Comparing Speaker-Dependent and Speaker-Adaptive Acoustic Models for Recognizing Dysarthric Speech

Frank Rudzicz
University of Toronto, Dept. of Computer Science
10 King's College Road
Toronto, Ontario, Canada
frank@cs.toronto.edu

ABSTRACT

Acoustic modeling of dysarthric speech is complicated by its increased intra- and inter-speaker variability. The accuracy of speaker-dependent and speaker-adaptive models are compared for this task, with the latter prevailing across varying levels of speaker intelligibility.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: Acoustic models; H.4.3 [**Communications Applications**]: Speech Recognition

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Dysarthria is a set of motor disorders resulting from general physical disabilities that limit speech intelligibility. Despite these difficulties, dysarthric speakers tend to prefer spoken expression over other physical modes to increase naturalness and speed [2]. Unfortunately, current automatic speech recognition (ASR) systems for the general public are not well-suited to dysarthric speech, rendering such software inaccessible to those who would benefit from it the most [1].

We compare two approaches to acoustic modeling to improve ASR accuracy on dysarthric speech. Speaker-dependent (SD) models are trained solely to an individual, whereas speaker-adaptive (SA) models are initialized by models trained on a larger population but are later adjusted to a single user. SD models tend to be more accurate than SA models as user-specific training increases, but are initially less accurate since they are initialized by real speech [3].

Raghavendra, Rosengren, and Hunnicutt [7] compared what they described as an SA phoneme recognizer and an SD word recognizer on dysarthric speech. They concluded that SA modeling is appropriate for mild or moderate dysarthria, with an empirical relative error reduction (RER) of 22%, but that severely dysarthric speakers are better served by speaker dependence, with 47% RER. Noyes and Frankish [6] reported SD models attaining between 75% and 99% word-level accuracy for impaired speakers on a small vocabulary, where human listeners could only correctly identify between 7% and 61%. Sawhney and Wheeler [8] found pronounced gains from SD models, with an RER of $\sim$ 22% over independent models using an unspecified segmental phoneme recognizer. These experiments, however, used no more than 5 test subjects each, with limited training data.

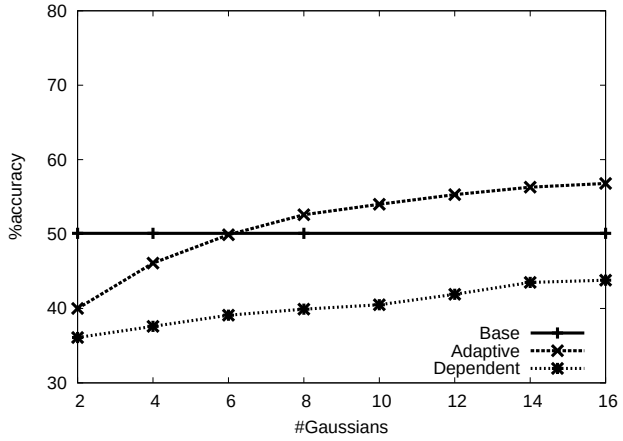
ASSETS'07, October 15–17, 2007, Tempe, Arizona, USA..
ACM 978-1-59593-573-1/07/0010.

2. EXPERIMENTAL SETUP

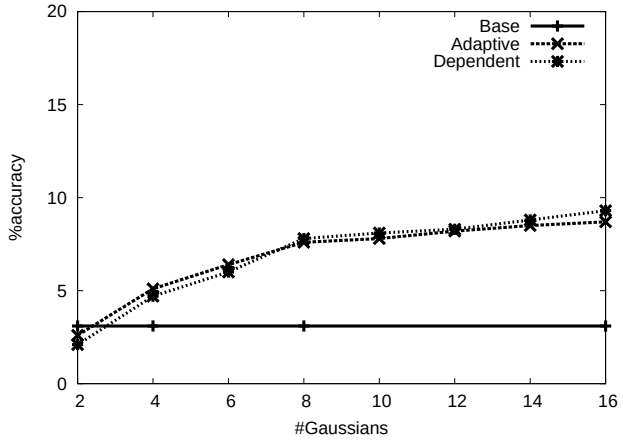
The following experiment is designed to test the accuracy of ASR as one increases the precision of acoustic models and the amount of training data beyond the examples cited above. The Nemours database provides annotated speech data from 11 male speakers with varying degrees of dysarthria, each producing 74 nonsense sentences of the form *The* N_0 *is* V*ing the* N_1 where N_0 and N_1 are monosyllabic nouns and V is a monosyllabic verb [5]. These target words were randomly selected without replacement in order to provide closed-set phonetic contrasts (e.g., place, manner, voicing). Additionally, one non-dysarthric control speaker repeated each sentence in the database.

