Articulatory Knowledge in the Recognition of Dysarthric Speech

Frank Rudzicz, *Student Member, IEEE*

***Abstract*—Disabled speech is not compatible with modern generative and acoustic-only models of speech recognition (ASR). This work considers the use of theoretical and empirical knowledge of the vocal tract for atypical speech in labeling segmented and unsegmented sequences. These combined models are compared against discriminative models such as neural networks, support vector machines, and conditional random fields. Results show significant improvements in accuracy over the baseline through the use of production knowledge. Furthermore, although the statistics of vocal tract movement do not appear to be transferable between regular and disabled speakers, transforming the space of the former given knowledge of the latter before retraining gives high accuracy. This work may be applied within components of assistive software for speakers with dysarthria.**

***Index Terms*—Articulatory models, discriminative methods, dysarthria.**

I. Introduction

THERE are several simplifying assumptions in automatic speech recognition (ASR) that have become particularly ingrained. One such assumption is that the acoustics of speech can be adequately described while being agnostic to non-surface phenomena. Although ASR takes a few important cues from the biological perception of speech, such as the Mel scale [1], it rarely models physical production explicitly. Secondly, modern ASR is often built assuming that models trained on a sufficiently large set of speakers will adequately capture enough inter-speaker variability to be usable by a typical user. The further one's voice deviates from this aggregate, however, the less likely an ASR system is to function as intended, as shown next.

Each of these simplifications can appear to be useful in certain contexts but their utility in the presence of more atypical patterns of production can be contentious, especially in cases of speech disorder. One group of such disorders, called dysarthria, is primarily an endogenous phenomenon distinguished by its aberrant mechanics of articulation resulting in highly unintelligible speech that is not accommodating to the traditional assumptions of speech recognition. This paper describes work whose goal is to improve speech recognition accuracies for dysarthric individuals by augmenting acoustic models with articulatory information. The relationships between acoustics and articulation are especially relevant for these speakers, for whom normal speech production is compromised. After an introduction to the effects of dysarthria, this paper presents a new database of dysarthric articulation and several experiments in articulatory modeling for recognition of atypical speech.

The purpose of this work is to discover how traditional acoustic modeling of dysarthric speech can be improved with articulatory information and to expand on recent work in this area [2], [3].

A. Dysarthria

Dysarthria is a set of congenital and traumatic neuromotor disorders that impair the physical production of speech. These impairments reduce or remove normal control of the primary vocal articulators but do not affect the regular comprehension or production of meaningful, syntactically correct language. Congenital causes of dysarthric speech are often caused by some sort of asphyxiation of the brain, inhibiting normal development in the speech–motor areas. Of these causes, cerebral palsy is among the most common, affecting approximately 0.5% of children in North America [4], 88% of whom are dysarthric throughout adulthood [5]. Later-onset causes are more typically traumatic, including cerebro-vascular stroke affecting approximately 1% of the population aged 45 to 64, and 5% of those aged 65+, with the severity of impairment varying with the amount of cerebral damage [5]. Other sources of dysarthria include multiple sclerosis, Parkinson's disease, myasthenia gravis (i.e., blocked acetylcholine receptors), and amyotrophic lateral sclerosis (ALS).

Neurological bases of dysarthria involve damage to the cranial nerves that control the articulatory musculature of speech [6]. For example, damage to the glossopharyngeal nerve typically reduces control over vocal fold vibration (i.e., phonation), resulting in either guttural or grating raspiness. Inadequate control of soft palate movement caused by disruption of the vagus cranial nerve may lead to a disproportionate amount of air being released through the nose during speech (i.e., hypernasality). More commonly, a lack of tongue and lip dexterity often produces heavily slurred speech and a more diffuse and less differentiable vowel target space [7]. The lack of articulatory control often leads to various involuntary sounds caused by velopharyngeal or glottal noise, or noisy swallowing problems [8].

Dysarthric speech can be up to 17 times slower than regular speech, at about 15 words per minute in severe cases [9]. Apart from being more laborious for the speaker and listener, slow

Manuscript received October 16, 2009; revised April 28, 2010 and August 10, 2010; accepted August 15, 2010. Date of publication September 07, 2010; date of current version March 30, 2011. This work was supported by Bell University Labs, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto. It incorporates some previously published work. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ruhi Sarikaya.

Thee author is with the Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: frank@cs.toronto.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2072499

speech has several acoustic consequences. For example, human listeners often mispartition words and syllables prolonged by lengthened vowels or extended occlusions preceding voiceless plosives [10]. Other types of disfluency commonly associated with dysarthria (especially in conjunction with apraxia) include hesitation (e.g., false starts) and repetition (e.g., stuttering), although these actually result from higher-level linguistic causes [11]. These sorts of disfluencies can produce severely atypical phrasing which is difficult to comprehend at the utterance level. Despite a great amount of inter-speaker variability, dysarthric individuals who can maintain a regular speaking rate are able to repeat individual speech units with fairly normal consistency [7].

Standardized Assessments: Clinical assessments of motor function and intelligibility in speakers with dysarthria are often used by speech therapists for rehabilitation [11]. The Frenchay Dysarthria Assessment, for example, is a standard series of tests that individually measure respiration, reflex, speaking rate, the strengths of various articulators, and word and phrase intelligibility on 9-point scales [12]. Since intelligibility correlates well with ASR accuracy [13], these assessments are used to find correlations between particular speech deficits and observations across several speech classification models. For instance, the degree of tongue disability is a theoretical indicator of poorer discrimination between front and back vowels.

B. Representations for Speech Production

Articulatory features (AFs) are quantized abstractions of speech production according to distinctive configurations of the vocal tract.[1] They provide an inventory of the types of sounds humans can produce [1], [18]. The study of AFs in recent phonetics dates back at least to Chomsky and Halle [19], who represented sounds of speech as vectors of binary features (e.g., nasal/non-nasal, voiced/voiceless). That work showed that some context-sensitive phonetic variation could be specified by transformational rules based on phoneme sequences and syntactic trees (e.g., /p/ is aspirated if it begins a syllable onset consonant cluster, as in *prim*, but not aspirated if it ends that onset, as in *spin*).

Here, articulatory features are collected into seven categories, each with a number of possible values. For example, a segment of speech can be concurrently voiced, nasal, and static, which represent values for three distinct features. Parallelizing streams of information in this manner allows asynchronous modulation of speech acts across phoneme boundaries, which can partially account for coarticulation effects and speaker variability [20], which are particularly exacerbated in dysarthric speech. Other useful properties reported of AFs include language-independence and reliable recovery from acoustics among regular speakers [21]. The features used here are based on those of Wester [22] and are listed in Table I.

In the absense of AF annotations, AF values can be derived directly from phoneme annotations. In this study, we assign to each MFCC frame of data a seven-dimensional vector of AF values based exclusively on the phoneme annotation at that frame. This assignment is derived directly from the phoneme-to-AF transformation table in Frankel *et al.* [21]. This incorporates recommendations by Wester *et al.* [23] in which the Front/Back feature includes the normally excluded *central* value, and diphthongs are split in half into their component vowels, which are mapped to their corresponding AFs. Unlike Frankel *et al.* [21], we label the Place feature of phonemes /b/ and /m/ as bilabial rather than labiodental.

A more empirical approach to production knowledge is derived from direct measurement of the vocal tract during speech with semi-invasive procedures such as electromagnetic articulography (EMA), magnetic resonance imaging (MRI), X-ray microbeam analysis [24], or electropalatograph. These procedures capture motions of external (e.g., lips) and internal (e.g., tongue, velum) actuators with sufficient temporal and spatial resolution to accurately reconstruct physical activity [25]. EMA is the source of kinematic data used in our experiments next. Here, the positions of the tongue, lips, and other articulators can be accurately inferred at a rate of 200 Hz to within 0.5 mm [26] relative to fixed transmitters around the speaker's head that produce alternating magnetic fields. These systems produce no audible noise, and the coils do not interfere with regular speech. Fig. 1 shows typical configurations of the EMA cube and the placement of the receiver coils.

TABLE I
ARTICULATORY FEATURES, A DESCRIPTION OF THEIR CHARACTERISTICS, AND THEIR POSSIBLE VALUES

Feature	Description (*and values*)
Manner (**M**)	high-level categorization of speech sound *approximant, fricative, nasal, retroflex, silence, stop, vowel*
Place (**Pl**)	location of primary constriction *alveolar, bilabial, dental, labiodental, velar, silence,nil*
High/Low (**HL**)	ventral position of the tongue *high, mid, low, silence, nil*
Front/Back (**FB**)	anterior position of the tongue *front, central, back, nil*
Voice (**V**)	presence/absence of glottal vibration *voiced, unvoiced*
Round (**R**)	circularity of the lips *round, non-round, nil*
Static (**S**)	movement of articulators (e.g., diphthong) *static, dynamic*

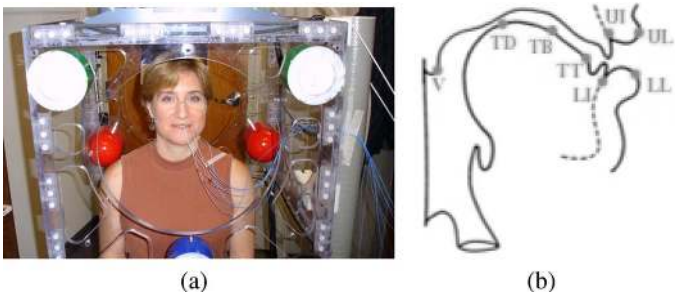

Fig. 1. Example configuration of electromagnetic articulography. (a) shows a subject connected within the recording environment, and (b) shows the typical locations of receiver coils on the midsagittal plane (i.e., **V** velum, **TD** tongue dorsum, **TB** tongue body, **TT** tongue tip, **UI** upper incisor, **LI** lower incisor, **UL** upper lip, and **LL** lower lip).

[1]Articulatory features are sometimes called *phonological features* in the literature (e.g., by Clements [14] and by King and Taylor [15]). However, the latter term has largely been superseded by the former in the literature (e.g., by Kirchhoff [16] and by Metze [17]). In this paper, the term *articulatory feature* must be differentiated from articulatory *measurements*, which refer to direct recordings of the vocal tract.

