PHONOLOGICAL FEATURES IN DISCRIMINATIVE CLASSIFICATION OF DYSARTHRIC SPEECH

Frank Rudzicz

Department of Computer Science, University of Toronto
frank@ai.toronto.edu

ABSTRACT

In an attempt to overcome problems associated with articulatory limitations and generative models, this work considers the use of phonological features in discriminative models for disabled speech. Specifically, we train feed-forward and recurrent neural networks, and radial basis and sequence-kernel support vector machines to abstractions of the vocal tract, and apply these models to phone recognition on dysarthric speech. The results show relative error reduction of between 1.5% and 10.9% with this approach against standard hidden Markov modeling, and increases in accuracy with speaker intelligibility across all classifiers. This work may be applied within components of assistive software for speakers with dysarthria.

Index Terms— dysarthria, neural networks, kernel methods

1. INTRODUCTION

Dysarthria comprises a group of neuromuscular disorders that can drastically limit speech intelligibility in congenital cases such as cerebral palsy or traumatic ones such as stroke. These disorders typically limit motor function generally, making other physical interaction (e.g., keyboard) slower, and less desirable than spoken expression [1]. Unfortunately, automatic speech recognition is currently ill-suited to dysarthric speech, rendering such software inaccessible to those who might most benefit from it. We have found that traditional generative approaches such as hidden Markov models (HMMs) trained for speaker independence may achieve word-level accuracy of less than 4.5% on severely dysarthric speech against 84.8% on non-disabled speech on short sentences [2].

Disabled speech is typically characterized by a limited range of motion in the speech articulators, which results in smaller vowel spaces and more inconsistent consonants, especially in clusters [3]. As these phones assimilate with one another, generative models assign more probability to overlapped spaces, hurting performance. In this paper we consider two discriminative families for stochastic classification, neural networks (NNs) and support vector machines (SVMs), on the task of differentiating phones at the frame level for disabled speech. Since this speech is characterized by differences in physical production, our goal is to determine whether abstract representations of dysarthric articulation are easily discriminable in disordered speech, and whether these are useful in speech recognition for this population generally.

1.1. Phonological features

Phonological features (PFs)[1] are quantized abstractions of speech production along particular vocal tract configurations. For example, the *Front/Back* feature specifies the sagittal position of the tongue during vowels, and *Static* specifies the rate of acoustic change (e.g., diphthongs are dynamic). 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 [4], which are particularly exacerbated in dysarthric speech. Other useful properties of PFs include noise-robustness, language-independence, and reliable recovery from acoustics among regular speakers [5]. The features used here are based on those of Wester [6], and listed in Table 1.

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

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

2. PHONOLOGICAL-ACOUSTIC MODELS

In this paper, acoustic observation vectors are frames of speech optionally surrounded by a window of varying length. Each PF is modeled by two NNs and two SVMs for each speaker, as described below. Additionally, for each of these four discriminative techniques, we construct three triphone classifiers. The first identifies triphones solely by acoustics, the second based solely on output from the 7 PF classifiers, and the third based on a combination of these. Nonexistent triphones in the training data are modeled by their monophonic progenitors, of which there are 61.

2.1. Neural Networks

Multilayer neural networks have rarely been applied to classification within dysarthric speech, despite their popularity in general. One study, however, showed that multilayer feed-forward NNs supplied with either Fourier spectral coefficients or formant frequencies could achieve a relative error reduction (RER) of up to 40% over a commercial HMM-based system for a cerebrally palsied speaker [7].

The two types of neural network we consider here are the feed-forward multi-layer perceptron (**MLP**) and the recurrent Elman network (**ELM**), which are primarily distinguished by the latter's time-delayed replication of the hidden layer as additional contextual in-

[1] PFs are often called *articulatory features*.

put. The output of each NN consists of n nodes, where n is the cardinality of the PF being modeled, and the i^{th} node is uniquely active when training the i^{th} value of that PF. The input to each of these NNs are 42-dimensional acoustic frame vectors (see §3.1), plus optional bounding context windows of 2 or 4 frames. Three additional networks perform triphone classification given either acoustic frame vectors, the output of the 7 PF classifiers, or both as input. All networks are fully connected between layers and select the class having the highest posterior probability.