We categorize each speaker according to his recognition rate on Nemours data using a baseline acoustic model trained on spoken transcripts of the Wall Street Journal (WSJ) [4]. The four speakers having word-level recognition rates below 10% with the baseline model are grouped as 'severe', the four with rates between 11% and 30% are grouped as 'moderate', and the three between 31% and 60% are grouped as 'mild'. The control speaker had a word-level recognition rate of 84.8%. In broad terms, these initial results show a pattern of distribution similar to that of subjective sentence-level intelligibility scores among human listeners.

Both the dependent and adaptive models for each speaker are triphone left-right Hidden Markov Models (HMMs) with Gaussian mixture output densities decoded with the Viterbi algorithm on a lexical-tree structure augmented with a context-free grammar. For each speaker, we initialize the HMM acoustic parameters of the dependent model randomly, and initialize the adaptive model with the common WSJ-trained baseline. We independently vary the number of Gaussians and the amount of training utterances in order to measure how precision and data coverage accommodate the variability of dysarthric speech, and apply the iterative Baum-Welch training algorithm on both models for each speaker. Word-level accuracy is measured using our automated system on test data.



Mildly dysarthric speakers



Severely dysarthric speakers

Figure 1: ASR accuracy measured against acoustic model precision (i.e., number of Gaussians). Baselines represent models trained on the WSJ corpus.

3. RESULTS AND DISCUSSION

Increasing the amount of training data from 20 to 132 training sentences per speaker does not show any definite improvement, with accuracy fluctuating around 3% from the mean across trials. The fact that accuracy does not increase suggests that there is not enough data in Nemours to represent intra-speaker variation, and that studies using fewer test subjects may also require more data.

Figure 1 shows accuracy increasing monotonically with the number of Gaussians for the mildly and severely dysarthric speakers. In all cases but the most severe, the adaptive models outperform their dependent counterparts and reduce relative error by up to 23.1% in the mild group, by 4.9% in the moderate group, and by 30.7% for the non-dysarthric speaker. This suggests that taking advantage of pre-existing models of the normal population may best suit dysarthric speakers with higher intelligibility. This tends to support the abstract conclusions of Raghavendra et al. [7], except that they also observed a clear superiority of dependency for severely dysarthric speakers. By contrast, we only observe slight SD gains over the baseline as the number of Gaussians increases, possibly due to the distribution of data.

In real-world applications, pre-training is sometimes not possible. Phonemic substitutions are the most common phenomena observed across all speakers in the baseline model, especially */ng/* for */n/* (125), */t/* for */uw/* (87), and */ey/* for */ih/* (84). Deletions were also common with this model, especially */b/* (118), */s/* (111), */w/* (60), */f/* (55) and */l/* (48). These observations suggest that ASR software might be made more accessible to dysarthric speakers by increasing robustness against consonant variations in general.

4. ONGOING WORK

Our current work involves acquiring more varied types of data upon which to perform machine learning. In particular, we will be acquiring time-aligned acoustic and electromagnetic midsagittal articulographic data with dysarthric and control speakers on more linguistically interesting texts amenable to n-gram language modeling. Future experiments will include alternatives to Gaussian mixtures in modeling HMM parameters, and more discriminative classification mechanisms such as recurrent neural networks.

5. ACKNOWLEDGMENTS

This work is funded by an NSERC Canada Graduate Scholarship and the University of Toronto.

6. REFERENCES

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

[2] 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)*, volume 1, pages 924–927, April 2003.

[3] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, April 2001.

[4] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf. The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, Apr 2003.

[5] 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.

[6] J. M. Noyes and C. R. Frankish. Speech recognition technology for individuals with disabilities. *Augmentative and Alternative Communication (AAC)*, 8(4):297–303, 1992.

[7] 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. *Augmentative and Alternative Communication (AAC)*, 17(4):265–275, December 2001.

[8] N. Sawhney and S. Wheeler. Using phonological context for improved recognition of dysarthric speech. Technical Report 6345, MIT Media Lab, 1999.