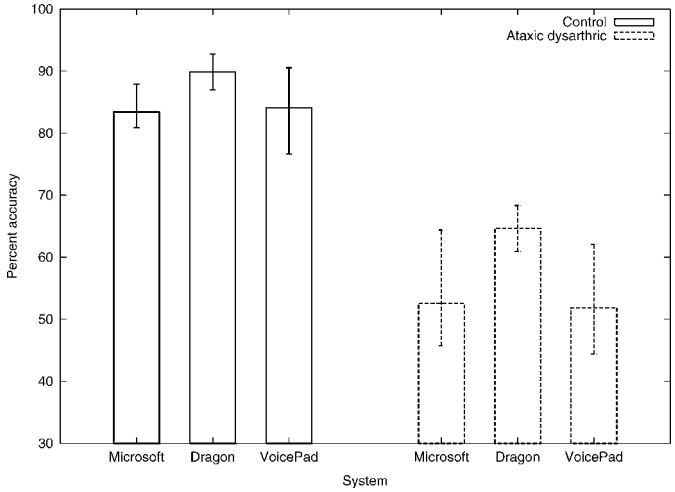

Fig. 2. Comparison of recognition accuracies for control and ataxic speakers across Microsoft Dictation, Dragon NaturallySpeaking, and KES VoicePad Platinum, from Hux *et al*. [29]. Boxes represent average accuracies, with errorbars representing minimum and maximum accuracy over five trials.

II. Previous Work

There have been a number of attempts at improving speech recognition for speakers with dysarthria, and other attempts at integrating articulatory knowledge into ASR, but these two efforts have so far not converged. The following subsections describe the state of the art in each sub-domain.

A. Speech Recognition for Speakers With Dysarthria

Early work in applying ASR to individuals with dysarthria almost exclusively involved the use of hidden Markov models (HMMs) whose parameters were trained to the general population. Usually, these involved small-vocabulary recognition tasks with word-recognition accuracies significantly lower for speakers with dysarthria, often at least 26.2% lower than the general population [27]. For example, given a vocabulary of 40 words, Noyes and Frankish [28] report mean word-recognition accuracies of 58.6% for speakers with dysarthria compared with 95% for the general population. Hux *et al.* [29] report similar divergences with continuous sentences in three commercial ASR dictation systems, namely Microsoft Dictation, Dragon NaturallySpeaking (DNS), and Kurzweil Education Systems' VoicePad Platinum. All systems performed significantly better with regular speech, averaging between 83.4% (Microsoft) and 89.9% (Dragon) word-recognition, compared with between 50.9% (VoicePad) and 64.7% (Dragon) for speakers with dysarthria. These results are shown in Fig. 2. Despite their relatively poor results, however, such commercial ASR systems have been shown to improve accuracy and speed in simple text-entry for physically disabled individuals relative to other modes of input (e.g., scan-and-switch) [30], [31].

Several projects have attempted to adapt to dysarthric speech without considering the causes or features of dysarthria. For example, feed-forward neural networks supplied with either Fourier spectral coefficients or formant frequencies have been shown to reduce error relative to commercial HMM-based systems by up to 40% on isolated word-recognition for cerebrally palsied speech [32]. Adapting HMM acoustic models trained to the general population given dysarthric data has also shown

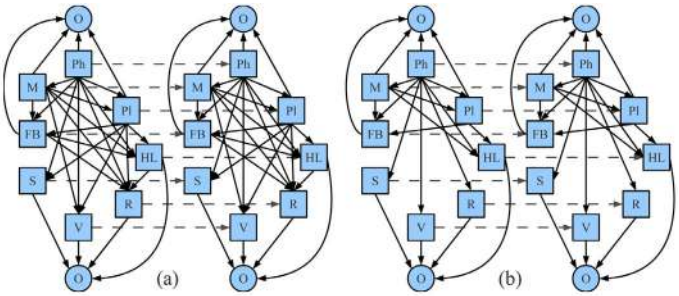

Fig. 3. Two-frame dynamic Bayes networks with articulatory features (DBN-F (default), left, and DBN-F (sparse), right). Nodes **Ph**, **Q**, and **O** represent phoneme, state, and MFCC observations. All other variables are highlighted in Table I. Inter-frame conditional links are dashed for clarity.

to improve accuracy, but not as much as training those models exclusively with dysarthric acoustics, especially in the more severe cases [10], [33].

More recently, attempts have been made to improve ASR accuracies by focusing on the types of errors made with dysarthric speech. Polur and Miller [34], for example, produced ergodic HMMs that allow for "backwards" state transitions. This ergodic structure is meant to capture aspects of dysarthric speech such as stuttering and disruptions during sonorants (e.g., pauses) and reveals small but definite improvements over the traditional baseline. Morales and Cox [35] improved word-error rates by approximately 5% on severely dysarthric speech and approximately 3% on moderately dysarthric speech by building weighted transducers into an ASR system according to observed phonetic confusion matrices. A commonality among all this work is that the actual articulatory behavior of the dysarthric speech has not been taken into account.

B. Speech Recognition With Articulatory Information

Discrete articulatory feature recognition has been applied to identifying values for concurrent features (similar to those in Table I), usually independently from phone recognition or more general ASR [36]. Neural network discriminative classifiers have been shown by King and Taylor [15], Kirchhoff [16], and Scharenborg *et al.* [37] to correctly identify approximately 53% of simultaneous multivalued AFs, on average, for non-dysarthric speech (e.g., from TIMIT). More recently, dynamic Bayes networks have been applied to this problem, on similar data, and using structures similar to the sparser variant in Fig. 3 [21]. This model correctly identified 57.8% of similar multivalued AFs on non-dysarthric speech.

Articulatory knowledge has had relatively little historical presence in ASR despite evidence that articulatory control is often far more speaker-invariant than the resulting acoustics [38]. Typically, such knowledge is manifested as decision trees that support state-tying in semi-continuous ASR systems [39]. Here, knowledge of common articulatory features (e.g., nasality in */m/* and */n/*) allows states in HMM models for different phones to be trained on shared data. There have, however, been a few attempts to build more explicit production knowledge into phoneme- and word-recognition systems. For example, appending articulatory measurements to acoustic observations has shown to reduce phone-error relatively by up to 17% on a speaker without dysarthria in a standard HMM

system (however, if those articulatory measurements are inferred from acoustics, this improvement disappeared) [40]. Similar work on incorporating AFs learned discriminatively with maximum mutual information into HMM systems have reduced word-error rates from 25% to 19.8% on English spontaneous scheduling tasks [17]. Along the same lines, systems incorporating discrete AFs derived by NNs from acoustics into HMM-based ASR have shown some improvement over the acoustic-only baselines [41], [16], although these results were often statistically insignificant except in the presence of extreme environmental noise [36].

More recently, Bayes networks have seen increased use in modeling interdependencies between articulation and acoustics in regular speech [42]. Stephenson *et al.* [43] showed that simple Bayes networks relating MFCC observations with Wisconsin's X-ray microbeam articulatory data [24] resulted in a 9% word-error rate reduction when compared with a baseline acoustic-only ASR system. Markov *et al.* [44] followed this work with a series of simpler Bayes networks that estimated the likelihood of acoustic observations given discretized articulatory parameters, achieving similar results when combined with an HMM-based ASR system.

A commonality among all of this work is its reliance on non-dysarthric data where articulatory and acoustic patterns are less disordered than in speakers with cerebral palsy and other neuromotor disabilities [20].

III. Data

Three speech databases are used in this study. The first consists of dysarthric acoustics, but without direct vocal tract measurements. The second includes vocal tract measurements, but only for speakers without dysarthria. The third database includes EMA recordings of speakers with dysarthria, and is currently being recorded at the University of Toronto. These databases are described next.

A. Nemours Database

The Nemours database is a popular source of phonemically annotated dysarthric acoustics consisting of 11 dysarthric males and one non-dysarthric male each uttering 74 syntactically invariant short sentences and two additional paragraphs [45]. Here, phonemic annotations were automatically derived by HMM-based forced alignment given known orthography and corrected manually by the authors of that database. Each speaker is also associated with a complete Frenchay assessment of motor function. Since no physiological information is included, articulatory features are derived directly from phonemic annotations as described in Section I-B and provide the bases for production knowledge in Section V.

B. MOCHA Database

The University of Edinburgh's MOCHA database consists of 460 sentences derived from TIMIT [46] uttered by a male and a female British speaker [47]. All acoustic data are temporally aligned with EMA and laryngograph measurements. For this study we use eight bivariate articulatory parameters, namely the upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI), tongue tip (TT), tongue blade (TB), tongue dorsum (TD), and velum (V). Each parameter is measured in the two dimensions of the midsagittal plane, resulting in a 16-dimensional articulatory configuration.

C. TORGO Database of Dysarthric Articulation

The TORGO database [48] is an ongoing project that consists of aligned acoustic and articulatory recordings for the purpose of learning statistical relationships between dysarthric and non-dysarthric speech production. This database currently consists of seven dysarthric subjects with either cerebral palsy or amyotrophic lateral sclerosis (ALS), and gender-matched controls. Each participant has recorded 3 hours of data (approximately 500 utterances from each speaker with dysarthria and 1200 from speakers without dysarthria) split across multiple sessions in two 3-D measurement environments, namely EMA and a 3-D reconstruction given binocular video recordings of phosphorescent facial markers [49]. In general, video provides more facial motion data (e.g., the depressor anguli oris muscle) but excludes any tongue motion. Since speakers with dysarthria are in the minority and susceptible to fatigue, collecting data from this population can be particularly challenging. Most published experiments typically include no more than three or four speakers with dysarthria [50], often producing only about 25 utterances each [32] for on the order of 100 samples in total. Similarly, Yunusova *et al.* [51] recorded 15 speakers with ALS and Parkinson's disease, but each speaker repeated only ten word stimuli each.

