APPLYING DISCRETIZED ARTICULATORY KNOWLEDGE TO DYSARTHRIC SPEECH

Frank Rudzicz

University of Toronto
Department of Computer Science
Toronto, Ontario, Canada M5S 3G4

ABSTRACT

This paper applies two dynamic Bayes networks that include theoretical and measured kinematic features of the vocal tract, respectively, to the task of labeling phoneme sequences in unsegmented dysarthric speech. Speaker dependent and adaptive versions of these models are compared against two acoustic-only baselines, namely a hidden Markov model and a latent dynamic conditional random field. Both theoretical and kinematic models of the vocal tract perform admirably on speaker-dependent speech, and we show that the statistics of the latter are not necessarily transferable between speakers during adaptation.

***Index Terms*—** Accessibility, dynamic Bayes nets, articulatory information, conditional random fields

1. INTRODUCTION

Dysarthria is a set of speech disorders affecting the physical production of speech but not the abstract understanding of language, as in aphasia or autism. In dysarthria, congenital or traumatic damage to the neuromotor cranial nerves restricts the motion of the speech articulators (e.g., tongue, lips), resulting in smaller vowel spaces, more atypical consonants, and generally unintelligible speech. Despite these difficulties, dysarthric speakers tend to prefer spoken expression over other physical modes for its relative naturalness and speed [1]. Since dysarthria is characterized by differences in physical production, our goal is to determine whether abstract and measured representations of dysarthric articulation are useful in speech recognition for this population.

In this paper we compare two models for labeling acoustic phoneme sequences in dysarthric speech that incorporate articulatory knowledge. The first is trained with a joint model of the phonological features of speech, and the second using measured articulatory motion of the tongue, lips and jaw. The relative performance of these methods should suggest whether measured data are necessarily preferable, or if theoretical knowledge of the production mechanism can sufficiently improve rates of recognition.

1.1. Production Knowledge

Despite their popular use, monophones and triphone segments can be decomposed into more fundamental units. Namely, phonological features (PFs) are quantized abstractions of several articulatory features of speech such as the sagittal position of the tongue, or vocalization. Because PFs can change asynchronously across phonetic boundaries and are more fine-grained than phonemic representations, their use has been shown to partially account for coarticulation effects and speaker variability [2], which are particularly exacerbated in dysarthric speech. Phonological features are language independent, reliably recoverable from acoustics among regular speakers, and robust to environmental noise [3, 4].

A more empirical approach to vocal tract knowledge is derived from actual measurement of the vocal tract during speech with semi-invasive procedures such as electromagnetic articulography (EMA), magnetic resonance imaging (MRI), X-ray microbeam analysis, ultrasound, or electropalatography. For EMA, inducing a small current into small receiver coils glued to the tongue, lips, and other articulators, these positions can be accurately inferred relative to fixed transmitters around the speaker's head that produce alternating magnetic fields. These systems produce no audible noise, and the coils interfere surprisingly little with regular speech.

1.2. Dynamic Bayes networks

Dynamic Bayes networks (DBNs) are directed acyclic graphs that generalize the powerful stochastic mechanisms of Bayesian learning to time sequences. Given an observation $Z_{1:T}^{(1:N)}$ of arbitrary length T, its likelihood is computed by 'unrolling' a 2-frame DBN to T frames, and multiplying all posteriors,

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

Here, conditional distributions, $B_\rightarrow$ are drawn over adjacent frames in time for the i^{th} variable at time t, $Z_t^{(i)}$, by $P(Z_t|Z_{t-1}) = \prod_{i=1}^{N} P(Z_t^{(i)}|Pa(Z_t^{(i)}))$, given the parents of z, $Pa(z)$. This temporal model generalizes both the hidden Markov model (HMM) and the Kalman filter [5]. 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 (EM) given hidden variables, enabling state-based methods [6].

2. EXPERIMENTS

Given mel-frequency cepstral coefficient (MFCC) observation sequences $\mathbf{o} = \{o_1, o_2, ..., o_T\}$, our task is to identify the aligned phonemic labels $\mathbf{l} = \{l_1, l_2, ..., l_T\}$ for whole, unsegmented utterances. The purpose is to study adaptation of vocal tract information to dysarthric speech, so we examine two DBN models, DBN-A and DBN-PF, that incorporate measured EMA data and phonological features, respectively.

