing certain types of dysarthria in which speech is broken into syllables either due to respiratory problems or to improve overall intelligibility (Ziegler & Maassen, 2004). The work highlighted in this article demonstrates that articulatory measurements are both theoretically and practically useful in removing uncertainty and error from the acoustics of speakers with dysarthria who have CP. However, such models cannot inherently account for more complex aspects of articulatory organization, for which parallel and self-organizing theories may be more appropriate (Smith & Goffman, 2004; van Lieshout, 2004). In order to study the long-term dynamics of dysarthria in particular, we require a framework of dynamic systems in which our data can be analyzed.

Future work should be based on the study of dysarthric data within the framework of task dynamics, as introduced in Experiment 2. Indeed, the quantal theory of speech is based on the empirical observation that acoustics depend on a relatively discrete set of distinctive underlying articulatory configurations (Stevens & Keyser, 2010). Articulatory behavior of speakers with dysarthria should be compared against the behavior of control speakers by applying and extending methods that automatically compute the parameters of second-order differential equations with principal differential analysis (Ramsay & Silverman, 2005). In practice, however, several other aspects of task dynamics are not represented by its fundamental underlying spring-mass equations. For each speaker and each linguistic unit (i.e., syllable), several parameters can be derived. By adapting the parameters of this system (specifically, those that relate tract-variable positions to acoustics as discussed in Experiment 2) and repeating the experiments conducted in this study, researchers are currently attempting to measure the usefulness of task dynamics in speech recognition for speakers with dysarthria. Specifically, we are currently studying a new method for automatically reevaluating competing acoustic-based hypotheses by a speech recognizer according to the likelihoods of their articulatory realizations (Rudzicz, 2010b). In that work, we are measuring articulatory likelihood by evaluating continuous task-dynamics trajectories within probability distributions determined by acoustic–articulatory inversion (Rudzicz, 2010a).

The state-based models that are used in speech recognition (specifically, in HMMs and DBNs) provide convenient statistical representations that adapt to speech data for the purposes of deciphering future acoustic observations. However, these models are not necessarily representative of biological speech production (or perception). In order to more completely model the speech apparatus, future extensions to the dynamic Bayes model should incorporate some manner of feedback between the acoustics and the articulatory variables, for example. The results of Experiment 2 also suggest that explicit statistical relationships between some underlying control mechanism (task dynamics or otherwise) might be appropriate, although alternative relationships to those evaluated here should be considered. As future models are developed with these types of additional constraints, we expect accuracy to increase further, although there is not enough evidence to predict what a maximum rate of accuracy might be for these populations nor to what degree such accuracy would validate the biological plausibility of those models.

In general, this research represents a confluence of disparate disciplines and related research areas within speech recognition. Greater interaction between artificial intelligence and speech science would likely result in further shared increases in perspective and knowledge in the future.

Acknowledgments

This work was funded by the University of Toronto, Bell University Labs, the Natural Sciences and Engineering Research Council of Canada (Grant CRDPJ 364360-07), and the Canada Research Chair program. We acknowledge Gerald Penn and Fraser Shein for their academic contributions throughout this process as well as Aravind Namasivayam and Talya Wolff during the collection of the TORGO database. We are grateful to the participants who expended considerable time and effort in contributing to the TORGO database.

References

Abbs, J. H., Folkins, J. W., & Sivarajan, M. (1976). Motor impairment following blockade of the infraorbital nerve: Implications for the use of anesthetization techniques in speech research. *Journal of Speech and Hearing Research, 19,* 19–35.

American Speech-Language-Hearing Association. (2011). *Speech-language pathology medical review guidelines.* Retrieved from www.asha.org/practice/reimbursement/SLP-medical-review-guidelines.

Augmentative Communication, Inc. (2007). *Section 3: Clinical aspects of AAC devices*. Retrieved from www.augcominc.com/whatsnew/ncs3.html.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1,* 295–311.

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

Chesta, C., Siohan, O., & Lee, C.-H. (1999). Maximum a posteriori linear regression for hidden Markov model adaptation. *EUROSPEECH-99: Proceedings, Sixth European Conference on Speech Communication and Technology, 1,* 211–214. Retrieved from www.isca speech.org/archive/eurospeech_1999/e99_0211.html.

Duffy, J. R. (1995). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Mosby.

