Learning mixed acoustic/articulatory models for disabled speech

Frank Rudzicz
Department of Computer Science
University of Toronto
Toronto, Canada M5S 3G4
frank@cs.toronto.edu

Abstract

This paper argues that automatic speech recognition (ASR) should accommodate dysarthric speech by incorporating knowledge of the production characteristics of these speakers. We describe the acquisition of a new database of dysarthric speech that includes aligned acoustics and articulatory data obtained by electromagnetic articulography. This database is used to train theoretical and empirical models of the vocal tract within ASR which 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.

1 Introduction

Despite advances in automatic speech recognition (ASR) for the general population, many of the services they provide remain effectively unusable by individuals with severe speech disorders. One group of such disorders, called dysarthria, is caused by neuromotor disorders such as cerebral palsy, multiple sclerosis, and stroke. People with dysarthric speech are able to form and comprehend full sentences but lack control over the muscles used in speech, which is in turn unintelligible. The neurological damage that causes dysarthria also affects other physical activity and can slow keyboard input 150 to 300 fold in severe cases compared with regular users [1, 2]. By contrast, dysarthric speech is often only between 10 and 17 times slower than regular speech, at about 15 words per minute in the most severe cases [3]. Speech recognition is therefore an important access method for those affected by dysarthria.

Unfortunately, standard modelling techniques have failed to yield acceptable rates of recognition for speakers with dysarthria. Numerous attempts at modifying acoustic hidden Markov models have not raised word-recognition rates above 10% for speakers with severe dysarthria in contexts where speakers without dysarthria obtained 84.9% word-level recognition [4]. This paper argues the position that since dysarthria is an endogenous phenomenon, ASR models designed for this population need to incorporate knowledge of dysarthric articulation. We present the first database of dysarthric speech which includes 3D measurements of the vocal tract in section 2. Sections 3 and 4 respectively show that the use of articulatory data can reduce entropy theoretically and error empirically relative to acoustic-only representations. We conclude by discussing several aspects of dysarthric speech to which modern approaches in ASR are not amenable.

1.1 Representations for speech production

Articulatory features (AFs) are quantized abstractions of speech production according to distinctive configurations of the vocal tract. Here, articulatory features are collected into seven categories, each with a number of possible values. Parallelizing streams of information in this manner allows asynchronous modulation of speech acts across phoneme boundaries, which can partially account for co-articulation effects and speaker variability [5], which are particularly exacerbated in dysarthric speech. The features used here are based on those of Wester [6] and are listed in table 1.

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*

Table 1: Articulatory features, a description of their characteristics, and their possible values.

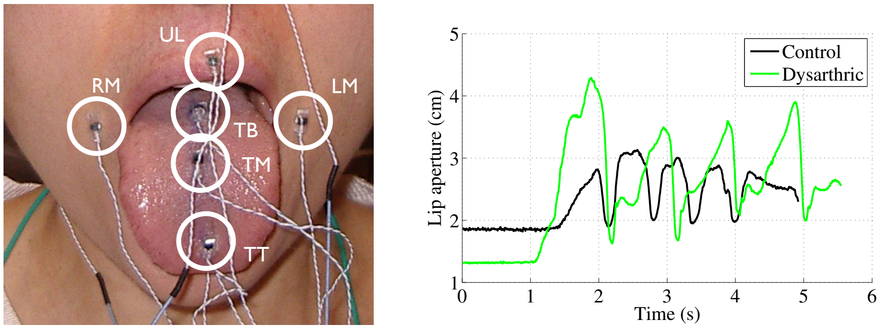
In the absence of AF annotations, values can be derived directly from phoneme annotations. Here, we assign to each MFCC frame a 7-dimensional vector of AF values based exclusively on the phoneme annotation at that frame. This assignment is derived directly from the phoneme-to-AF transformation table of Frankel *et al.* [5]. This incorporates recommendations by Wester *et al.* [7] in which the Front/Back feature includes the normally excluded *central* value, and diphthongs are split into their component vowels, which are mapped to their corresponding AFs. Unlike Frankel *et al.* [5], we label the Place feature of phonemes */b/* and */m/* as bilabial rather than labiodental.