In this paper, we concentrate on our EMA recordings, which constitute approximately 60% of our data. Unlike the MOCHA database, our recordings include points outside the midsaggital plane, namely the two lip corners and one point behind each ear, but not on the velum. In addition to typical issues of speech data collection such as the need to suppress environmental noise, the development of the TORGO database has incurred some additional challenges specific to the population. Decreased control of salivation and an increased risk of a severe gag reflex among cerebrally palsied participants can make placing coils on the tongue very difficult, so approximately 12% of EMA data from dysarthric individuals does not include the rearmost tongue positions. Involuntary movement such as shaking or extension of the neck also presents a problem for video recording, as the points on the face become occluded.

Stimuli are read by the participants from an LCD screen and are randomized at runtime within smaller collections to ensure direct comparability between speakers who complete data at different rates. Single-word stimuli include repetitions of the English digits, the international radio alphabet, the 20 most frequent words in the British National Corpus, and words selected by Kent *et al.* to demonstrate relevant articulatory contrasts (e.g., alveolar-palatal fricatives, front-back vowels, stop-nasals) [52]. These contrasts are especially relevant given speakers with articulatory disorders. Single-word stimuli are useful to study variation in isolation without boundary detection. Sentence stimuli are derived from the Yorkston–Beukelman assessment of intelligibility [53] and the TIMIT database [46]. Sentences in the Yorkston–Beukelman assessment are designed to highlight perceptual contrasts in speech that are relevant to speaker intelligibility. We complement these with sentences from TIMIT in

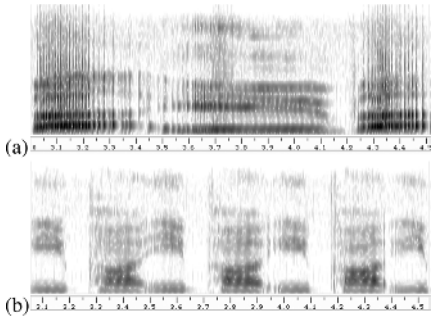

Fig. 4. Repetitions of */iy pcl p ah/* over 1.5 s by (a) a male speaker with athetoid CP, and (b) a female control in the TORGO database. Dysarthric speech is notably slower and more strained than regular speech.

order to more readily compare results performed with our database with those performed with others. The use of sentences in general allow for the use of higher-level syntax and language modeling in ASR.

Fig. 4 exemplifies some typical acoustic contrasts between dysarthric and non-dysarthric speech in TORGO. On average, dysarthric vowels are 116.7 ms while control vowels are 45.5 ms. This might partially be explained by an increase of brief staccato gaps in exhalation during sonorants. Dysarthric vowel acoustics are also slightly more variable, with an average variance across the first seven mel-scaled frequency cepstral coefficients of 12.1, against 9.8 in control data. Notably, speakers with dysarthria mispronounce plosives in word-initial, -medial, and -final positions 16%, 20%, and 19% of the time, respectively, and substitutions in this class are exclusively from unvoiced to voiced. By comparison, only 5% of corresponding plosives are mispronounced in regular speech, either dropped in the final position or incorrectly voiced in word-medial positions. Also, our dysarthric data often includes many deleted affricates in word-final and fricatives in word-initial positions, almost all of which are static and alveolar. This does not occur in the corresponding non-dysarthric data.

All data are being phonemically annotated to the TIMIT phone set [46] by a speech-language pathologist to allow supervised frame-level training of phone-dependent acoustic and kinematic models. These annotations are further checked by two naïve listeners for consistency, although automatic phonemic labeling by forced alignment on similar data has been shown to be sufficient for certain tasks [54]. Additionally, all dysarthric participants are diagnosed by a speech-language pathologist according to the standardized Frenchay Dysarthria Assessment (Section I-A). The following experiments make use of data from two speakers with dysarthria (male and female) and two speakers without (male and female) whose data are fully annotated at the time of this writing.

IV. Classification Methods

Throughout the following experiments we apply five classification methods which are described next.

A. Hidden Markov Models (HMMs)

The default baseline is a tristate left-to-right triphone HMM with observation likelihoods at each state computed over mixtures of 16 Gaussians through marginalization amenable to normal expectation-maximization training with Baum–Welch and Viterbi decoding. An HMM is evaluated by the Forward algorithm, in which the probability of the observation sequence $\mathbf{o}$ is modeled by

$$P(\mathbf{o}) = \sum_{\forall \mathbf{q}} P(\mathbf{q}) P(\mathbf{o} \,|\, \mathbf{q}) \tag{1}$$

which sums over all possible sequences of hidden states $\mathbf{q}$. These quantities are also used in computing the objective function during Baum–Welch training [18]. Prior to training each HMM, the Gaussian mixtures for all states are first initialized to a common Gaussian mixture obtained by performing k-means clustering with full covariance over all data for the associated triphone. If fewer than five examples of the triphone exist, data for the associated monophonic root are used instead. This approach to dealing with sparse triphone data is taken for all other classification methods as well.

B. Latent-Dynamic Conditional Random Fields (LDCRFs)

The discriminative latent-dynamic conditional random field is a sequence classifier differing from the HMM in that its estimation of the distribution over a sequence of labels $\mathbf{l}$ (where the ith label $l_i \in \mathcal{L}$ for some vocabulary of labels $\mathcal{L}$) does not model the observation prior $P(\mathbf{o})$, as shown in (2). This model extends traditional conditional random fields in that it models an intrinsic sequential substructure using hidden states, and differs from "hidden state" CRFs in that labels are assigned dynamically on a frame-by-frame basis, rather than once to the entire sequence [55].

In CRFs, the parameter set θ defines the weights ($\theta_k \in \theta$) applied to *feature functions* f_k of the graphical model, which are analogous to state and observation variables in HMMs (see Lafferty *et al.* [56]). In fact, the parameters θ are analogous to logarithms of the conditional probabilities present between variables in HMMs (i.e., transition probabilities and state-specific observation probabilities) and are initialized randomly. In this approach, we wish to measure the likelihood of a particular labeling $\mathbf{l}$ of an observation sequence $\mathbf{o}$ given some parameterization θ. This quantity must be computed over all possible sequences of hidden states (where $\mathbf{q}$ is a particular state sequence) that produce that label sequence, where each state q_i comes from the set $\mathcal{Q}_{l_i}$ of states associable with a particular label l_i at time i. For example, an LDCRF model for phoneme $/m/$ might have three hidden states (i.e., $|\mathcal{Q}_m| = 3$) which are distinguished from the states in the other phoneme models. In other words,

$$P(\mathbf{l} \,|\, \mathbf{o}, \theta) = \sum_{\mathbf{q}: q_i \in \mathbf{Q}_{l_i}} P(\mathbf{l}|\mathbf{q}, \mathbf{o}, \theta) P(\mathbf{q}|\mathbf{o}, \theta) \tag{2}$$

where $P(\mathbf{q}|\mathbf{o}, \theta)$ is the standard conditional random field formulation that defines state and transition functions [56], [55], namely

$$P(\mathbf{q}|\mathbf{o}, \theta) = \frac{\exp(\sum_k \theta_k F_k(\mathbf{q}, \mathbf{o}))}{\sum_{\mathbf{r}} \exp(\sum_k \theta_k F_k(\mathbf{r}, \mathbf{o}))} \tag{3}$$

where $F_k(\mathbf{q}, \mathbf{o})$ is the sum over all state transition feature functions applicable to $\mathbf{q}$ and observation feature functions applicable to $\mathbf{o}$.

TABLE II
NUMBER OF HIDDEN UNITS PER NN, GIVEN TARGET FEATURE

Feature	# hidden units	Feature	# hidden units
Manner	300	Voice	100
Place	200	Round	100
High/Low	100	Static	100
Front/Back	200		

Given a training set of labeled sequences $(\mathbf{o}_i, \mathbf{l}_i)$ where $i = 1 \ldots N$, we apply conjugate gradient ascent to find the optimal parameter values $\theta^* = \arg\max_\theta L(\theta)$ given the following objective function:

$$L(\theta) = \sum_{i=1}^{N} \log P(\mathbf{l}_i|\mathbf{o}_i, \theta) - \frac{1}{2\sigma^2}||\theta||^2 \tag{4}$$

which is the log-likelihood of the parameterization given by the conditional log-likelihood of each training sequence $\log P(\mathbf{l}_i|\mathbf{o}_i, \theta)$ and the Gaussian prior likelihood of θ with variance σ^2. If the parameter space θ is uniformly distributed, as we assume here, σ^2 approaches infinity and we discount the second term. Further details on training LDCRFs can be found in Morency *et al.* [55].

The label sequence hypothesis $\mathbf{l}^*$ is obtained by marginalizing over the sets of states $\mathbf{Q}_{l_t}$ given the label l_i at time t

$$\mathbf{l}^* = \arg\max_{\mathbf{l}} \sum_{\mathbf{q}:\ \forall q_t \in \mathbf{Q}_{l_t}} P(\mathbf{q}|\mathbf{o}, \theta^*). \tag{5}$$

C. Neural Networks (NN)

Despite their general popularity, NNs are rarely studied with regards to dysarthric acoustics, with some exceptions [32]. The two types of NN we consider here are the feed-forward multi-layer perceptron (MLP) and the recurrent Elman network (ELM) [57], which are primarily distinguished by the latter's time-delayed replication of the hidden layer as additional contextual input. The output of each AF NN consists of n nodes, where n is the cardinality of the class being modeled (i.e., either AF or phone), and the ith node is uniquely active when training the ith value of that class. Given the presence of 21 464 triphones in our data, this approach is not tenable for NNs that recognize triphones. Here, 15 output neurons are used in which each of the 2^{15} possible binary output combinations are mapped to a unique triphone (or a "null" triphone not considered in classification). The sizes of hidden layers in AF neural networks are based empirically on similar work on non-dysarthric speech [37], [21] and shown in Table II. All NN triphone classifiers contain 500 hidden units.