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

Freund, H.-J., Jeannerod, M., Hallett, M., & Leiguarda, R. (2005). *Higher-order motor disorders: From neuroanatomy and neurobiology to clinical neurology*. New York, NY: Oxford University Press.

Fukuda, T., & Nitta, T. (2003). Noise-robust automatic speech recognition using orthogonalized distinctive phonetic feature vectors. *EUROSPEECH 2003–INTERSPEECH 2003: Eighth European Conference on Speech Communication and Technology, 3,* 2189–2192. Retrieved from www.isca-speech.org/archive/eurospeech_2003/e03_2189.html.

Goldberger, J., Gordon, S., & Greenspan, H. (2003). An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. *Proceedings of Ninth IEEE International Conference on Computer Vision, 1,* 487–493. doi:10.1109/ICCV.2003.1238387

Goto, Y., Hochberg, M. M., Mashao, D. J., & Silverman, H. F. (1995). Incremental MAP estimation of HMMs for efficient training and improved performance. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995), 1,* 457–460. doi:10.1109/ICASSP.1995.479627

Gracco, V. L. (1995). Central and peripheral components in the control of speech movements. In F. Bell-Berti & L. J. Raphael (Eds.), *Producing speech: Contemporary issues. For Katherine Safford Harris* (pp. 417–431). Woodbury, NY: American Institute of Physics Press.

Guenther, F. H., & Perkell, J. S. (2004). A neural model of speech production and its application to studies of the role of auditory feedback in speech. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 29–49). New York, NY: Oxford University Press.

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. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), 1,* 924–927. doi:10.1109/ICASSP.2003.1198933

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing: A guide to theory, algorithm and system development*. Upper Saddle River, NJ: Prentice Hall.

Huber, M. F., Bailey, T., Durrant-Whyte, H., & Hanebeck, U. D. (2008). On entropy approximation for Gaussian mixture random vectors. *Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 1,* 181–188. doi:10.1109/MFI.2008.4648062

Hustad, K. C., Gorton, K., & Lee, J. (2010). Classification of speech and language profiles in 4-year-old children with cerebral palsy: A prospective preliminary study. *Journal of Speech, Language, and Hearing Research, 53,* 1496–1513.

Kent, R. D., & Rosen, K. (2004). Motor control perspectives on motor speech disorders. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 285–311). New York, NY: 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.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., . . . Wolf, P. (2003). *The CMU SPHINX-4 Speech Recognition System.* Unpublished manuscript. Retrieved from www.cs.cmu.edu/~rsingh/homepage/papers/icassp03-sphinx4_2.pdf.

Lazo, A. C., & Rathie, P. N. (1978). On the entropy of continuous probability distributions. *IEEE Transactions on Information Theory, 23,* 120–122.

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

Mefferd, A. S., & Green, J. R. (2010). Articulatory-to-acoustic relations in response to speaking rate and loudness manipulations. *Journal of Speech, Language, and Hearing Research, 53,* 1206–1219.

Menéndez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzjo, J. E., & Bunnell, H. T. (1996). The Nemours database of dysarthric speech. *Proceedings of the Fourth International Conference on Spoken Language Processing, 3,* 1962–1965. doi:10.1109/ICSLP.1996.608020

Miyazawa, Y. (1993). An all-phoneme ergodic HMM for unsupervised speaker adaptation. *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2,* 574–577. doi:10.1109/ICASSP.1993.319372

Murphy, K. P. (2002). *Dynamic Bayesian networks: Representation, inference and learning* (Unpublished doctoral dissertation). University of California at Berkeley.

Nam, H., & Goldstein, L. (2006). *TADA (TAsk Dynamics Application) manual*. New Haven, CT: Haskins Laboratories.

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. *Proceedings of the 15th International Congress of Phonetic Sciences, 1,* 2253–2256.

O'Shaughnessy, D. (2000). *Speech communications: Human and machine*. New York, NY: IEEE Press.

Raghavendra, P., Rosengren, E., & Hunnicutt, S. (2001). An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication, 17,* 265–275.

Ramsay, J. O., & Silverman, B. W. (2005). Principal differential analysis. In J. O. Ramsay & B. W. Silverman (Eds.), *Functional data analysis* (pp. 327–348). New York, NY: Springer.