2 The TORGO database of dysarthric articulation

The TORGO database of dysarthric articulation consists of time-aligned measurements of the vocal tract and the resulting acoustics for 7 speakers with cerebral palsy, one with amyotrophic lateral sclerosis, and age- and gender-matched controls. Vocal tract movement is measured by two systems. The first infers 3D positions of facial markers given stereo video sequences. The second uses electromagnetic articulography (EMA), in which the speaker is placed within a low-amplitude electromagnetic field. Tiny sensors within this field allow the inference of articulator positions and velocities to within 1 mm of error [8]. Here, measurement coils are placed as in other studies (e.g., the University of Edinburgh's MOCHA database [9]), namely on the upper and lower lip (UL and LL), lower incisor (LI), and tongue tip, middle, and back (TT, TM, and TB). Unlike many other studies, we record 3D rather that 2D midsagittal data, and also include left and right mouth corners (LM and RM) to measure lip rounding. The TORGO database is the first to measure dysarthric data with EMA and will be made public by the Linguistic Data Consortium in 2011. Figure 1(a) shows an example of this setup. Figure 1(b) shows the degree of lip aperture (i.e., the distance between UL and LL) over time for two speakers, one of whom has dysarthria. Here, the dysarthric speech is notably slower and has more excessive movement.

All articulatory data are smoothed with third-order median filtering in order to minimize measurement 'jitter' and are aligned with associated acoustic data by the recording mechanism. Acoustic noise is removed with spectral subtraction using minimum mean-square error estimation. Speech stimuli include random repetitions of phonetically balanced short sentences originally used in the TIMIT database [10], and pairs of monosyllabic words identified by Kent [11] as having relevant articulatory contrasts (e.g., *beat* versus *meat* as a stop-nasal contrast). Phoneme boundaries and pronunciation errors are transcribed by a speech-language pathologist to the TIMIT phone set.

Table 2 shows pronunciation errors according to manner of articulation for dysarthric speech as determined by a registered speech-language pathologist. Plosives are mispronounced most often, with substitution errors exclusively caused by errant voicing (e.g. */d/* for */t/*). By comparison, only 5% of corresponding plosives are mispronounced in regular speech. Furthermore, the prevalence



(a) (b)

Figure 1: (a) Electromagnetic articulograph system. (b) Lip aperture over time for four iterations of */ah p iy/* given a dysarthric and control speaker.

of deleted affricates in word-final positions, almost all of which are alveolar, does not occur in the corresponding non-dysarthric speech.

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

Table 2: Percentage 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.

3 Entropy reduction using articulation

Measuring the statistical disorder in the acoustic and articulatory data, as well as the *a posteriori* disorder of one given the other can suggest the relative merits of incorporating knowledge of articulatory behaviour into ASR systems for dysarthric speakers. Since our observations are continuous, we must use *differential entropy* defined by

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

where $f(X)$ is the probability density function of X. The differential entropy has known forms for a number of distributions $f(X)$, such as the multivariate normal,

$$\begin{aligned} f_X(x_1, ..., x_N) &= \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{N/2}\left|\Sigma\right|^{1/2}} \\ H(X) &= \tfrac{1}{2}\ln\left((2\pi e)^N \left|\Sigma\right|\right), \end{aligned} \tag{1}$$

where μ and Σ are the mean and covariances of the data. However, since our acoustic and articulatory data follow non-Gaussian distributions, we choose to represent these spaces by mixtures of Gaussians and we estimate the differential entropy of these distributions by the method of merging Gaussians incorporating equation 1. Namely,

$$\tilde{H}(X) = \sum_{i=1}^{L} \omega_i \left(-\log \omega_i + \tfrac{1}{2}\log((2\pi e)^N \left|\Sigma_i\right|\right),$$

where ω_i is the weight of the $i^{th}(1 \leq i \leq L)$ Gaussian and Σ_i is that Gaussian's covariance matrix [12], with L empirically set to 32. While differential entropies *can* be negative and not invariant under change of variables, other properties of entropy are retained [12], such as the chain rule for conditional entropy

$$H(Y \mid X) = H(Y, X) - H(X),$$

which describes the uncertainty in Y given knowledge of X. Here, we quantize entropy with the *nat*, which is the natural logarithmic unit.

