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

Frank Rudzicz
University of Toronto, Department of Computer Science

NSERC CRSNG

UNIVERSITY of TORONTO

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

1. Introduction

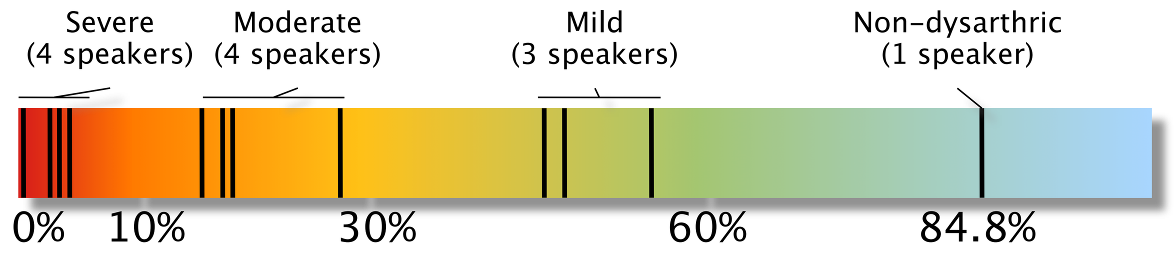
- **Dysarthria** is a set of neuromuscular motor disorders that limit speech **intelligibility**.
- Dysarthric speakers often prefer spoken expression over other physical means to increase naturalness and speed.
- Automatic speech recognition (**ASR**) is essentially inaccessible for individuals with dysarthria.
- We compare the following types of acoustic model:
  - **Speaker-dependent** (**SD**): Trained solely to an individual.
  - **Speaker-adaptive** (**SA**): Initialized by models trained on a larger population, later adjusted to a single user.
- SD models tend to become more accurate as user-specific training increases, but are initially less accurate than SA models.

2. Previous Work

- Raghavendra et al. [4] compared a SA phoneme- and a SD word-recognizer on dysarthric speech.
  - They concluded that SA is appropriate for **mild** or **moderate** dysarthria, with empirical relative error reduction (**RER**) of **22%**.
  - **Severely** dysarthric speakers are better served by SD, with **47%** RER.
- Noyes and Frankish [3] report SD models attaining between **75%** and **99%** word accuracy for impaired speakers on a small vocabulary.
  - Humans are accurate between **7%** and **61%** of the time.
- Sawhney and Wheeler [5] found pronounced gains from SD models, with an RER of ~**22%** over independent models using unsupervised segmental phoneme recognizer.
- Most work suffers from using too few (≤ 5) speakers for training.

3. Data

- We use the annotated Nemours database [1].
- This contains 11 dysarthric male speakers, each producing 74 nonsense sentences of the form The ($\mathbf{N_0}$) is (**V**)ing the ($\mathbf{N_1}$).
- Target words were randomly selected without replacement to provide closed-set phonetic contrasts (e.g., place, manner, voicing).
- One non-dysarthric speaker repeated each sentence in the database.
- Speakers are **grouped** according to recognition rate with baseline acoustic models trained on spoken Wall Street Journal (WSJ) transcripts [2].



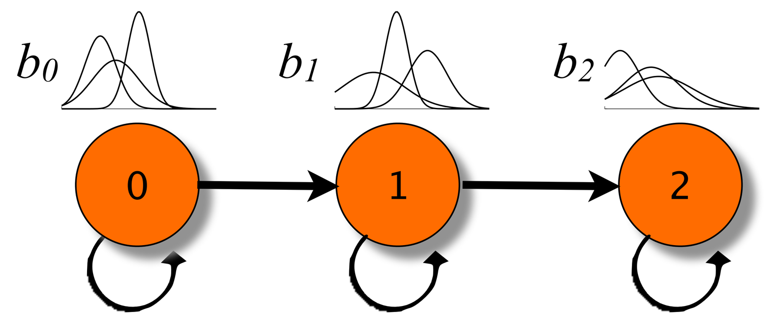
- Subjective sentence-level human intelligibility scores are similarly distributed.