Activation functions at each node are tan-sigmoid (i.e., $a(x) = [2/(1+e^{-2x})] - 1$) in the hidden layer, and linear in the output layer, given a weighted sum of all inputs $x = \sum_j \omega_j a_j$, where a_j is the activation of node j and ω_j is the weight of the connection from node j to the current node, as usual. All NN training is performed by resilient back-propagation, which adjusts update values according to sign changes in partial derivatives. Here, the degree of updates is reduced if weights oscillate over several iterations and is increased when weights continually change in the same direction. This approach is faster than standard steepest descent on our data, while only requiring a modest increase in memory.

All networks are fully connected between layers and select the class having the highest posterior probability.

D. Support Vector Machines (SVMs)

General maximum margin classifiers are of increasing interest in ASR due to their robustness against both sparse data [58] and rapid transient changes in acoustic sequences [59]. SVMs explicitly minimize an upper bound on the expected classification error by orienting a hyperplane between classes such that the norm of its orthogonal vector maximizes the margin between the nearest data. We use a soft-margin SVM here and extend the process to k-class discrimination by training $k(k-1)/2$ binary classifiers, each delineating two class regions [60].

SVMs depend on kernel functions κ to describe the distance between two points of data. We consider two of these that differ slightly in the form of their input. The first kernel is a symmetric radial basis function (RBF), that generalizes to nonlinear decision boundaries using the following function:

$$\kappa_{\mathrm{RBF}}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{y})^2}{2\sigma^2}\right) \tag{6}$$

given vectors $\mathbf{x}$ and $\mathbf{y}$, and width parameter σ.

The second kernel κ_{DTW} is a sequence kernel that can be generalized to arbitrary sequences $\mathbf{u}$ and $\mathbf{v}$ having non-equal lengths, as proposed recently by Wan and Carmichael [58]. This kernel exploits the notion of distance between sequences inherent in dynamic time warping (DTW), and converts it to a form amenable for use in SVMs. The approach is to convert local Euclidean distances between frame vectors to angles by projecting these d-dimensional vectors onto a unit hypersphere H centered α units from their origin in the $(d+1)$st dimension. Namely, every vector u_i is converted to the unit vector $\hat{u_i}$ sharing an origin with H by

$$\hat{u_i} = \frac{1}{\sqrt{u_i^2 + \alpha^2}} \begin{bmatrix} u_i \\ \alpha \end{bmatrix}. \tag{7}$$

Given two unit vectors, $\hat{u_i}$ and $\hat{v}_j$ that define points on the surface of H, the angle between them is by definition

$$d_s(\hat{u_i}, \hat{v_j}) = \theta_{\hat{u_i}, \hat{v_j}} = \arccos(\hat{u_i}, \hat{v_j}). \tag{8}$$

Now, given these local distances, we apply *symmetric* DTW on whole sequences $\mathbf{u}$ and $\mathbf{v}$ and get the minimum global distance from the nonlinear aligned Viterbi path Γ with

$$D_{\mathrm{global}}(\mathbf{u}, \mathbf{v}) = \min_{\Gamma} \frac{1}{||\Gamma||} \sum_{p=1}^{||\Gamma||} d_s(\hat{u_p}, \hat{v_p}). \tag{9}$$

This distance is then converted to the kernel

$$\kappa_{\mathrm{DTW}}(\mathbf{u}, \mathbf{v}) = \cos D_{\mathrm{global}}(\mathbf{u}, \mathbf{v}) \tag{10}$$

which is symmetric if the symmetric version of DTW is used, which is a requirement for use in SVM classification. In order

TABLE III
CLASSIFIER ACCURACIES AVERAGED OVER SPEAKERS WITH DYSARTHRIA (BEST OF ROW IN BOLD) FOR AF RECOGNITION

		Average accuracy (%)									
			DBN-F			NN		SVM			
Feature	Input	HMM	default	sparse	LDCRF	MLP	ELM	RBF	DTW	μ	σ
Manner		23.8	36.5	32.1	**69.1**	22.1	30.2	66.8	65.4	43.3	19.0
Place		33.9	39.6	34.7	**58.8**	35.5	41.9	**58.3**	56.5	44.9	10.4
Hi/Low		48.6	52.9	49.0	56.2	53.0	**58.7**	55.7	55.9	53.8	3.3
Voice	MFCC (window)	76.1	77.8	76.3	79.2	78.7	**81.3**	76.8	78.1	78.0	1.6
Front/Back		49.0	48.4	49.4	54.0	48.2	52.1	**55.1**	**55.7**	51.5	2.9
Round		60.4	64.5	60.6	64.8	68.9	**69.7**	55.3	54.0	62.3	5.4
Static		61.3	65.2	63.6	**70.2**	64.2	66.5	67.3	69.2	65.9	2.8

for the quadratic programming problem to have a definite solution, the kernel must either be a valid dot product [61], or satisfy Mercer's condition, which is to say that given a real-valued kernel $\kappa(x, y)$, all square integrable functions $g(x)$ will give $\int\int \kappa(x,y)g(x)g(y)dx\,dy \geq 0$ [62]. While the cosine over an aggregate of sequences is not strictly a dot-product, it has been shown to be empirically useful in speech classification nonetheless [58].

E. Dynamic Bayes Networks (DBNs)

We are not bound to learning relationships between inputs and outputs by training the parameters of an otherwise "black box," but are free to explicitly provide the topological relationships between relevant variables in our models, which can include measurements of kinematic data. Bayes networks provide a popular statistical framework that allows us to determine precise instantaneous conditional relationships. Traditional Bayesian learning is restricted to universal or immutable relationships and does not model dynamic systems or time-varying relationships. Dynamic Bayes networks (DBNs) are directed acyclic graphs connecting random variables that generalize the stochastic mechanisms of Bayesian learning to time sequences. Given an N-variable observation sequence $Z_{1:T}^{(1:N)}$ of arbitrary length T, its likelihood is computed by "unrolling" a 2-frame DBN to T frames, and multiplying all posteriors

$$P(Z_{1:T}^{(1:N)}) = \prod_{i=1}^{N} P_{B_1}\left(Z_1^{(i)}|\text{par}\left(Z_t^{(i)}\right)\right) \times \prod_{t=2}^{T}\prod_{i=1}^{N} P_{B_arrow}\left(Z_t^{(i)}|\text{par}\left(Z_t^{(i)}\right)\right) \quad (11)$$

where conditional distributions, B_arrow are drawn over adjacent frames in time for the ith state at time t, $Z_t^{(i)}$ by $P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^{(i)}|\text{par}(Z_t^{(i)}))$, given the parents of $Z_t^{(i)}$, $\text{par}(Z_t^{(i)})$. This temporal model generalizes both the hidden Markov model and the Kalman filter [63]. Given a specified topology between variables and a data set D, the posterior distribution over the model parameters θ is learned either with maximum likelihood for fully observed sequences, or with expectation-maximization given hidden variables, enabling state-based methods [64].

In all graphical depictions of DBNs, filled and empty nodes represent observed and hidden variables, respectively. Square and round nodes are discrete and continuous, respectively.

V. EXPERIMENT SET 1: ACOUSTICS ALONE

We begin by considering the effects of dysarthria in systems trained solely from acoustic data, which is a considerably more common scenario than one in which kinematic data are available. However, given phonemic annotations, we can infer articulatory features as representative of articulatory knowledge, as described in Section I-B. We train each classifier both to identify articulatory features from acoustics and to identify phones given both acoustics and their identified AFs. In all cases, acoustic data are sampled at 16 kHz and converted to 42-dimensional feature vectors of Mel-frequency cepstral coefficients (MFCC) consisting of 0th- to 12th-order cepstral coefficients, log energy, and δ and $\delta\delta$ coefficients. Neither δ nor $\delta\delta$ observations are appended to AF components, due to the relative parsimony of tracking changes in step functions. We apply tenfold cross-validation on random permutations of 90% training and 10% test data for each speaker in the Nemours database. Training sets consist of approximately 93 000 frames per speaker on average.

We test two topologies of AF variables within DBNs. The first is based on similar work by Frankel *et al.* [21], and is shown in Fig. 3(a). The second is a sparser version of that DBN with certain conditional dependencies removed, as shown in Fig. 3(b). All AFs are observed in the DBN during training but inferred during testing.

A. AF Classification With Acoustics

Frame-level accuracies for each AF averaged over all speakers in the Nemours database are summarized in Table III for each classifier. Both the LDCRF and SVM methods are exceptionally proficient at classifying *Manner* and *Place*, which are highly related, and poor at classifying the *Round* AF despite its low cardinality. This suggests that there is some other aspect of those AFs that affects discriminability, at least for SVMs. The *nil* class is the most poorly recognized in three of the four AFs having it. The most frequently confused pairs for each AF are shown in Table IV, which is generally consistent with the literature for speakers without dysarthria [16].

In general, SVM methods outperform NN on average by 4.9% to 9.3% absolute and provide a 19.8% relative error reduction on dysarthric speech. On the control subject, AF models achieved 74.3% accuracy for MLP, and 77.6% for RBF, on average. Results of the SVM methods with this speaker were comparable though slightly lower than in similar research on non-dysarthric AF recognition by SVM [65], although that work included far more training data. Other research on speaker-independent recurrent neural networks for AF recognition on regular speech

TABLE IV
MOST FREQUENT ERRORS FOR EACH AF ([ACTUAL] → [HYPOTHESIS] (% TOTAL ERROR))

Feature	1^{st}	2^{nd}
Manner	[vowel]→[approx.] (12%)	[vowel]→[retro.] (8%)
Place	[nil]→[alv.] (10%)	[nil]→[dental] (7%)
Hi/Low	[nil]→[low] (14%)	[mid]→[low] (11%)
Voice	[unvoiced]→[voiced] (68%)	[voiced]→[unvoiced] (32%)
Front/Back	[nil]→[central] (19%)	[nil]→[back] (17%)
Round	[non]→[nil] (26%)	[nil]→[non] (22%)
Static	[stat.]→[dyn] (54%)	[dyn]→[stat] (46%)

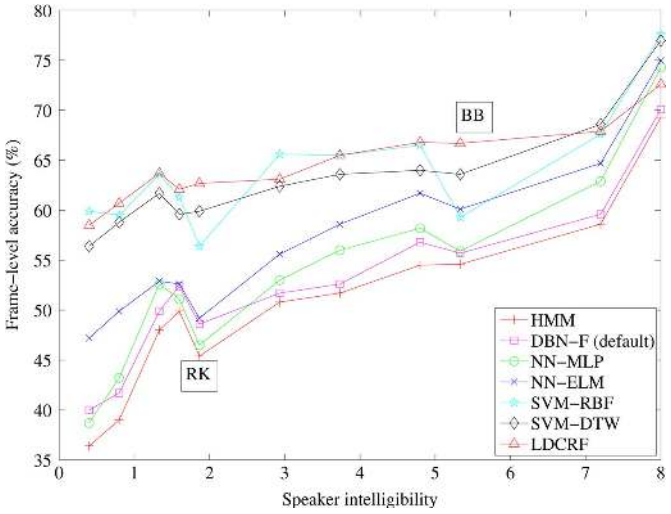

Fig. 5. Average classifier accuracy against assessed intelligibility level.

report frame-level accuracies between 85.9% and 91.8% given ∼2.2 million frames [21].