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

Rosen, K., & Yampolsky, S. (2000). Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication, 16,* 48–60. doi:10.1080/07434610012331278904

Rudzicz, F. (2007). Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. *Proceedings of the Ninth International ACM SIGACCESS Conference on Computers and Accessibility (Assets07), 1,* 255–266. doi:10.1145/1296843.1296899

Rudzicz, F. (2009). Applying discretized articulatory knowledge to dysarthric speech. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09), 1,* 4501–4504. doi:10.1109/ICASSP.2009.4960630

Rudzicz, F. (2010a). Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics. *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP10), 1,* 4198–4201. Retrieved from www.cs.toronto.edu/~frank/Download/Papers/rudzicz_icassp10.pdf.

Rudzicz, F. (2010b). Correcting errors in speech recognition with articulatory dynamics. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), 1,* 60–68. Retrieved from www.aclweb.org/anthology-new/P/P10/P10-1007.pdf?CFID=99183903&CFTOKEN=86059735.

Rudzicz, F. (2010c). Towards a noisy-channel model of dysarthria in speech recognition. *Proceedings of the First Workshop on Speech and Language Processing for Assistive Technologies (SLPAT) at the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010), 1,* 80–88. Retrieved from www.cs.toronto.edu/~frank/Download/Papers/rudzicz_naacl10.pdf.

Rudzicz, F. (2011a). Articulatory knowledge in the recognition of dysarthric speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19,* 947–960.

Rudzicz, F. (2011b). *Production knowledge in the recognition of dysarthric speech* (Unpublished doctoral dissertation). University of Toronto, Ontario, Canada.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2011). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*. Advance online publication. doi:10.1007/s10579-011-9145-0

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (pp. 129–144). Berlin, Germany: Springer-Verlag.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333–382. doi:10.1207/s15326969eco0104_2

Seikel, J. A., King, D. W., & Drumright, D. G. (Eds.). (2005). *Anatomy & physiology for speech, language, and hearing*, (3rd ed.). Clifton Park, NJ: Delmar.

Shannon, C. E. (1949). *A mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Sharma, H. V., & Hasegawa-Johnson, M. (2010). State-transition interpolation and map adaptation for HMM-based dysarthric speech recognition. *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), 1,* 72–79. Retrieved from www.isle.illinois.edu/pubs/2010/sharma10slpat.pdf.

Shevell, M., Miller, S. P., Scherer, S. W., Yager, J. Y., & Fehlings, M. G. (2011). The Cerebral Palsy Demonstration Project: A multidimensional research approach to cerebral palsy. *Seminars in Pediatric Neurology, 18,* 31–39. doi:10.1016/j.spen.2011.02.004

Smith, A., & Goffman, L. (2004). Interaction of motor and language factors in the development of speech production. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 227–252). New York, NY: Oxford University Press.

Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics, 38,* 10–19. doi:10.1016/j.wocn.2008.10.004

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

van Lieshout, P. (2004). Dynamical systems theory and its application in speech. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 51–82). New York, NY: Oxford University Press.

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,* 283–303.

van Lieshout, P. H. H. M., & Moussa, W. (2000). The assessment of speech motor behavior using electromagnetic articulography. *The Phonetician, 81,* 9–22.

Vertanen, K. (2006). *Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments* (Technical report). Cambridge, United Kingdom: Cavendish Laboratory.

Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs: A review. In *ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition (Adaptation-2001)* (pp. 11–19). Retrieved from www.isca-speech.org/archive_open/adaptation/adap_011.html.

Wrench, A. (1999). MOCHA-TIMIT [Articulatory database]. Edingburgh, United Kingdom: Centre for Speech Technology Research, University of Edinburgh. Retrieved from www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Wrench, A., & Richmond, K. (2000). Continuous speech recognition using articulatory data. *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), 4,* 145–148. Retrieved from www.isca-speech.org/archive/icslp_2000/i00_4145.html.

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

Ziegler, W., & Maassen, B. (2004). The role of the syllable in disorders of spoken language production. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 415–447). New York, NY: Oxford University Press.

Zierdt, A., Hoole, P., & Tillmann, H. G. (1999). Development of a system for three-dimensional fleshpoint measurement of speech movements. *Proceedings of the 14th International Conference of Phonetic Sciences (ICPhS99), 1,* 73–75. Retrieved from www.phonetik.uni-muenchen.de/~hoole/pdf/3d_icphs99.pdf.

Zue, V., Seneff, S., & Glass, J. (1989). Speech database development: TIMIT and beyond. *In ESCA Tutorial and Research Workshop on Speech Input / Output Assessment and Speech Databases (SIOA-1989), 2,* 35–40. Retrieved from www.isca-speech.org/archive_open/sioa_89/sia_2035.html.