4. Model and Training Mechanism

- Both the SD and SA models are continuous 3-state triphone Hidden Markov Models (**HMMs**) decoded by the Viterbi algorithm.
- Emission probabilities b_i are Gaussian mixture models (**GMMs**), with K Gaussians N_k.

$$b_i(x) = \sum_{k=1}^{K} c_k N_k(x; \mu_k, \Sigma_k)$$

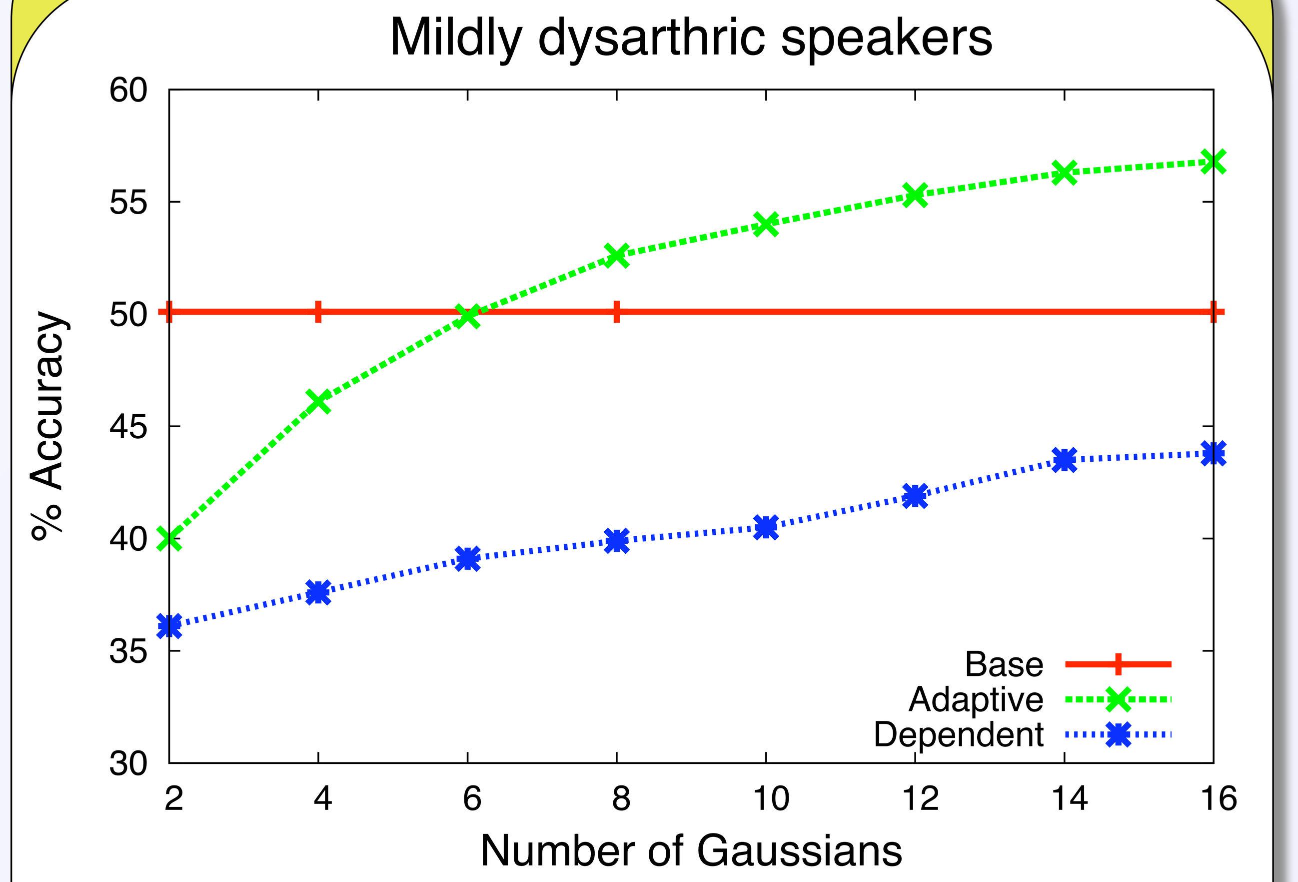
- Language model contains lexical tree structures augmented with a context-free grammar.