1) Effects of Dysarthria: Fig. 5 shows the overall accuracy of each classification technique according to speaker intelligibility as determined by the Frenchay Dysarthria Assessment (see Section I-A). These results show a general success of SVM and LDCRF methods across all speakers, especially the less intelligible ones, and a global increase in accuracy with intelligibility. Two speakers perturb this trend, however, with noticeable drops in accuracy as indicated for speakers "RK" and "BB" in the figure. These two individuals share exceptionally poor tongue elevation and lateral movement relative to the rest of the group which seems to account for their especially low accuracy with *High/Low* and *Front/Back* AFs, which are predicated on tongue movement and position. According to their Frenchay assessments, "RK" and "BB" both had scores of 0/9 for tongue elevation and scores of 0/9 and 1/9 for lateral tongue movement, respectively. Only two other speakers, "SC" and "BK," had similarly poor assessments of tongue control, with the latter also having the lowest intelligibility of all speakers. Table V shows the recognition accuracies for the two AFs under consideration against the average of all other AFs given an HMM system. Here, the four speakers identified as having particularly bad tongue movement have recognition accuracies for *Front/Back* and *High/Low* that are all between 5.3% and 10.2% lower than for other AFs, on average. By contrast, *Front/Back* and *High/Low* AFs are better recognized than other AFs, on average, for all speakers without the identified tongue deficit.

TABLE V
RECOGNITION ACCURACIES (% CORRECT) OF *FRONT/BACK* AND *HIGH/LOW* AFS COMPARED WITH THE AVERAGE RECOGNITION ACCURACIES ACROSS ALL OTHER AFS FOR FOUR SPEAKERS AND THE AVERAGE OF ALL OTHER SPEAKERS GIVEN AN HMM RECOGNITION SYSTEM

	Front/Back	High/Low	avg. other AF
BK	31.2	32.5	37.8
SC	35.3	34.7	41.3
RK	37.1	36.9	47.1
BB	48.6	49.0	55.8
avg. of others	55.3	54.5	54.2

TABLE VI
PHONE CLASSIFICATION ACCURACIES (%) AT THE FRAME LEVEL AVERAGED OVER SPEAKERS WITH DYSARTHRIA GIVEN VARIOUS TYPES OF OBSERVATION. ESTIMATED AFS ARE COMBINED WITH MFCC OBSERVATIONS EITHER BY USING AF ESTIMATORS OF THE SAME TYPE (MFCC+AF) OR BY USING THE LDCRF AF ESTIMATOR ($\text{MFCC} + \text{AF}_{\text{LDCRF}}$)

	Input type			
	MFCC	AF	MFCC+AF	MFCC+AF_{LDCRF}
HMM	33.8	7.4	36.3	37.6
DBN-F (default)	34.1	7.8	37.1	37.9
DBN-F (sparse)	33.4	7.5	37.0	38.1
LDCRF	41.2	16.0	41.5	41.5
NN-MLP	31.9	5.8	34.8	35.3
NN-ELM	36.7	11.7	40.2	40.7
SVM-RBF	38.4	16.2	38.7	40.1
SVM-DTW	39.6	17.9	41.0	41.3

Within these AFs, follow-up analysis revealed linear correlation coefficients up to 0.95 between increased formant deviation and decreased tongue function. While overall intelligibility may be useful in predicting general trends in Fig. 5, it is an aggregate measure of the functions of component articulators, and may be overridden for speakers having more localized disabilities.

B. Phone Recognition With Acoustics

Finally, we consider whether AFs are useful in identifying phones. For each of our modeling techniques, we construct three triphone classifiers that differ by the nature of their observations. Each of these is trained either with acoustics, with estimated AFs, or with acoustics and estimated AFs concatenated together. Here, AF estimates are derived both from the outputs of models having the same type as the phone classifier, or from the outputs of the LDCRF model which represents the best average AF estimates achievable. No other heterogeneous combination of models is attempted. Given that the LDCRF is the most accurate AF classifier, we find it unlikely that other combinations would yield much greater accuracies.

All models are applied over whole unsegmented utterances as continuous tasks. Specifically, each frame of speech is classified by NN and SVM methods given short windows of input observations, as described earlier. Connected-state models of the same type (i.e., either HMM, LDCRF, and DBN) are connected together so that all phonemes are equally likely to follow all others. This frame-based approach is taken to evaluate these models as substitutes of standard acoustic models, as is our intention. The use of language models is explored in Section VI-D. Accuracy is measured at the frame level by converting estimated triphones to their monophonic roots.

The results in Table VI indicate relative error reductions of 8.8% and 11.2% merely by replacing an HMM model with an

SVM-DTW and an LDCRF, respectively, given only dysarthric acoustics, which is significant at the 99% confidence level. Here, relative error reduction is the absolute difference between the error rates of the two systems under comparison divided by the higher error rate of the two. Extending observation vectors to include AFs reduces error relatively by between 0.5% and 7.1% over associated acoustic-only models, which represent significant improvements at the 99% confidence level for all models except LDCRF. This result shows a clear benefit of incorporating AFs into the input of all but one type of acoustic model. Since the seven AFs are so rarely unanimously correct, they alone cannot be used to infer the respective phone in practice, and further research should investigate whether it is more useful to limit the use of AFs to some subset. No explicit weighting was applied between the MFCC and AF components of heterogeneous vectors, but the relative importance of these parts and their covariances are inferred during training by each of these classifiers implicitly.

VI. Experiment Set 2: Initialization With Articulatory Models

There is increasing evidence that replacing the Gaussian mixture observation densities of HMMs with limited Bayes nets representing spacial vocal tract kinematics can improve accuracy over acoustic-only models for speakers without dysarthria [44]. Although it is impractical to perform articulography on each speaker we wish to model, we can make use of publicly available databases such as MOCHA or TORGO to provide baseline kinematic knowledge that we can adapt to speakers for whom only acoustic data are available. This scenario is explored in this section.

We conflate the instantaneous EMA position data from the MOCHA and TORGO databases (see Section III-C) by first reducing their dimension to $N_p = 4$ or $N_p = 8$ principal components by singular value decomposition specific to each phone in which $K = 4$, $K = 8$, or $K = 16$ mean vectors are computed according to the sum-of-squares error function. During training, the DBN variable $\mathbf{A}$ is the observed index of the mean vector nearest to the current frame of EMA data at time t. During inference, this variable is hidden and we marginalize 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. Similarly, we follow the same procedure on the velocities and accelerations of the articulators, producing indices $\mathbf{A_v}$ and $\mathbf{A_a}$. These variables are used in alternative DBN topologies DBN-A2 and DBN-A3. In the first, the observation vector is trisected, with each 14-dimensional vector (i.e., MFCC, δ, and $\delta\delta$) being conditioned on $\mathbf{P}$, $\mathbf{Q}$, and one of $\mathbf{A}$, $\mathbf{A_v}$ and $\mathbf{A_a}$. The second alternative structure, DBN-A3, conditions $\mathbf{A_a}$ on $\mathbf{A_v}$, and $\mathbf{A_v}$ on $\mathbf{A}$ and conditions the 42-dimensional observation vector on all variables. The three kinematic DBN topologies are shown in Fig. 6.

The MOCHA database uniquely includes velum position and the TORGO database uniquely includes left and right lip corners. Both databases include three midsagittal tongue positions, upper and lower lip, and lower incisor positions.

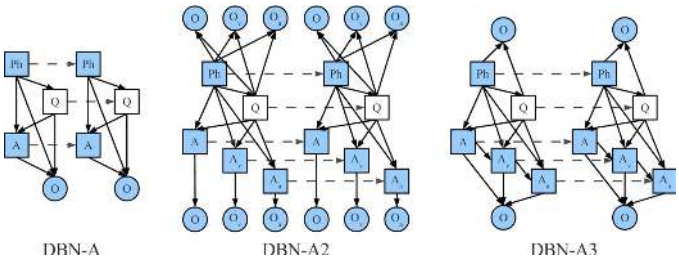

Fig. 6. Two-frame dynamic Bayes networks with EMA measurements differing by their connectivity. Nodes $\mathbf{Ph}, \mathbf{Q}, \mathbf{O}, \mathbf{A}, \mathbf{A_v}$, and $\mathbf{A_a}$ represent phoneme, state, MFCC observations, and EMA position, velocity, and acceleration, respectively. Inter-frame conditional links are dashed for clarity.

TABLE VII
Accuracies of Frame-Level Phone Recognition Across Kinematic DBNs With Varying Quantities of Principal Components N_p and Gaussians K for Speaker-Dependent, Non-Dysarthric Speech. Data Are Obtained From the MOCHA and TORGO Databases

		DBN-A		DBN-A2		DBN-A3	
N_p	K	MOC.	TOR.	MOC.	TOR.	MOC.	TOR.
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
	16	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

A. Recognition With Non-Dysarthric Speech

The three DBN models are compared on non-dysarthric speech across the number of principal components N_p and the number of Gaussians K used in quantization. Reducing dimensionality across heterogeneous acoustic/articulatory observations in this way has previously been shown to preserve important features of both articulation and acoustics [40], [66]. Results of frame-level phone recognition are summarized in Table VII. Across all topologies and data, $N_p = 16$ 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.