We measure the differential entropy of acoustics ($H(Ac)$), of articulation ($H(Ar)$), and of acoustics given knowledge of the vocal tract ($H(Ac \mid Ar)$) in order to obtain theoretical estimates as to the utility of articulatory data. Table 3 shows these quantities across six speakers in the TORGO database. As expected, the acoustics of dysarthric speakers are much more disordered than for non-dysarthric speakers, however there is very little difference in the entropy of articulation among the speakers in terms of their statistical disorder. Although dysarthric speakers clearly lack articulatory dexterity, this implies that they nonetheless articulate with a level of consistency similar to their non-dysarthric counterparts, which is borne out in the clinical literature [13]. This consistency allows for significant reductions to the entropy of the resulting dysarthric acoustics, despite the fact that this equivocation $H(Ac \mid Ar)$ is an order of magnitude lower than for non-dysarthric speakers.

	Speaker	$H(Ac)$	$H(Ar)$	$H(Ac \mid Ar)$
Dysarthric	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	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

Table 3: Differential entropy, in nats, across dysarthric and control (non-dysarthric) speakers for acoustic ac and articulatory ar data.

In order to better understand these results, we compare the distributions of the vowels in acoustic space across dysarthric and non-dysarthric speech. Vowels in acoustic space are characterized by the steady-state positions of the first two formants (F1 and F2) as determined automatically by applying the pre-emphasized Burg algorithm [14]. We fit Gaussians to the first two formants for each of the vowels in our data, as exemplified in figure 2 and compute the entropy 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 acoustically consistent as non-dysarthric speakers, their targets in that space are not as discernible. Some research has shown larger variance among dysarthric vowels relative to our findings [15], which may partially be due to our use of natural connected speech as data, rather than restrictive consonant-vowel-consonant non-words.

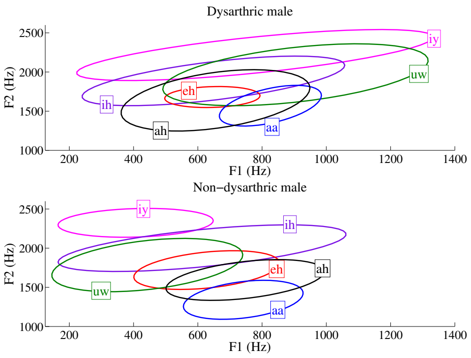


Figure 2: Contours showing first standard deviation of the first two formants for distributions of the six most frequent vowels in both dysarthric and non-dysarthric speech in the TORGO database.

4 Articulatory models in ASR for dysarthric speech

In this section we compare two discriminative methods (i.e., neural networks and support vector machines) that take acoustics alone against generative dynamic Bayes networks that incorporate articulatory information on the task of frame-level phone recognition with dysarthric speech. Given phonemic annotations for 46 phonemes in TORGO, we infer articulatory features as representative of articulatory knowledge, as described in section 1.1. These systems are compared against a tristate left-to-right HMM baseline with observation likelihoods at each state computed over mixtures of 16 Gaussians in typical Baum-Welch training and Viterbi decoding. 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.

In all cases, acoustic data are sampled at 16kHz and converted to 42-dimensional feature vectors of Mel-frequency cepstral coefficients (MFCC) consisting of 0^{th}- to 12^{th}-order cepstral coefficients, log energy, and δ and $\delta\delta$ coefficients. We conflate all dysarthric data together and all non-dysarthric data together and apply 10-fold cross-validation on random permutations of 90% training and 10% test data on each set.