Activation functions in hidden layer units are tan-sigmoid (i.e., $\left[2/\left(1+e^{-2x}\right)\right]-1$), and linear in the output layer, given weighted sums of activations x. The size of the hidden layer in each network varies with the size of the PF being classified, and is based empirically on ratios for similar tasks in related work [8, 5], as shown in Table 2. All NN triphone classifiers contain 500 hidden units.

Mann.	*Place*	*Hi/Low*	*Voi.*	*Fr/Back*	*Ro.*	*Stat.*
300	200	100	100	200	100	100

Table 2. *Number of hidden units per NN, given target feature.*

A problematic consequence of using tan-sigmoid activations, especially given large input vectors, is that the gradient can have very small magnitude, which can slow training. Instead, 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 gradient descent on our data, while only requiring a modest increase in memory.

2.2. Support Vector Machines

Support vector machines are general maximum margin classifiers that are of increasing interest in speech recognition due to their robustness to both sparse data [9] and rapid transient changes in acoustic sequences [10]. SVMs explicitly minimize the empirical classification error by orienting a hyperplanar decision boundary between classes such that its orthogonal vector represents a maximized margin between the nearest data. In these experiments 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 [11].

We consider two SVM kernels that differ in the form of their input. The first is a *radial basis function* (**RBF**), that generalizes to non-linear decision boundaries using the following kernel:

$$K_{RBF}\left(\mathbf{x},\mathbf{y}\right) = \exp\left(-\frac{(\mathbf{x}-\mathbf{y})^2}{2\sigma^2}\right), \qquad (1)$$

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

The second kernel, K_{DTW}, is a *sequence kernel* proposed by Wan and Carmichael [9] that explicitly operates on time-series, and can be generalized to arbitrary sequences $\mathbf{u}$ and $\mathbf{v}$ having non-equal lengths, although sequence lengths are forced to be equal here. 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 $(d+1)^{th}$ dimension. Namely, every vector u_i is converted to the unit vector $\hat{u_i}$ sharing an origin with H by

$$\hat{u_i} = \frac{1}{\sqrt{u_i^2+\alpha^2}}\begin{bmatrix} u_i \\ \alpha \end{bmatrix}. \qquad (2)$$

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

$$d_s(\hat{u_i},\hat{v_j}) = \theta_{\hat{u_i},\hat{v_j}} = \arccos(\hat{u_i},\hat{v_j}). \qquad (3)$$

Now, given these local distances, we apply the *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

$$D_{global}(\mathbf{u},\mathbf{v}) = \min_{\Gamma}\frac{1}{||\Gamma||}\sum_{p=1}^{||\Gamma||} d_s(\hat{u_p},\hat{v_p}). \qquad (4)$$

This distance is then converted to the kernel

$$K_{DTW}\left(\mathbf{u},\mathbf{v}\right) = \cos D_{global}(\mathbf{u},\mathbf{v}), \qquad (5)$$

which is symmetric if the symmetric version of DTW is used, which is a requirement for use in SVM classification. The kernel must also either satisfy Mercer's condition or be shown to be a valid dot product. 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 [9].

2.3. Hidden Markov Model

The baseline triphone classifier consists of standard tri-state left-right hidden Markov models (HMMs) with continuous 16-Gaussian mixture output densities decoded with the Viterbi algorithm and conditioned on the triphone label. Each model is speaker-dependent and trained with the iterative Baum-Welch algorithm [2].

3. EXPERIMENTS

The following subsections describe the materials, procedures, and results of experiments related to the performance of discriminative models on dysarthric speech.