B. Retraining Dysarthric Acoustics

We retrain models initialized on non-dysarthric data given new dysarthric acoustics. We retrain each kinematic DBN with dysarthric acoustics by making indices $\mathbf{A}$, $\mathbf{A_v}$, and $\mathbf{A_a}$ hidden after training on non-dysarthric acoustic/articulatory data (MOCHA and TORGO), and retraining on dysarthric acoustics (Nemours and TORGO). All HMM and kinematic DBN models are trained with EM and smoothed junction-tree inference, given their hidden variables. When retraining the HMM, DBN, NN, and LDCRF models to dysarthric speech, we initialize new instantiations with the distributions learned on regular speech and retrain on speaker-specific acoustics until convergence. All training of the fully observed DBN-F is with maximum likelihood, so adaptation involves concatenating the non-dysarthric and dysarthric training data and learning once. SVM models from previous sections are not included here, due to the dissimilar manner in

TABLE VIII
AVERAGE FRAME ACCURACY (%) OF CORRECTLY LABELED PHONES OF SPEAKER-DEPENDENT AND SPEAKER-RETRAINED (EMA-INITIALIZED) MODELS, ACCORDING TO THE SEVERITY OF DYSARTHRIA

		sev	mod	mild	ctrl
HMM	Depend.	14.1	27.8	51.6	72.8
	Retrain.	16.8	32.1	58.9	-
LDCRF	Depend.	15.2	28.0	51.8	73.5
	Retrain.	16.8	32.4	59.1	-
DBN-F	Depend.	15.0	28.0	51.6	73.3
	Retrain.	16.7	32.3	59.4	-
DBN-A	Depend.	16.4	31.1	54.2	73.6
	Retrain.	16.2	31.7	58.3	-
DBN-A2	Depend.	16.3	31.1	54.3	73.6
	Retrain.	16.3	31.9	58.4	-
DBN-A3	Depend.	16.4	31.3	54.5	73.8
	Retrain.	16.5	32.0	58.7	-
NN-MLP	Depend.	15.5	28.6	51.4	72.6
	Retrain.	16.0	29.0	58.6	-
NN-ELM	Depend.	15.6	30.5	51.2	72.7
	Retrain.	16.1	30.7	57.5	-

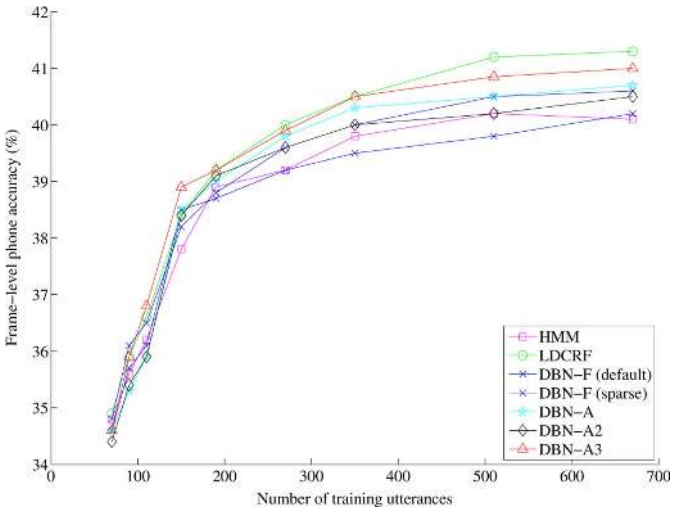

Fig. 7. Labeling accuracy with increasing amount of dysarthric retraining.

which those models are trained. In all cases, training data include all phones observed during testing and is applied to the 46 phones that MOCHA, Nemours, and TORGO have in common. Data are randomly split into 90% training and 10% test data. We split all dysarthric data from Nemours and TORGO into three categories according to the level of intelligibility as determined by the Frenchay assessment [12]. Individuals with intelligibility levels between 0 and 25% are "severe," between 25% and 62.5% are "moderate," and between 62.5% and 87.5% are "mild." are considered severely and moderately dysarthric, respectively.

Table VIII shows the frame-level accuracy of unsegmented phone labeling on speaker-dependent and speaker-retrained distributions for each model, according to the severity of dysarthria. Here, DBN-A, -A2, and -A3 are trained to mixtures of 16 Gaussian clusters determined by unreduced (16-dimensional) articulatory data. These results show an increasing benefit of retrained over dependent training on dysarthric speech as intelligibility increases, with absolute rates of improvement of 0.86%, 1.96%, and 6.03% on severely, moderately, and mildy dysarthric speech, respectively. Although speaker-dependent kinematic models are more successful than other models, they do not adapt as well as the DBN-F or LDCRF models.

These results are generally consistent with similar work that retrained acoustic-only DBNs to Japanese kinematic data [44] over 1 or 2 iterations of EM. That work showed error reduction of between 0.7% and 3.8% on phone classification among a selection of alternative DBNs relative to a baseline DBN. The performances of DBN-F and HMM are also consistent with similar work on non-dysarthric models [21].

C. Effect of Sample Size

We examine the effect of increased sample size by retraining non-dysarthric models with cross-sections of data selected uniformly at random among all speakers with dysarthria in Nemours and TORGO, and testing on proportionally increasing test sets. Fig. 7 suggests that as the amount of dysarthric speech is increased, the LDCRF model outperforms all others, with an absolute error reduction of 1.2% over HMM with 670 training utterances for retraining.

D. Use of Language Models

Although this work is concentrated on articulatory enhancements to acoustic models, in practice the latter are rarely used alone without some contextual information. Often, bigrams are used in order to weigh the likelihood of transitioning from one phoneme or word to another. Since our data consist of many single-word utterances, we consider phoneme bigrams in which the probability of one phoneme p_t following another p_{t-1} at time t is given by

$$P(p_t|p_{t-1}) = \frac{N_{(p_{t-1},\, p_t)}}{N_{(p_{t-1})}}$$

where $N_{(p_{t-1})}$ is the total number of occurrences (i.e., whole sequences of frames) of p_{t-1} in the data and $N_{(p_{t-1},\, p_t)}$ is the total number of times p_t immediately follows p_{t-1} in the data. We gather these counts from TIMIT which includes 2472 unique bigrams covering 172 460 adjacent pairs of phonemes, as determined by the included phonetic annotations. Similarly, the unigram probability of phoneme p_t is determined from the same data by

$$P(p_t) = \frac{N_{(p_t)}}{\sum_\rho N_{(\rho)}}$$

where ρ is iterated over all 61 phonemes in the training data.

In order to implement systems that incorporate either bigram and unigram information, we first train individual HMM and DBN-A models for each phoneme, as before, where training data consist of whole sequences of phonemes. The result is 61 HMMs and 61 DBN-A models, each consisting of three states with reflexive and left-to-right transitions. We first connect 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. First, the probabilities associated with these transitions are their bigram probabilities of equation VI-D. Expectation–maximization is then performed for two iterations on each of the large connected HMM and DBN-A models in order to learn reflexive transition

TABLE IX
AVERAGE FRAME-LEVEL ACCURACY (%) OF UNSEGMENTED PHONEME LABELING GIVEN ERGODIC HMMS AND DBN-AS WITH UNIGRAM AND BIGRAM PHONEME TRANSITION PROBABILITIES

Severity	HMM		DBN-A	
	unigram	bigram	unigram	bigram
sev	17.2	20.8	17.4	21.0
mod	33.4	37.3	34.1	37.9
mild	60.1	63.5	60.5	63.7
ctrl	74.0	74.2	74.2	74.6

probabilities on the last state for each phoneme without over-fitting. This is a common approach producing all-phoneme ergodic models [67]. This process is then repeated, but with initial transition probabilities between phoneme models being derived from their unigram probabilities (equation VI-D).

Given these connected models, the same data as in Section VI-B is used to measure the average proportion of correctly labeled phones given phoneme models trained by the speaker-dependent method. Table IX shows the frame-level phoneme recognition accuracies of each model across the same speaker intelligibility levels of Table VIII. While there are clear improvements in accuracy, these are still lower than one would expect if full word-level bigrams were used, given more training data. Trigram models were not attempted due in part to this relative sparsity of data and to inherent constraints of the implementation.

VII. DISCUSSION

Preceding experiments have concentrated on recognition tasks across an inventory of classifiers. This section explores possible explanations for some of the behavior observed in those experiments.

A. Synthesizing Dysarthric Acoustics

We compare the generative abilities of DBN-A and DBN-F on our data. We iteratively set **Ph** to each phone in the available DBN-A and DBN-F models and marginalize over all other variables to get the distribution on **O** from which we sample virtual data for each phone. These generated likelihood functions are fitted with Gaussians and compared with the true MFCC distributions of each phone by means of Kullback–Leibler relative divergence. The likelihood functions generated by DBN-F diverge from true distributions by a factor of 0.22016 on regular speech and by 0.2246 on dysarthric speech. However, while virtual DBN-A data diverge from true data by a factor of 0.1690 for regular speech, speaker-retrained DBN-As for dysarthric speech diverge by 0.3378, on average, from true phone MFCC distributions. This disparity is exemplified in Fig. 8.

B. Statistical Transformation of Articulator Space

In order to better understand some recognition results, we relate the distributions of the vowels in acoustic and articulatory spaces across dysarthric and non-dysarthric speech. Vowels in acoustic space are characterized by the steady-state positions of the first two formants as determined automatically by applying pre-emphasis and the Burg algorithm [68]. Vowels in articulatory space are characterized by the positions of the articulators when their accelerations are minimum. We fit Gaussians to these data, as exemplified in Fig. 9 for the most frequent vowels in TORGO and compute the entropy of the data within these distributions. Surprisingly, the entropies of these distributions were relatively consistent across dysarthric (34.6 nats) and non-dysarthric (33.3 nats) speech, with some exceptions (e.g., *iy*). However, vowel spaces overlap considerably more in the dysarthric case signifying that, while speakers with CP can be nearly as consistent as speakers without dysarthria in the acoustic space, the locations of their targets in that space are not as discernible. Moreover, we note linear correlation coefficients of over 0.95 between F2 standard deviation and the extent of tongue protrusion, as determined by the Frenchay assessment described above.