4.1 Dynamic Bayes networks

Bayes networks provide a popular statistical framework that describes 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,

$$P(Z_{1:T}^{(1:N)}) = \prod_{i=1}^{N} P_{B_1}(Z_1^{(i)}|par(Z_t^{(i)}))\times \\ \prod_{t=2}^{T}\prod_{i=1}^{N} P_{B_\rightarrow}(Z_t^{(i)}|par(Z_t^{(i)})), \qquad (2)$$

where conditional distributions, $B_\rightarrow$ are drawn over adjacent frames in time for the i^{th} state at time t, $Z_t^{(i)}$ by $P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^{(i)}|par(Z_t^{(i)}))$, given the parents of $Z_t^{(i)}$, $par(Z_t^{(i)})$. This temporal model generalizes both the hidden Markov model and the Kalman filter [16]. 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 [17].

We test two DBN topologies with AF variables. The first AF DBN is based on similar work by Frankel *et al.* [5], and the second is a sparser version of that DBN with certain conditional dependencies removed, as shown in figure 3(a). All AFs are observed in the DBN during training but inferred during testing.

We also test three DBN topologies trained directly from EMA measurements. Here, we conflate the instantaneous EMA position data 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 DBN-A2, 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 $\mathbf{A_a}$. In 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 figure 3(b).

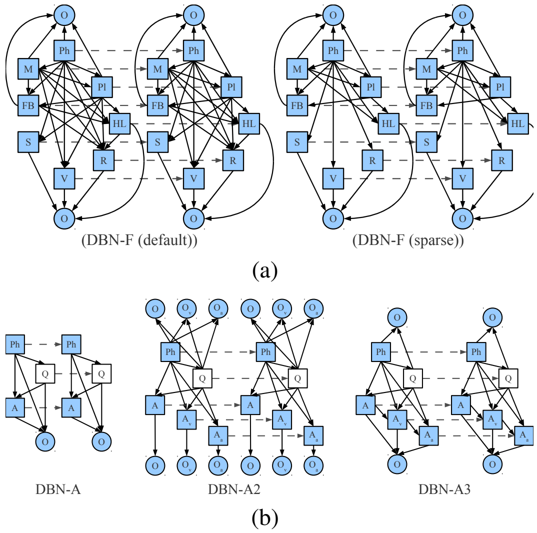


Figure 3: Two-frame dynamic Bayes networks with (a) theoretical articulatory features, and (b) EMA measurements differing by their connectivity. Nodes **Ph**, **Q**, **O**, **A**, $\mathbf{A_v}$, and $\mathbf{A_a}$ represent phoneme, state, MFCC observations, and EMA position, velocity, and acceleration, respectively. Inter-frame conditional links are dashed for clarity. Filled and empty nodes represent observed and hidden variables, respectively. Square and round nodes are discrete and continuous, respectively.

4.2 Neural networks

Despite their general popularity, neural networks (NNs) are rarely studied with regards to dysarthric acoustics, with some exceptions [18]. The two types of three-layer NN we consider here are the feed-forward multi-layer perceptron (MLP) and the recurrent Elman network (ELM) [19], which are primarily distinguished by the latter's time-delayed replication of the hidden layer as additional contextual input. The output of each NN consists of 46 nodes (i.e., the number of phonemes under consideration), and the i^{th} node is uniquely active when training the i^{th} value of that phoneme. The (non-contextual) input to each NN consists of 210 nodes in which the first 42 nodes are the values for the current MFCC vector, as defined above, and the subsequent 168 nodes are the values for the two previous and two following MFCC vectors in the sequence. In other words, input consists of the current MFCC and 2 frames of context on either side. All networks are fully connected between layers, with 500 hidden nodes as in similar work on non-dysarthric speech [5], and select the class having the highest posterior probability.