3.1. Material

The Nemours database [12] provides phonetically annotated speech from 11 dysarthric male speakers with either cerebral palsy or traumatic brain injury, and a non-dysarthric male control. Each speaker produces 74 nonsensical sentences consisting of words randomly selected without replacement from closed sets. All speech is sampled at 16 kHz where half-overlapping 16 ms Hamming windows are 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. Target phonetic features are derived from TIMIT-phoneset annotations. Additionally, Nemours provides intelligibility assessments of each speaker as determined by the standardized Frenchay Dysarthria Assessment [13]. This assessment is administered by speech pathologists and measures the motor function of the articulators (e.g., tongue, lips) and speech intelligibility along a normalized 0 (no function) to 8 (normal) scale.

Since dysarthric speakers are both relatively rare and susceptible to fatigue, collecting data from this population can be particularly challenging. Most studies will typically include no more than 3 or 4 dysarthric speakers [14], often producing only about 25 utterances each [7]. Although Nemours is a relatively large database given its type, we apply K-fold cross-validation to 10 random permutations of

90% training and 10% test data for each speaker to partially account for sparsity. Each training set consists of ∼ 93K frames, on average, which is within the range used in studies of phonological features for regular speakers [8].

3.2. Results and Discussion

Frame-level accuracies for each PF averaged over all dysarthric speakers is summarized in Table 3 for each classifier. Interestingly, while the NN methods predictably become more accurate as PF cardinality decreases, the SVM methods are exceptionally proficient at classifying *Manner* and *Place*, which are highly related, and poor at classifying the *Round* PF despite its low cardinality. This suggests that there is some other aspect of those PFs that affects discriminability, at least for SVMs. The *nil* class is the most poorly recognized in three of the four PFs having it.

In general, SVM methods outperform NN on average by 4.9% to 9.3% absolutely and provide a 19.8% relative error reduction on dysarthric speech. On our control subject, PF models achieved 74.3% accuracy for **MLP**, and 77.6% for **RBF**, on average. Other research on speaker-independent recurrent neural networks for PF recognition on regular speech report frame-level accuracies between 77.2% and 89.1% given ∼2.2 million frames of the OGI Numbers database [15].

	Accuracy (%)					
Feature	MLP	ELM	RBF	DTW	Avg.	Card.
Manner	22.1	30.2	**66.8**	65.4	46.1	7
Place	35.5	41.9	**58.3**	56.5	48.1	7
Hi/Low	53.0	**58.7**	55.7	55.9	55.8	4
Voice	78.7	**81.3**	76.8	78.1	78.7	2
Front/Back	48.2	52.1	**55.1**	**55.7**	52.8	4
Round	68.9	**69.7**	55.3	54.0	62.0	3
Static	64.2	66.5	67.3	**69.2**	66.8	2

Table 3. *Classifier accuracy on PFs averaged over all speakers (best of row in bold), with overall accuracy and cardinality.*

Figure 1 shows the overall accuracy of each classification technique according to speaker intelligibility as determined by the Frenchay Dysarthria Assessment (§3.1). These results show a general preference for SVM 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, according to their assessments, which seems to account for an especially low accuracy with *High/Low* and *Front/Back* PFs for these speakers. Within these PFs, follow-up analysis revealed linear correlation coefficients up to 0.94 between increased formant deviation and decreased tongue function. While overall intelligibility may be useful in predicting general trends in Figure 1, it is an aggregate measure of the functions of component articulators [13], and may be superseded for speakers having more localized disabilities.