In an attempt to tease apart the acoustic targets in dysarthric speech, and to give them meaningful conditioning articulatory variables within the DBN framework, we learn statistical mappings between dysarthric and non-dysarthric speech. Namely,

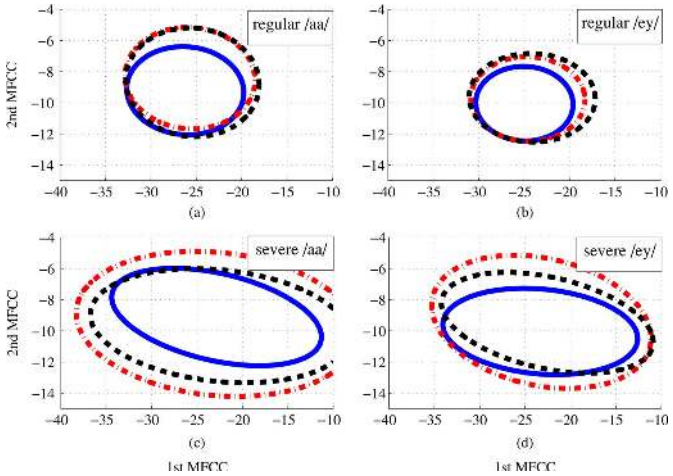

Fig. 8. Contours representing two standard deviations of Gaussians fitted to real data (solid line), samples from DBN-F (dashed line), and samples from DBN-A (dash-dotted line) on the first two mel-frequency cepstral coefficients. Subfigures represent (a) regular speech (/aa/), (b) regular speech (/ey/), (c) severely dysarthric speech (/aa/), and (d) severely dysarthric speech (/ey/).

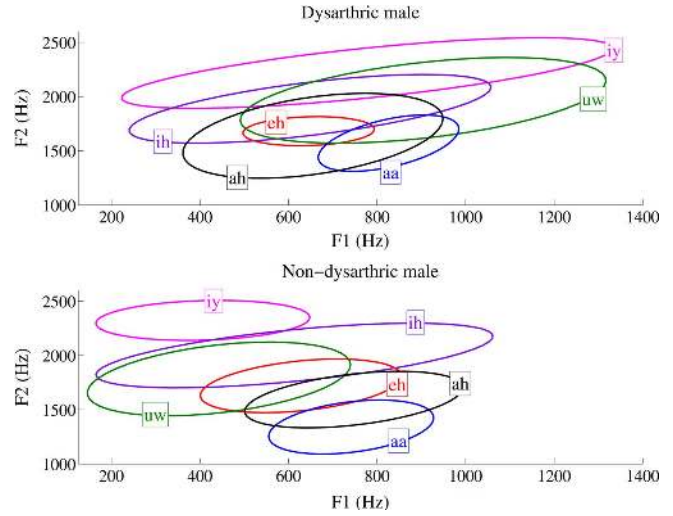

Fig. 9. Contours showing first standard deviation in F1 versus F2 space for distributions of the six of the most frequent vowels in continuous speech for the dysarthric and non-dysarthric males from the TORGO database.

TABLE X
TRAINING DATA ARE A COMBINATION OF TRANSFORMED REGULAR ACOUSTICS AND ARTICULATION, AND DYSARTHRIC ACOUSTICS AND ARTICULATION

Training Data	Retraining Data	Testing Data	Accuracy (%)
Trans. acous. ∪ Trans. artic.	-	Trans. acous.	72.9
		Dys. acous.	72.6
	Dys. acous.	Trans. acous.	73.7
		Dys. acous.	73.4
	Dys. acous. ∪ Dys. artic.	Trans. acous.	74.3
		Dys. acous.	74.2

we learn two functions, f and g, which produce the expected frames in the acoustic and articulatory spaces of a speaker with dysarthria given corresponding frames for a regular speaker. For each function, we define Gaussian distributions $\mathbf{N}(\cdot)$ for each phone p by the means of the regular and dysarthric speech, respectively, $\mu_p^{(x)}$ and $\mu_p^{(y)}$, and the covariances, $\Sigma_p^{(xx)}$, of the regular speech. We can then apply the following statistical transformation function between non-dysarthric acoustic vectors $\mathbf{x}$ and their dysarthric counterparts y:

$$f(\mathbf{x}) = E(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{P} h_i \left[\mu_i^{(y)} + \Sigma_i^{(yx)} \left(\Sigma_i^{(xx)}\right)^{-1} \left(\mathbf{x} - \mu_i^{(x)}\right)\right] \tag{12}$$

where

$$h_i(x) = \frac{\alpha_i N\left(x; \mu_i^{(x)}, \Sigma_i^{(xx)}\right)}{\sum_{j=1}^{P} \alpha_j N\left(x; \mu_j^{(x)}, \Sigma_j^{(xx)}\right)} \tag{13}$$

where α_p is the proportion of the occurrences of phone p in the data, and $\Sigma_p^{(yx)}$ is the cross-covariance matrix in phone p across speakers with and without dysarthria. The function g is identical in articulatory space, but with vectors defined by articulator positions from EMA. We learn cross-covariance matrices on aligned sequences from both sets of speakers. Since each speaker in the TORGO database recites the same set of phrases, we achieve frame-by-frame alignment by applying dynamic time warping on corresponding acoustic segments of pre-annotated speech, and applying the resulting alignment on the raw articulatory data. This is effectively the reverse of the approach suggested by Hosom *et al.*, who propose transforming dysarthric acoustic space to regular acoustic space in order to be made more intelligible [69].

Once we have the transformed acoustic and articulatory spaces of a control subject that resemble those of our speaker with dysarthria, we quantize the latter using k-means clustering and train the DBN-A model as described in Section VI. We then update this model given either dysarthric acoustics only (see Section VI-B), or aligned dysarthric acoustics and quantized articulation. These three models are then tested with either additional transformed acoustics, or actual dysarthric acoustics. These results are shown in Table X. Notably, models tested with the transformed speech show slightly higher accuracies of recognition than models tested on the target dysarthric speech, which may be an artifact of supersegmental effects of dysarthria on intelligibility. We note that models initialized with transformed regular speech perform better than any dependent or retrained combination for dysarthric test data in Section VI-B.

VIII. CONCLUSION

This paper summarizes an extensive series of experiments concerning the recognition of dysarthric speech given knowledge of speech production. Our purpose is to discover which combinations of articulatory knowledge and modeling give improved accuracies of recognition for individuals with speech disabilities. In general, these experiments include both theoretical and empirical representations of the vocal tract, with data obtained from the MOCHA database and from our own collection of dysarthric and non-dysarthric speech. In situations where no kinematic data are available, incorporating theoretical articulatory knowledge into generative dynamic Bayes networks shows some improvement in phone recognition over traditional HMM models, but far greater improvements are possible through the application of discriminative methods, particularly latent-dynamic conditional random fields. However, generative DBN models that are trained by aligned kinematic electromagnetic articulographic data give the greatest improvement over standard models, also outperforming acoustic-only discriminative methods. We have also explored a few aspects of dysarthric and articulatory data, including the severity of disablement and the statistical transformation between regular and dysarthric kinematics in retraining.

Although our results may be applicable to improving current ASR systems for the dysarthric population, these successes are tempered by the relatively unconstrained nature of the underlying statistical methods and the short-time observation windows. Several fundamental phenomena of dysarthria such as increased disfluency, longer sonorants, and reduced pitch control [48] cannot be readily represented in any of the methods described here. We are currently studying the articulatory dynamics of dysarthria in particular, and speech generally, within the context of dynamical systems. Specifically, we are exploring task-dynamic theory as a combined model of skilled articulator motion and the planning of vocal tract configurations [70], [71]. This theory introduces the notion that the dynamic patterns of speech are the result of overlapping gestures, which are high-level abstractions of goal-oriented reconfigurations of the vocal tract such as bilabial closure or velar opening. 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 [72], [73]. We believe that such a high-level model of the vocal tract may better represent co-articulatory and long-distant effects in dysarthric speech.

REFERENCES

[1] D. O'Shaughnessy, *Speech Communications—Human and Machine*. New York: IEEE Press, 2000.
[2] F. Rudzicz, "Applying discretized articulatory knowledge to dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP09)*, Taipei, Taiwan, Apr. 2009, pp. 4501–4504.
[3] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP09)*, Taipei, Taiwan, Apr. 2009, pp. 4605–4608.

[4] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. S. Huang, "Audiovisual phonologic-feature-based recognition of dysarthric speech," *Abstract*, 2006.

[5] "Augmentative communication incorporated (ACI)," Section 3: Clinical Aspects of AAC Devices 2007 [Online]. Available: http://www.augcominc.com/whatsnew/ncs3.html

[6] K. L. Moore and A. F. Dalley, *Clinically Oriented Anatomy*, 5th ed. Philadelphia, PA: Lippincott, Williams, and Wilkins, 2005.

[7] R. D. Kent and K. Rosen, "Motor control perspectives on motor speech disorders," in *Speech Motor Control in Normal and Disordered Speech*, B. Maassen, R. Kent, H. Peters, P. V. Lieshout, and W. Hulstijn, Eds. Oxford, U.K.: Oxford Univ. Press, 2004, pp. 285–311, ch. 12.

[8] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augment. Altern. Commun.* vol. 16, no. 1, pp. 48–60, Jan. 2000 [Online]. Available: http://dx.doi.org/10.1080/07434610012331278904

[9] R. Patel, "Control of prosodic parameters by an individual with severe dysarthria," Univ. of Toronto, Toronto, ON, Canada, Tech. Rep., Dec. 1998 [Online]. Available: http://vismod.media.mit.edu/pub/masters_paper.doc

[10] P. Raghavendra, E. Rosengren, and S. Hunnicutt, "An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems," *Augment. Altern. Commun. (AAC)*, vol. 17, no. 4, pp. 265–275, Dec. 2001.

[11] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *J. Commun. Disorders*, vol. 33, no. 5, pp. 391–428, 2000.