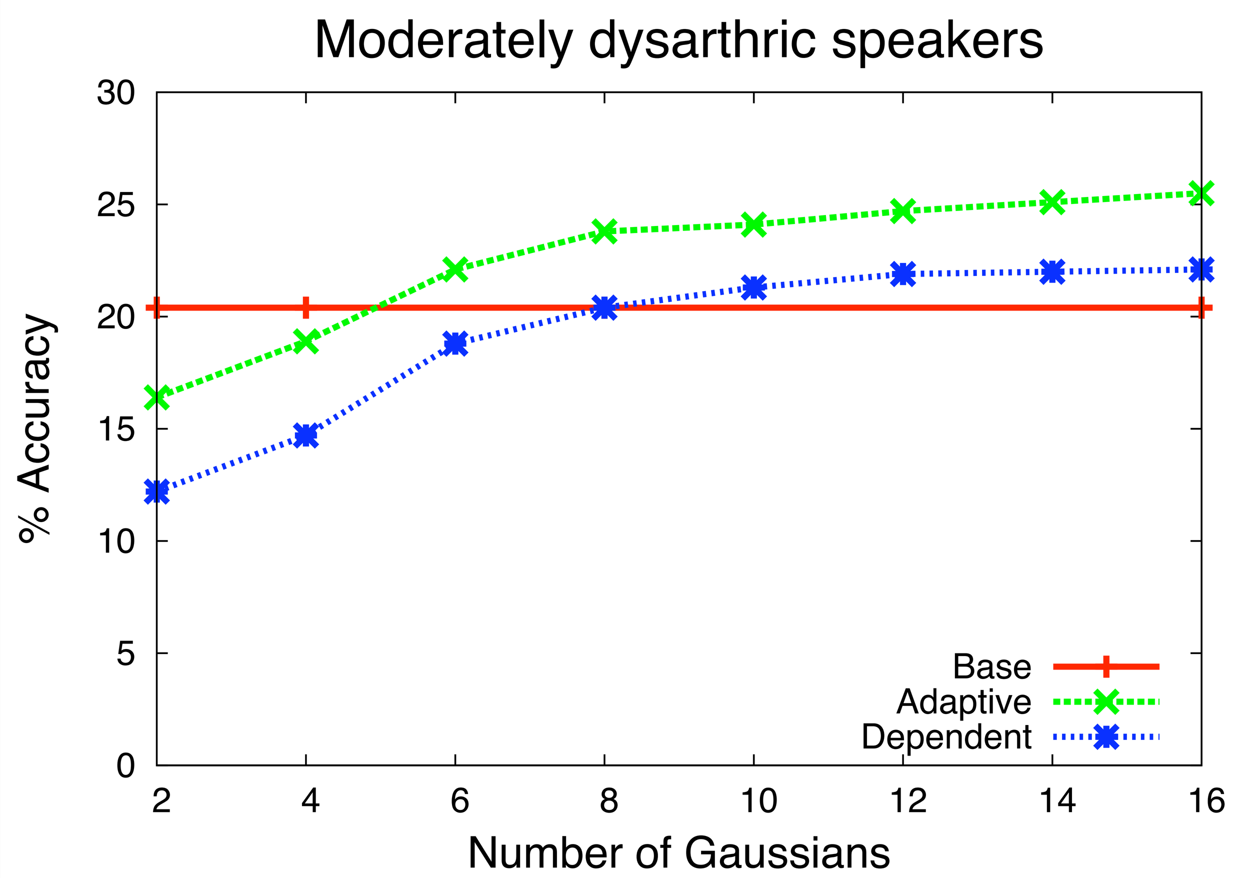


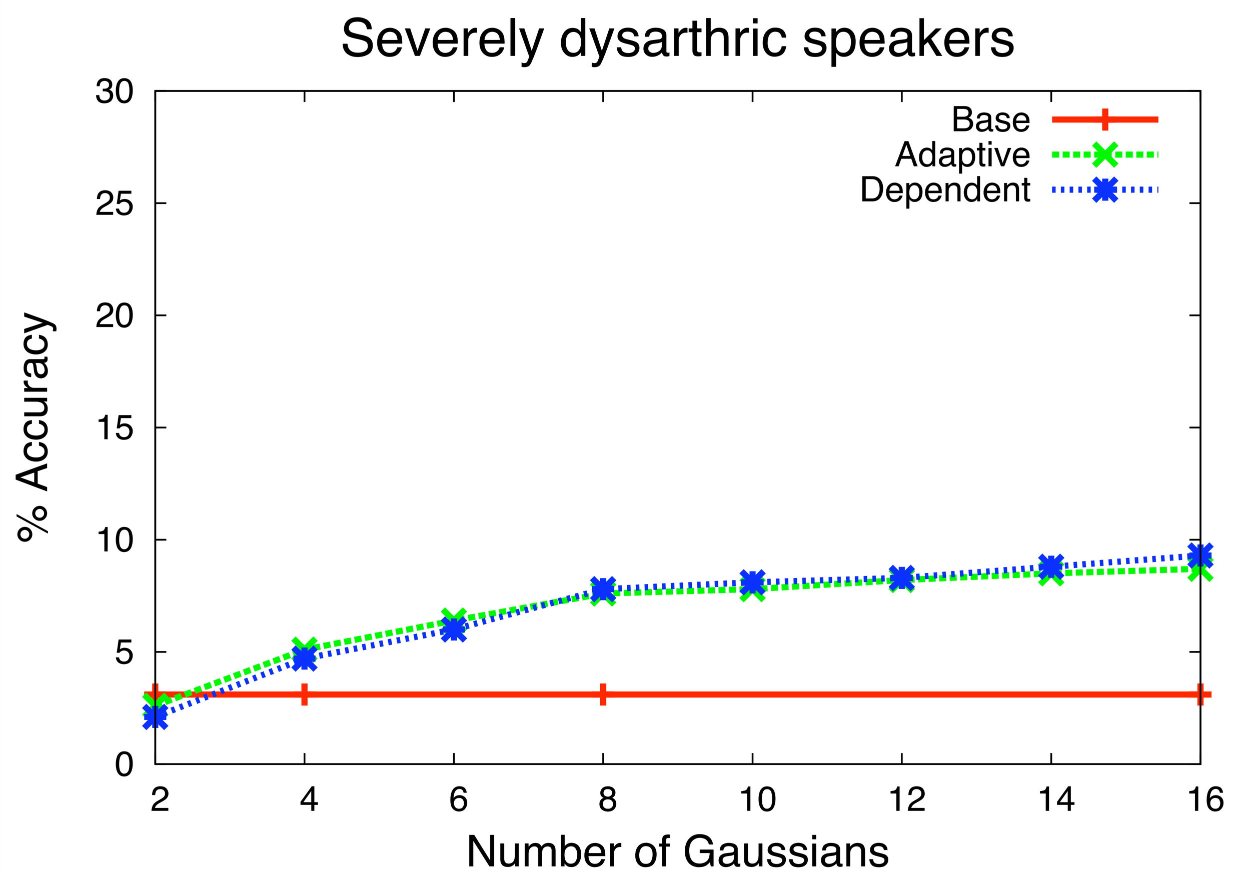
3-state triphone (e.g. /ae-s+eh/)

- **Baseline**: Use WSJ corpus, don't train.
- **Dependent Training**: Initialize b_i randomly.
- **Adaptive Training**: Initialize with WSJ corpus.
- For training, we independently vary the number of Gaussians in b_i, and apply the iterative Baum-Welch training algorithm on each speaker.
- **Word-level accuracy** is measured using our automated system on test data.

5. Results







- Increasing the amount of training data from 20 to 132 training sentences per speaker does not show any definite improvement (accuracy fluctuates around ±3% from mean).

6. Discussion

- Pre-existing models from the non-dysarthric population may best suit dysarthric speakers with higher intelligibility.
- Our results support Raghavendra et al. [4], except we do not observe a clear superiority of SD models for severely dysarthric speakers.
  - In contrast, we measure only slight SD gains as the number of Gaussians increases.
- **Phonemic substitution** is the most common phenomenon across all speakers, especially /ng/ → /n/ (**125**), /t/ → /uw/ (**87**), /ey/ → /ih/ (**84**)
- **Deletions** mostly involve dropped consonants /b/ (**118**), /s/ (**111**), /w/ (**60**), /f/ (**55**), /l/ (**48**)
- There is not enough data to represent intra-speaker variation. What are the alternatives?

7. Current Work

- We are designing a **generic classifier framework** that includes neural networks and support vectors.
  - Experiments will explore alternatives to GMM emission probabilities (e.g., Bayes nets).
- **Data collection** combines acoustics and kinetics using electromagnetic midsagittal articulography.
  - This will incorporate physical models into ASR and contain more linguistically varied texts amenable to syntactic and semantic language modeling.
- **Future work** includes development of a general dictation system accessible to dysarthric speakers.

References

[1] 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, Oct. 1996.

[2] 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.

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

[4] 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, 17(4):265-275, December 2001.

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

 ASSETS'07, October 15-17, 2007, Tempe, Arizona, USA.

Contact email: frank@cs.toronto.edu