Activation functions at each node are tan-sigmoid (i.e., $a(x) = \left[2/\left(1+e^{-2x}\right)\right] - 1)$ 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.

4.3 Support vector machines

General maximum margin classifiers are of increasing interest in ASR due to their robustness against both sparse data and rapid transient changes in acoustic sequences [20]. Here, we use a soft-margin SVM and extend the process to k-class discrimination by training $k(k-1)/2$ binary classifiers, each delineating two class regions [21].

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 non-linear decision boundaries as follows:

$$\kappa_{RBF}\left(\mathbf{x},\mathbf{y}\right)=\exp\left(-\frac{\left(\mathbf{x}-\mathbf{y}\right)^2}{2\sigma^2}\right), \tag{3}$$

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 [20]. 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

$$\hat{u_i}=\frac{1}{\sqrt{u_i^2+\alpha^2}}\left[\begin{array}{c}u_i\\ \alpha\end{array}\right]. \tag{4}$$

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

$$d_s(\hat{u_i},\hat{v_j})=\theta_{\hat{u_i},\hat{v_j}}=\arccos(\hat{u_i},\hat{v_j}). \tag{5}$$

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

$$D_{global}(\mathbf{u},\mathbf{v})=\min_{\Gamma}\frac{1}{||\Gamma||}\sum_{p=1}^{||\Gamma||}d_s(\hat{u_p},\hat{v_p}). \tag{6}$$

This distance is then converted to the kernel

$$\kappa_{DTW}\left(\mathbf{u},\mathbf{v}\right)=\cos D_{global}(\mathbf{u},\mathbf{v}), \tag{7}$$

which is symmetric if the symmetric version of DTW is used, which is a requirement for use in SVM classification. In order for the quadratic programming problem to have a definite solution, the kernel must either be a valid dot product, 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 $\int\int\kappa(x,y)g(x)g(y)dx\;dy\geq 0$. 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 [20].

As for the NN models, input consists of the current MFCC frame and 2 frames of context on either side.

4.4 Experiments

All models are applied over whole unsegmented utterances as continuous tasks. Each frame of speech is classified by NN and SVM methods given short windows of input observations, as described above. Connected-state models (i.e., HMM and DBN) are connected together so that all phonemes are equally likely to follow all others in order to evaluate these models as substitutes for standard acoustic models. Accuracy is measured at the frame level.

The results in table 4 indicate that the DBN-A model reduces error relative to the HMM baseline by 13.5% for the dysarthric speakers, and 8.2% for the non-dysarthric speakers, which is significant at the 95% confidence level. Moreover, each DBN model performs better than all discriminative models for the dysarthric group, despite the DBN being a generative model. This result shows a clear benefit of incorporating articulatory knowledge during the training process.

Model	Dysarthric	Non-dysarthric
HMM	34.3	73.1
NN-MLP	41.9	74.4
NN-ELM	42.2	75.1
SVM-RBF	42.3	74.7
SVM-DTW	42.0	74.5
DBN-F (default)	42.6	74.8
DBN-F (sparse)	42.4	74.3
DBN-A	43.2	75.0
DBN-A2	42.6	75.0
DBN-A3	42.7	75.3

Table 4: Phone classification accuracies (%) at the frame level for models of speakers with and without dysarthria.

5 Discussion

Since speakers with dysarthria are relatively sparse in the population, the standard approach of merely collecting more data may not be applicable. In this paper we've proposed that buttressing acoustic data with knowledge of the vocal tract can offset the need for more data. Section 3 showed that articulatory knowledge can theoretically be used to reduce the statistical disorder in the acoustics from speakers with dysarthria and section 4 confirmed this empirically.