Appendix. Hidden Markov models (HMMs).

Definition

The parameters of the HMM include a state space Q (where q_t is the state at time t), an initial state distribution $\pi_i = P(q_0 = i)$ describing the probability of starting a sequence in a particular state i, a state transition matrix $A(q_i, q_j)$ describing the a priori probability of transitioning from state q_i to q_j, and a probability $B_i(o)$ of observing vector o while in state i. Φ is defined as the set of all adjustable parameters, namely $\Phi = \{a_{ij} = A(q_i, q_j), B_i(o), \pi_i\}$.

Except where noted, the probability $B_i(o)$ of observing a set of cepstral features o while in temporal state i is described by a mixture of Gaussians. This distribution weights and combines several standard Gaussian (normal) distributions to allow for more complex probability spaces and is defined by

$$B_i(o; \omega, \mu, \textstyle\sum) = \sum_{m=1}^{M} \omega_m \mathbf{N}(o; \mu_m, \textstyle\sum_m), \tag{7}$$

where μ_m and $\sum_m$ are the mean and covariance of the mth Gaussian, respectively, and ω_m is the weight of that Gaussian. The result is a valid probability subject to the constraint that $\sum_m \omega_m = 1$. Crucially, the total number of Gaussians used in $B_i(o)$ is a manually specified parameter that we can empirically alter.

It is desirable to find values of Φ such that $P(O; \Phi)$ is maximized for a large collection of observation sequences, O . Since the hidden states are inaccessible, Φ cannot be estimated with maximum likelihood estimation. Instead, expectation-maximization (EM) is used as a "hill-climbing" approach that iteratively approximates the values of the hidden states given the current best parameters of Φ, then updates Φ. Prior to training each HMM, the Gaussian mixtures for all states are initialized to a common Gaussian mixture obtained by k-means clustering with full covariance over all data for the associated phoneme.

Adaptation in HMMs

In maximum a posteriori (MAP) estimation, given a parameter space Φ defined on HMMs as described above, prior knowledge can characterize a probability density $p(\Phi)$. Given a set of observation sequences X, the MAP estimate for the ideal parameters is

$$\hat{\Phi} = \arg\max_{\Phi}\ p(\Phi|\mathrm{X}) = \arg\max_{\Phi}[p(\mathrm{X}|\Phi)p(\Phi)]. \tag{8}$$

This estimate reduces to the maximum likelihood estimate if $p(\Phi)$ is uniform, that is, when there is no prior knowledge. Since we use continuous Gaussian mixture HMMs, we assume that the different components are mutually independent, which is standard practice (Huang et al., 2001) and allows us to split the optimization problem into subcomponents. For example, to obtain a more appropriate weight, $\hat{\omega}_i[m]$ for the mth Gaussian in the ith state, we use the Lagrange method

$$\frac{\delta}{\delta\hat{\omega}_i[m]}\left(\log\ p_{\omega_i}(\vec{\omega}_i) + \sum_{m=1}^{M}\sum_{t}\xi_t(i, m)\log\hat{\omega}_i[m]\right) + \lambda = 0, \forall m, \tag{9}$$

with the constraint that $\sum_{m=1}^{M} \hat{\omega}_i[m] = 1$. In equation 9 , λ is the Lagrange multiplier, and $\xi_t(i, m)$ is the probability that the observation at time t was generated by the mth Gaussian of the ith state. The solution is

$$\hat{\omega}_i[m] = \frac{\alpha_i[m] - 1 + \sum_t \xi_t(i, m)}{\sum_{l=1}^{M}(\alpha_i[l] - 1 + \sum_t \xi_t(i, l))}. \tag{10}$$

Optimization with respect to the means and covariances of the Gaussians is accomplished in the same manner (Goto, Hochberg, Mashao, & Silverman, 1995; Woodland, 2001). Here, MAP adaptation is embedded within maximum likelihood regression, as described by Chesta, Siohan, and Lee (1999).

Vocal Tract Representation in the Recognition of Cerebral Palsied Speech

Frank Rudzicz, Graeme Hirst, and Pascal van Lieshout

J Speech Lang Hear Res 2012;55;1190-1207; originally published online Jan 23, 2012;
DOI: 10.1044/1092-4388(2011/11-0223)