Dysarthric data are obtained from the Nemours database [7] which includes phonetically annotated speech from 11 male speakers with either cerebral palsy or traumatic brain injury. Each speaker produces 74 nonsensical sentences consisting of words randomly selected without replacement from closed sets. All speech is sampled at 16kHz and converted to 42-dimensional MFCC feature vectors consisting of 0^{th}- to 12^{th}-order cepstral coefficients, log energy, and all δ and $\delta\delta$ variants. Additionally, Nemours provides intelligibility assessments of each speaker as determined by the Frenchay Dysarthria Assessment, which measures the speech-motor function of the articulators and speech intelligibility along a normalized 0 (no function) to 8 (normal) scale [8].

2.1. Vocal tract information

The University of Edinburgh's MOCHA database consists of 460 English sentences, almost entirely from TIMIT [9], for each of the two available speakers, who are non-dysarthric, with acoustic data temporally aligned with EMA measurements [10]. We use eight of the male speaker's articulatory parameters, namely the upper lip, lower lip, upper incisor, lower incisor, tongue tip, tongue blade, tongue dorsum, and velum. Each parameter is measured in the two dimensions of the midsaggital plane, comprising a 16-dimensional kinematic articulation vector. This space is then optionally reduced to 4 or 8 principal components by singular value decomposition specific to each phone, in which 4, 8, or 16 mean vectors are computed according to the sum-of-squares error function to be indicative of the resulting clusters of data. During training, the index **A** of the nearest of these mean vectors to the frame of EMA data at time t is applied to the DBN as an observed variable. During inference, this variable is hidden and we marginalize over all values when computing the likelihood. Figure 1a shows the chosen topology of DBN-A.

There is currently no available database of EMA recordings for dysarthric subjects, so we adapt DBN-A to dysarthric acoustics by making the index **A** hidden after training on MOCHA, and retraining on Nemours data in the same acoustic feature space. Since phonological features are derived from phone-level annotations, there is no such restriction. The PFs used in this study are based on those of Wester [11], and are listed in Table 1. The chosen topology of DBN-PF, based on that of Frankel *et al*. [4], is shown in Figure 1b.

Feature	Values (with Cardinality)
*Manner (***M***)*	approximant, fricative, nasal, retroflex, silence, stop, vowel (7)
*Place (***Pl***)*	alveolar, bilabial, dental, labiodental, silence, velar, nil (7)
*High/Low (***HL***)*	high, mid, low, silence, nil (5)
*Voice (***V***)*	voiced, unvoiced (2)
*Front/Back (***FB***)*	front, central, back, nil (4)
*Round (***R***)*	round, non-round, nil (3)
*Static (***S***)*	static, dynamic (2)

Table 1. *Phonological features and their possible values.*

2.2. Acoustic baseline

We train two speaker-dependent baseline models to the mapping between phone label sequences $\mathbf{l}$ and MFCC observations $\mathbf{o}$, each with disjoint sets of hidden states $\mathbf{Q}_l$ associated with each phone $l \in L$, and model parameters θ. The first is a 3-state hidden Markov

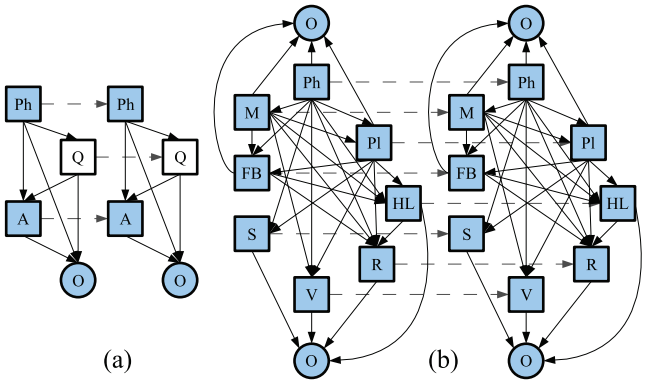

Fig. 1. *Two-frame dynamic Bayes networks with (a) EMA measurements (DBN-A) and (b) phonological features (DBN-PF). Filled and empty nodes are observed and hidden variables, respectively, and square and round nodes are discrete and continuous variables, respectively. Nodes* **Ph**, **Q**, **A** *and* **O** *represent phoneme, state, EMA and MFCC observations. All other variables are highlighted in Table 1. Inter-frame conditional links are in grey for clarity.*

model with the same topology as DBN-A in §2.1, except the articulatory variable is replaced by a hidden discrete variable, yielding a typical 16-Gaussian mixture output density through marginalization amenable to normal EM training and Viterbi decoding. These mixtures are initialized by k-means clustering with full covariance.