However, there are a number of supersegmental aspects of dysarthric speech that cannot be encapsulated by the frame-based approach taken here and in ASR research generally. For instance, sonorants uttered by speakers with dysarthria in TORGO are significantly slower than their control counterparts at the 95% confidence interval for */eh/* and at the 99.5% confidence interval for all other phonemes. Furthermore, our dysarthric data contains considerable involuntary velopharyngeal noise (often accompanied by difficult respiration), audible swallowing, and stuttering. In clinical circles, some of these can be partially explained by disruptions of the pathways of verbal motor commands that include sensory feedback [22], including exteroceptive stimuli (auditory and tactile), and interoceptive stimuli (particularly proprioception). Barlow argues 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 [23]. Since dysarthria represents the distortion of the *execution* of those plans, a model of ASR that explicitly encodes the long-term abstract dynamics of speech [24] may be the most fertile ground for future research.

Acknowledgments

This research is funded by NSERC and the University of Toronto.

References

[1] John-Paul Hosom, Alexander B. Kain, Taniya Mishra, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. Intelligibility of modifications to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 924–927, April 2003.

[2] Karen Hux, Joan Rankin-Erickson, Nancy Manasse, and Elizabeth Lauritzen. Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication (AAC)*, 16(3):186 –196, January 2000.

[3] Rupal Patel. Control of prosodic parameters by an individual with severe dysarthria. Technical report, University of Toronto, December 1998.

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

[5] Joe Frankel, Mirjam Wester, and Simon King. Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language*, 21:620–640, 2007.

[6] Mirjam Wester. Syllable classification using articulatory - acoustic features. In *Proceedings of Eurospeech 2003*, pages 233–236, Geneva, Switzerland, 2003.

[7] Mirjam Wester, Joe Frankel, and Simon King. Asynchronous articulatory feature recognition using dynamic Bayesian networks. In *Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop*, volume 104, pages 37–42, Kyoto, Japan, 2004.

[8] Yana Yunusova, Jordan R. Green, and Antje Mefferd. Accuracy Assessment for AG500, Electromagnetic Articulograph. *Journal of Speech, Language, and Hearing Research*, 52:547–555, April 2009.

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

[10] Victor Zue, Stephanie Seneff, and James Glass. Speech Database Development: TIMIT and Beyond. In *Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases (SIOA-1989)*, volume 2, pages 35–40, Noordwijkerhout, The Netherlands, 1989.

[11] Ray D. Kent, Gary Weismer, Jane F. Kent, and John C. Rosenbek. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499, 1989.

[12] Marco F. Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *Proceedings of the 2008 IEEE International Conference on In Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188, Seoul, South Korea, 2008.

[13] Ray D. Kent and Kristin Rosen. Motor control perspectives on motor speech disorders. In Ben Maassen, Raymond Kent, Herman Peters, Pascal Van Lieshout, and Wouter Hulstijn, editors, *Speech Motor Control in Normal and Disordered Speech*, chapter 12, pages 285–311. Oxford University Press, Oxford, 2004.

[14] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: the art of scientific computing*. Cambridge University Press, second edition, 1992.

[15] Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, September 2007.

[16] Kevin Patrick Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California at Berkeley, 2002.

[17] Zoubin Ghahramani. Learning dynamic Bayesian networks. In *Adaptive Processing of Sequences and Data Structures*, pages 168–197. Springer-Verlag, 1998.

[18] Gowtham Jayaram and Kadry Abdelhamied. Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development*, 32(2):162–169, 1995.

[19] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[20] Vincent Wan and James Carmichael. Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2005)*, September 2005.

[21] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. In *Proceedings of Neural Information Processing Systems 2003*, 2003.

[22] Vincent L. Gracco. Central and peripheral components in the control of speech movements. In Fredericka Bell-Berti and Lawrence J. Raphael, editors, *Introducing Speech: Contempory Issues, for Katherine Safford Harris*, chapter 12, pages 417–431. American Institute of Physics press, 1995.

[23] H.B. Barlow. Unsupervised learning. *Neural Computation*, 1(3):295–311, 1989.

[24] Frank Rudzicz. Correcting errors in speech recognition with articulatory dynamics. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala Sweden, July 12-14 2010.