[12] P. M. Enderby, *Frenchay Dysarthria Assessment*. London, U.K.: College Hill, 1983.

[13] L. J. Ferrier, H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit, "Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition," *Augment. Alternat. Commun.* vol. 11, no. 3, pp. 165–175, Jan. 1995 [Online]. Available: http://dx.doi.org/10.1080/07434619512331277289

[14] G. N. Clements, "The geometry of phonological features," *Phonology Yearbook* vol. 2, pp. 225–252, 1985 [Online]. Available: http://www.jstor.org/stable/4419958

[15] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 333–353, Oct. 2000.

[16] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Univ. of Bielefeld, Bielefeld, Germany, July 1999.

[17] F. Metze, "Discriminative speaker adaptation using articulatory features," *Speech Commun.*, vol. 49, no. 5, pp. 348–360, 2007.

[18] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ: Prentice-Hall, Apr. 2001.

[19] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper & Row, 1968.

[20] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, and B. Woods, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2007)*, Honolulu, HI, Apr. 2007, pp. IV-621–IV-624.

[21] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," *Comput. Speech Lang.*, vol. 21, pp. 620–640, 2007.

[22] M. Wester, "Syllable classification using articulatory—Acoustic features," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 233–236.

[23] M. Wester, J. Frankel, and S. King, "Asynchronous articulatory feature recognition using dynamic Bayesian networks," in *Proc. Inst. Electron., Inf. Commun. Eng. Beyond HMM Workshop*, Kyoto, Japan, 2004, vol. 104, pp. 37–42.

[24] J. R. Westbury, *X-Ray Microbeam Speech Production Database User's Handbook*. Madison, WI: Waisman Center on Mental Retardation & Human Development, 1994.

[25] P. H. H. M. van Lieshout, A. Bose, P. A. Square, and C. M. Steele, "Speech motor control influent and dysfluent speech production of an individual with apraxia of speech and Broca's aphasia," *Clinical Linguist. Phonetics*, vol. 21, no. 3, pp. 159–188, Mar. 2007.

[26] Y. Yunusova, J. R. Green, and A. Mefferd, "Accuracy assessment for AG500, electromagnetic articulograph," *J. Speech, Lang., Hearing Res.*, vol. 52, pp. 547–555, Apr. 2009.

[27] C. Coleman and L. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria," *Augment. Altern. Commun.*, vol. 7, no. 1, pp. 34–42, Mar. 1991.

[28] J. M. Noyes and C. R. Frankish, "Speech recognition technology for individuals with disabilities," *Augment. Altern. Commun. (AAC)*, vol. 8, no. 4, pp. 297–303, 1992.

[29] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: Case study of dysarthric speech," *Augment. Altern. Commun. (AAC)* vol. 16, no. 3, pp. 186–196, Jan. 2000 [Online]. Available: http://dx.doi.org/10.1080/07434610012331279044

[30] C. Havstam, M. Buchholz, and L. Hartelius, "Speech recognition and dysarthria: A single subject study of two individuals with profound impairment of speech and motor control," *Logopedics Phoniatrics Vocology*, vol. 28, no. 10, pp. 81–90, Aug. 2003.

[31] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. ONeill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Med. Eng. Phys.*, vol. 29, no. 5, pp. 586–593, Jun. 2007.

[32] G. Jayaram and K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks," *J. Rehabil. Res. Develop.*, vol. 32, no. 2, pp. 162–169, 1995.

[33] E. Sanders, M. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of Dutch dysarthric speech: A pilot study," in *Proc. 7th Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002.

[34] P. D. Polur and G. E. Miller, "Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals," *Med. Eng. Phys.*, vol. 28, no. 8, pp. 741–748, Oct. 2006.

[35] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP J. Adv. Signal Process.*, 2009.

[36] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 723–742, Feb. 2007.

[37] O. Scharenborg, V. Wan, and R. K. Moore, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Commun.*, vol. 49, no. 10–11, pp. 811–826, Oct.–Nov. 2007.

[38] O. Fujimura, "Relative invariance of articulatory movements: An iceberg model," in *Invariance and Variability of Speech Processes*, J. Perkell and D. Klatt, Eds. Hillsdale, NJ: Erlbaum, 1986, pp. 226–242, ch. 11.

[39] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (Version 3.4)," . Cambridge, U.K., 2006.

[40] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. Int. Conf. Spoken Lang. Process.*, Beijing, China, 2000.

[41] T. Fukuda, W. Yamamoto, and T. Nitta, "Distinctive phonetic feature extraction for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2003)*, Hong Kong, Apr. 2003, vol. 2, pp. 25–28.

[42] A. V. Nefian, L. Liang, X. Liu, X. Pi, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, 2002.

[43] T. A. Stephenson, M. Magimai-Doss, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 189–203, May 2004.

[44] K. Markov, J. Dang, and S. Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Commun.*, vol. 48, no. 2, pp. 161–175, Feb. 2006.

[45] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzjo, and H. Bunnell, "The nemours database of dysarthric speech," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, Philadelphia, PA, Oct. 1996.

[46] V. Zue, S. Seneff, and J. Glass, "Speech database development: TIMIT and beyond," in *Proc. ESCA Tutorial and Research Workshop Speech Input/Output Assessment and Speech Databases (SIOA-1989)*, Noordwijkerhout, The Netherlands, 1989, vol. 2, pp. 35–40.

[47] A. Wrench, "The MOCHA-TIMIT Articulatory Database," Univ. of Edinburgh. Edinburgh, U.K., Nov. 1999 [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

[48] F. Rudzicz, P. van Lieshout, G. Hirst, G. Penn, F. Shein, and T. Wolff, "Towards a comparative database of dysarthric articulation," in *Proc. 8th Int. Seminar Speech Production (ISSP'08)*, Strasbourg, France, Dec. 2008.

[49] M. Craig, P. van Lieshout, and W. Wong, "Suitability of a UV-based video recording system for the analysis of small facial motions during speech," *Speech Commun.*, vol. 49, no. 9, pp. 679–686, Sep. 2007.
[50] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2006)*, May 2006, vol. 3, pp. 1060–1063.
[51] Y. Yunusova, G. Weismer, J. R. Westbury, and M. J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *J. Speech, Lang. Hear. Res.*, vol. 51, pp. 596–611, Jun. 2008.
[52] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *J. Speech Hear. Disorders*, vol. 54, pp. 482–499, 1989.
[53] K. M. Yorkston and D. R. Beukelman, *Assessment of Intelligibility of Dysarthric Speech*. Tigard, OR: C.C. Publications, 1981.
[54] X. Menendez-Pidal and H. T. Bunnell, "Automatic phoneme labeler in the TIMIT database," *J. Acoust. Soc. Amer.* vol. 101, no. 5, pp. 3200–3201, 1997 [Online]. Available: http://link.aip.org/link/?JAS/101/3200/5
[55] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, Jun. 2007.
[56] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML-2001)*, San Francisco, CA, 2001, pp. 282–289, Morgan Kaufmann.
[57] J. L. Elman, "Finding structure in time," *Cognitive Sci.*, vol. 14, no. 2, pp. 179–211, 1990.
[58] V. Wan and J. Carmichael, "Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech 2005)*, Sep. 2005.
[59] P. Niyogi and C. Burges, Detecting and interpreting acoustic features with support vector machines Univ. of Chicago, Chicago, IL, Tech. Rep. TR-2002-02, 2002.
[60] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-classification by pairwise coupling," *Proc. Neural Inf. Process. Syst. 2003*, 2003.
[61] S. Russell and P. Norvig, *Articial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice-Hall, 2003.
[62] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.
[63] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. dissertation, Univ. of California at Berkeley, Berkeley, CA, 2002.
[64] Z. Ghahramani, "Learning dynamic Bayesian networks," in *Adaptive Processing of Sequences and Data Structures*. Berlin, Germany: Springer-Verlag, 1998, pp. 168–197.
[65] U. V. Chaudhari and M. Picheny, "Articulatory feature detection with support vector machines for integration into asr and phone recognition," in *IEEE Workshop Autom. Speech Recognition Understanding, 2009. ASRU 2009*, Merano, Italy, Nov. 2009, pp. 93–98.
[66] T. Fukuda and T. Nitta, "Noise-robust automatic speech recognition using orthogonalized distinctive phonetic feature vectors," in *Proc. Eurospeech-2003*, 2003, pp. 2189–2192.
[67] Y. Miyazawa, "An all-phoneme ergodic hmm for unsupervised speaker adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. ICASSP-93*, 1993, vol. 2, pp. 574–577, Aoruk.
[68] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[69] J.-P. Hosom, A. B. Kain, T. Mishra, J. P. H. van Santen, M. Fried-Oken, and J. Staehely, "Intelligibility of modifications to dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, Apr. 2003, vol. 1, pp. 924–927.
[70] E. M. Saltzman, *Task Dynamic Co-Ordination of the Speech Articulators: A Preliminary Model*. Berlin, Germany: Springer-Verlag, 1986, pp. 129–144.
[71] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecol. Psychol.*, vol. 1, no. 4, pp. 333–382, 1989.
[72] K. N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, P. B. Denes and E. E. David, Jr., Eds. New York: McGraw-Hill, 1972, pp. 51–66.
[73] K. N. Stevens and S. J. Keyser, "Quantal theory, enhancement and overlap," *J. Phonetics* vol. 38, no. 1, pp. 10–19, 2010 [Online]. Available: http://www.sciencedirect.com/science/article/B6WKT-4TWTJPV-1/2/4ef6e69008a78a141452409938d4d421

F. Rudzicz (S'09) received the B.Sc. degree in computer science from Concordia University, Montreal, QC, Canada, the M.Sc. degree in electrical and computer engineering from McGill University, Montreal, and the Ph.D. degree from the Department of Computer Science, University of Toronto, Toronto, ON, Canada.

His expertise includes parsing in natural language processing, acoustic modelling, multimodal interaction, and speech production.

Dr. Rudzicz is the recipient of a MITACS Accelerate Canada award and an NSERC Canada Graduate Scholarship.