The second acoustic baseline is the discriminative latent-dynamic conditional random field (LDCRF) that differs from the HMM primarily in that it does not model the prior $P(\mathbf{o})$, as defined in eq. 2. This model differs from both simple CRFs and 'Hidden state' CRFs in that the LDCRF models the intrinsic sequential substructure using hidden states and assigns labels dynamically on a frame-by-frame basis, rather than to the entire sequence [12].

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

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

$$L(\theta) = \sum_{i=1}^{N} \log P(\mathbf{l}_i|\mathbf{o}_i,\theta) - \frac{1}{2\sigma^2}||\theta||^2. \tag{3}$$

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,

$$\mathbf{l}^* = \arg\max_{\mathbf{l}} \sum_{\mathbf{q}:\forall q_t \in \mathbf{Q}_{l_t}} P(\mathbf{q}|\mathbf{o},\theta^*)\,. \tag{4}$$

Data from each speaker are randomly split into 90% training data and 10% test data. The HMM and DBN-A models are trained with EM and smoothed junction-tree inference since these models contain hidden variables. When adapting these models and the LDCRF to dysarthric speech, we initialize with the distributions learned on non-dysarthric speech and train on speaker-specific acoustics. All training of the fully observed DBN-PF is with maximum likelihood, so adaptation involves concatenating the MOCHA and Nemours training data and learning from scratch. In all cases, training data includes all phones observed during testing and is applied to the 46 phones that MOCHA and Nemours have in common. We split

dysarthric data by speaker into three categories according to the level of intelligibility as determined by the Frenchay assessment [8]. 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'.

3. RESULTS AND DISCUSSION

Table 2 shows the accuracy of unsegmented phone labeling on speaker-dependent and speaker-adaptive distributions for each model, according to the severity of dysarthria. Here, DBN-A is trained to mixtures of 16 Gaussian clusters determined by unreduced (16-dimensional) articulatory data. Results are surprisingly close across all methods, although given the current setup, a few distinctions are evident. First, we see an increasing benefit of adaptive over dependent training on dysarthric speech as intelligibility increases, with absolute improvements of 2.0%, 4.3%, and 7.5% on severely, moderately, and mildly dysarthric speech, respectively. We also note that although DBN-A performs admirably for the control ('ctrl') speaker, it is less successful in adapting to dysarthric speech.

		sev	mod	mild	ctrl
HMM	Depend.	14.1	27.8	51.6	72.8
	Adapt.	16.8	32.1	58.9	-
LDCRF	Depend.	15.2	28.0	51.8	73.5
	Adapt.	16.8	32.4	59.1	-
DBN-PF	Depend.	15.0	28.0	51.6	73.3
	Adapt.	16.7	32.3	59.4	-
DBN-A	Depend.	-	-	-	73.6
	Adapt.	16.2	31.7	58.3	-

Table 2. *Average proportion of correctly labelled phones of speaker-dependent and speaker-adaptive models, according to the severity of dysarthria. Dashes represent non-applicable configurations.*

We compare the results of DBN-PF against those of Frankel *et al*. who use a similar Bayesian structure to label frames in regular speech [4]. That research obtained between 78% and 79.4% frame-level classification accuracy in terms of unanimously correct PF labels. Key differences between their work and ours is that our networks contain two additional features, namely High/Low and the phone variable, and our data sets for regular speech differ significantly. Frankel *et al*. use the OGI Numbers corpus [13] and almost 2.2 million frames of training data, while we only train with approximately 740 thousand frames of the MOCHA database. For prudence, we evaluate our DBN-PF structure on roughly half of the TIMIT database of connected speech (approximately 2.4 million frames), and achieve 77.4% phone-level accuracy, which is within our expected range of performance. Future implementations of DBN-PFs will explicitly disallow simultaneously incompatible variable assignments (e.g., *Manner*=vowel and *Voice*=unvoiced).