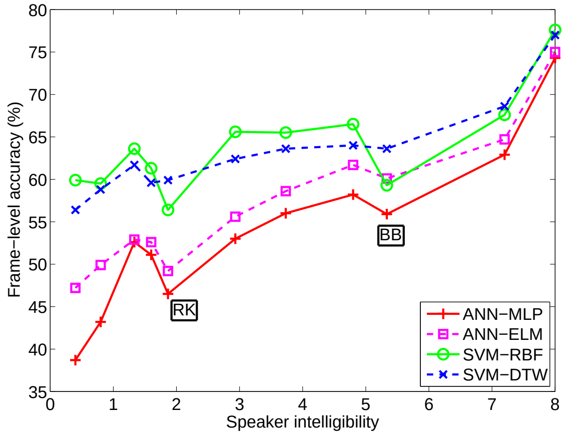

Fig. 1. *Average classifier accuracy against assessed intelligibility level.*

Finally, we consider whether PFs are useful in identifying phones. For each of our four modeling techniques, we construct three triphone classifiers as described in §2. These are then applied over whole utterances, and the selected triphones are converted into their monophonic representations. All hidden Markov models are connected to all others by means of transition probabilities learned from maximum-likelihood bigrams. Table 4 shows that merely replacing an HMM with an SVM model on acoustic data reduces frame error relatively by 6.9% to 8.8%, while including PFs gives between 1.5% and 10.9% relative error reduction over all methods, on average. However, since the seven PFs are rarely unanimously correct, they alone cannot be used to infer the respective phone in practice. Further research should investigate whether it would be useful to restrict analysis to some subset of the seven PFs used in this study.

	Accuracy (%)				
Input	MLP	ELM	RBF	DTW	HMM
MFCC	31.9	36.7	38.4	39.6	33.8
PF	5.8	11.7	16.2	17.9	-
MFCC ∪ PF	34.8	40.2	38.7	41.0	-

Table 4. *Average phone classification accuracy.*

3.3. Ongoing Work

The target PFs in this work are derived from phonetic annotations, as is generally the case in the literature [5], which in some sense does not take advantage of the suprasegmental and asynchronous properties of articulation. We are currently assembling a database of dysarthric speech focused on cerebral palsy. This data will consist of more dysarthric speakers, each producing more speech than is currently publicly available. The stimuli for this database includes meaningful phrases, and more syntactic variability in order to explore other types of constraints on the classification process. These constraints may include hybrid word or sentence models [16, 17] that use language modeling, parsing, and other features of the vocal tract, but the eventual goal is to move away from quantized models such as PFs. Also, since hybrid NN/HMM models have improved word recognition rates on acoustic-only regular speech [16, 17], ongoing work involves embedding the phonologically informed methods here into hybrid HMM models for disabled speech, and to examine the effects of disablement on those systems.

4. ACKNOWLEDGEMENTS

This work is funded by a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada, Bell University Labs, and the University of Toronto.

5. REFERENCES

[1] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, April 2003, pp. 924–927.

[2] F. 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.

[3] N. Thubthong, P. Kayasith, S. Manochiopinig, W. Leelasiriwong, and O. Rukkharangsarit, "Articulation analysis of Thai cerebral palsy children with dysarthric speech," in *Proceedings of the 6th Symposium on Natural Language Processing*, 2005.

[4] 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 *Proceedings of ICASSP 2007*, Honolulu, April 2007.

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

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

[7] G. Jayaram and K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks," *Journal of rehabilitation research and development*, vol. 32, no. 2, pp. 162–169, 1995.

[8] O. Scharenborg, V. Wan, and R. K. Moore, "Towards capturing fine phonetic variation in speech using articulatory features," *Speech Communication*, vol. 49, no. 10-11, pp. 811–826, October-November 2007.

[9] V. Wan and J. Carmichael, "Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data," in *Proceedings of INTERSPEECH 2005*, September 2005.

[10] P. Niyogi and C. Burges, "Detecting and interpreting acoustic features with support vector machines," University of Chicago, Tech. Rep. TR-2002-02, 2002.

[11] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," in *Proceedings of Neural Information Processing Systems 2003*, 2003.

[12] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzjo, and H. Bunnell, "The Nemours Database of Dysarthric Speech," in *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA, USA, October 1996.

[13] P. M. Enderby, *Frenchay Dysarthria Assessment*. College Hill Press, 1983.

[14] 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 *Proceedings of ICASSP 2006*, vol. 3, May 2006, pp. 1060–1063.

[15] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Germany, July 1999.

[16] 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," *Medical Engineering and Physics*, vol. 28, no. 8, pp. 741–748, October 2006.

[17] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10-11, pp. 827–835, 2007.