We consider alternative network structures that explicitly augment DBN-A with variables representing the velocity ($\mathbf{A_v}$) and acceleration ($\mathbf{A_a}$) of EMA coils. The first, DBN-A2, conditions $\mathbf{A_v}$ and $\mathbf{A_a}$ on the phone and state, and trisects the observation vector $\mathbf{O}$ into three 14-dimensional vectors (MFCC, δ, and $\delta\delta$) that are each conditioned on the phone, state, and on one appropriate kinematic variable (either $\mathbf{A}$, $\mathbf{A_v}$, or $\mathbf{A_a}$). 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 $\mathbf{O}$ on all other variables. These models are compared in Table 3 on the control speaker across the number of principal components, N_p, and the number of Gaussians, K. Although DBN-A3 appears somewhat more accurate, it is comparably very slow to train. These results generally agree with similar work that adapted acoustic-only DBNs to Japanese kinematic data [14] over 1 or 2 iterations of EM. That work showed relative error reduction of between 0.7% and 3.8% on phone classification among a selection of alternative speaker-dependent DBNs. We adapt both DBN-A2 and DBN-A3 to each severely dysarthric speaker, and observe phone-level accuracy between 15.9% and 16.2%, showing no significant improvement over DBN-A for this group.

		DBN-A	DBN-A2	DBN-A3
$N_p = 4$	K=4	57.6	56.9	57.8
	K=8	66.8	66.5	66.8
	K=16	68.9	69.1	69.3
$N_p = 8$	K=4	63.3	63.4	63.8
	K=8	71.0	71.1	71.3
	K=16	72.4	72.2	72.7
$N_p = 16$	K=4	64.7	65.1	65.2
	K=8	72.5	72.4	72.7
	K=16	73.6	73.6	74.0

Table 3. *Accuracies of three EMA-informed DBNs across varying quantities of principal components, N_p, and Gaussians, K for speaker-dependent, regular speech.*

We examine the effect of increased sample size by adapting non-dysarthric models to cross-sections of data selected uniformly at random among all dysarthric speakers in Nemours, and testing on proportionally increasing test sets. Figure 2 suggests that as the amount of dysarthric speech is increased, the LDCRF model outperforms all others, with a relative error reduction of almost 2% over HMM with 670 training utterances for adaptation. The LDCRF is the only method of the four that employs discriminative training.

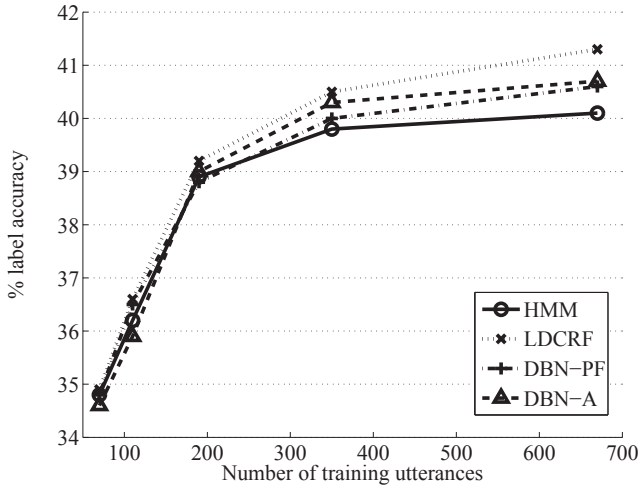

Fig. 2. *Labeling accuracy of four models as amount of dysarthric data is increased.*

Finally, we compare the generative abilities of DBN-A and DBN-PF relative to our data. We iteratively set **Ph** to each phone in the available DBN-A and DBN-PF models and marginalize over all other variables to get the distribution on $\mathbf{O}$ from which we sample synthetic acoustic data for each phone. These generated distributions are fitted with Gaussians and compared with the true

MFCC distributions of each phone with Kullback-Leibler relative divergence. The distributions generated by DBN-PF diverge from true distributions by an factor of 0.22016 on regular speech and by 0.2246 on dysarthric speech. However, while virtual DBN-A data diverges from true data by a factor of 0.1690 for regular speech, speaker-adaptive DBN-As for dysarthric speech diverge by 0.3378, on average, from true phone MFCC distributions. This disparity appears to suggest that statistical relations between acoustics and kinematics do not necessarily translate across speakers, at least on the available data, which is exemplified in Figure 3. We are currently recording our own database of articulatory kinematics in dysarthric speech with 12 individuals having either cerebral palsy or amyotrophic lateral sclerosis, along with matched controls [15]. Applying this database to our models will allow us to pursue 'speaker-dependent' and 'severity-dependent' kinematic models and will help determine to what extent the observed divergence of adaptive DBN-A is caused by the articulatory irregularities in dysarthria rather than simple inter-speaker variation.

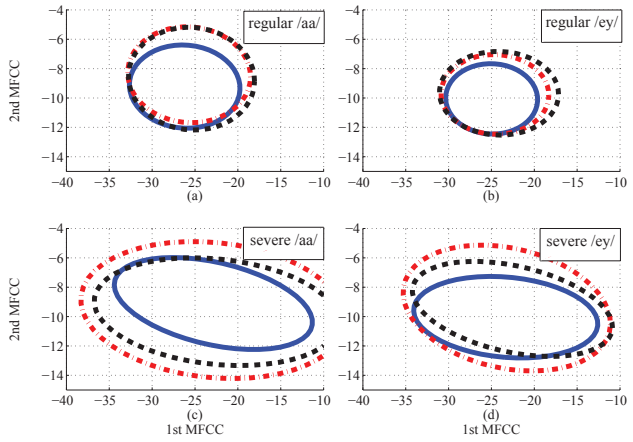

Fig. 3. *Contours representing 2 standard deviations of Gaussians fitted to real data (solid line), samples from DBN-PF (dashed line), and samples from DBN-A (dash-dotted line) on the first two mel-frequency cepstral coefficients. Subfigures represent (a) regular speech (/aa/), (b) regular speech (/ey/), (c) severely dysarthric speech (/aa/), and (d) severely dysarthric speech (/ey/).*

Although we have examined frame-based vocal tract models over whole utterances, several endogenous factors of dysarthria cannot be readily represented in any of the methods described here. For example, increased dysfluency, longer sonorants, reduced pitch control, and greater articulatory variance are typical features of dysarthric speech [15] and we are currently working to integrate some of these phenomena within the DBN framework. Further study with regards to the articulatory dynamics of dysarthria in particular, and speech generally, might in the long term result in more robust rates of recognition, and therefore an improved quality of life for those more dependent on the accuracy of assistive software.

4. ACKNOWLEDGEMENTS

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

5. 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)*, April 2003, vol. 1, pp. 924–927.

[2] Karen Livescu, Ozgur Cetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Haggerty, and Bronwyn Woods, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop," in *Proceedings of ICASSP 2007*, Honolulu, April 2007.

[3] Katrin Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, Germany, July 1999.

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

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

[6] Zoubin Ghahramani, "Learning dynamic bayesian networks," in *Adaptive Processing of Sequences and Data Structures*. 1998, pp. 168–197, Springer-Verlag.

[7] Xavier Menendez-Pidal, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzjo, and H.T. Bunnell, "The Nemours Database of Dysarthric Speech," in *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA, USA, October 1996.

[8] Pamela M. Enderby, *Frenchay Dysarthria Assessment*, College Hill Press, 1983.

[9] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development: Timit and beyond," in *Proceedings of SIOA-1989*, 1989, vol. 2, pp. 35–40.

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

[11] Mirjam Wester, "Syllable classification using articulatory - acoustic features," in *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 233–236.

[12] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2007.

[13] Ronald A. Cole, Krist Roginski, and Mark Fanty, "The OGI numbers database," Tech. Rep., Oregon Graduate Institute, 1995.

[14] Konstantin Markov, Jianwu Dang, and Satoshi Nakamura, "Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework," *Speech Communication*, vol. 48, no. 2, pp. 161–175, February 2006.

[15] Frank Rudzicz, Pascal van Lieshout, Graeme Hirst, Gerald Penn, Fraser Shein, and Talya Wolff, "Towards a comparative database of dysarthric articulation," in *Proceedings of the eighth International Seminar on Speech Production (ISSP'08)*, Strasbourg France, December 2008.