ADAPTING ACOUSTIC AND LEXICAL MODELS TO DYSARTHRIC SPEECH

Kinfe Tadesse Mengistu and Frank Rudzicz

University of Toronto
Department of Computer Science
{kinfe, frank}@cs.toronto.edu

ABSTRACT

Dysarthria is a motor speech disorder resulting from neurological damage to the part of the brain that controls the physical production of speech and is, in part, characterized by pronunciation errors that include deletions, substitutions, insertions, and distortions of phonemes. These errors follow consistent intra-speaker patterns that we exploit through acoustic and lexical model adaptation to improve automatic speech recognition (ASR) on dysarthric speech. We show that acoustic model adaptation yields an average relative word error rate (WER) reduction of 36.99% and that pronunciation lexicon adaptation (PLA) further reduces the WER by an average of 8.29% relative on a large vocabulary task of over 1500 words for 6 speakers with severe to moderate dysarthria. PLA also shows an average relative WER reduction of 7.11% on speaker-dependent models evaluated using 5-fold cross-validation.

Index Terms— dysarthria, dysarthric speech, pronunciation lexicon adaptation, speech recognition

1. INTRODUCTION

Dysarthria encapsulates all neuro-motor articulatory disorders that result in acoustically unintelligible speech. It is often accompanied by other physical handicaps that limit the ability of individuals to interact with computers and the environment. Therefore, persons with dysarthria would benefit greatly from automatic speech recognition (ASR) applications. However, speaker-independent (SI) speech recognition systems remain ill-suited to this population because of the considerable deviation of dysarthric speech from the assumed norm in these systems.

Several existing attempts to apply ASR to dysarthric individuals involved small-vocabulary recognition tasks where the word error rate (WER) for dysarthric speech was shown to be significantly higher than for normal speech. For instance, in [1] it was reported that for 10 adults with dysarthria and 13 control subjects, the dysarthric subjects had significantly fewer stimuli recognized by the computer than the non-dysarthric speakers. Experiments with commercial ASR systems have shown promising performance for speakers with moderate to mild dysarthria on limited vocabulary tasks, but not for severely dysarthric speakers [2, 3]. For example, in a mixed read-speech and novel dictation context, commercial systems from Microsoft, Dragon, and VoicePad recognized approximately 85% of words uttered by a non-dysarthric speaker on average, but only between 51.87% and 64.68% of words spoken by a person with mild dysarthria [4]. For individuals with severe dysarthria, the vocabulary size of ASR systems tends to be extremely restricted. The STARDUST project [5], for instance, developed speech-controlled interfaces with a limited vocabulary of 10 isolated command words. In this paper, we aim to build a relatively large vocabulary ASR system, consisting of over 1500 words for speakers with severe to moderate dysarthria.

Although dysarthric speech deviates considerably from normal speech in many ways, it is nonetheless characterized by highly consistent articulatory errors [6]. These tend to result in relatively predictable errors of phoneme omission, substitution, addition, and distortion. Our motivation is therefore to exploit the intra-speaker consistency of these errors to improve recognition performance of a large vocabulary ASR system for dysarthric individuals. We use a 3-level cascaded adaptation procedure. First, we use maximum likelihood linear regression (MLLR) adaptation followed by maximum *a posteriori* (MAP) estimation to adapt a SI model to the vocal characteristics of a dysarthric speaker. We then analyze the pronunciation deviations of each dysarthric subject from the canonical form and build an associated speaker-specific pronunciation lexicon that incorporates the erroneous pronunciations of the speaker.

2. DESCRIPTION OF DATA

The TORGO database of dysarthric speech consists of aligned acoustic and articulatory recordings. At present, the database consists of 15 subjects of which eight (5 males, 3 females) are dysarthric, and seven (4 males, 3 females) are control subjects [7]. We are currently negotiating with the Linguistic Data Consortium to make all of the data described here available in early 2011.

Thanks to the Natural Sciences and Engineering Research Council of Canada and the University of Toronto for funding.

All dysarthric participants have been diagnosed by a speech-language pathologist according to the Frenchay Dysarthria Assessment [8], which evaluates the overall clinical intelligibility and the motor functions of the articulators. According to this assessment, four subjects (i.e., F01, M01, M02, and M04) are severely dysarthric. One subject (M05) is moderate-to-severely dysarthric, and one female subject (F03) is moderately dysarthric. The remaining two subjects have very mild dysarthria and are not considered as dysarthric speakers in this paper as their measured intelligibility ranks among the non-dysarthric speakers in this database.

Three hours of speech is recorded from each subject in multiple sessions, in which an average of 415 utterances are recorded from each dysarthric speaker and 800 from each control subject. The single word stimuli in the database include repetitions of English digits, the international radio alphabets, the 20 most frequent words in the British National Corpus (BNC), and a set of words selected by Kent *et al.* to demonstrate relevant phonetic contrasts [9]. The sentence stimuli are derived from the Yorkston-Beukelman assessment of intelligibility [10] and the TIMIT database [11]. In addition, each participant is asked to describe the contents of a few photographs that are selected from standardized tests of linguistic ability in his/her own words so as to include dictation-style speech into the database.

3. ACOUSTIC CHARACTERISTICS OF DYSARTHRIC SPEECH

We have empirically observed that the articulatory and pronunciation errors in dysarthric speech appear to be repeatable within individual speakers, which confirms existing clinical work [6]. It should therefore be possible to construct lexicons specific to individual dysarthric speakers, given their particular erroneous pronunciation patterns.

We listened to 25% of speech data from each dysarthric subject (this partition is later used as an adaptation-set) and we carefully analyzed the pronunciation deviations of each subject from the norm. Specifically, the desired phoneme sequence as determined by the CMU pronunciation dictionary[1] was compared against the actual phoneme sequences observed, and the deviations were recorded. These deviant pronunciations were then encoded into the generic pronunciation lexicon as alternatives to create a speaker-dependent lexicon. The observed deviations in the utterances of the six dysarthric subjects are grouped into the following classes:

- Final consonant deletion: Omission of word-final consonants that require more articulatory control (i.e., stops and fricatives). Examples include:
  feed → *[f iy], read* → *[r iy], beat* → *[b iy],*
  sheet → *[sh iy], thread* → *[th r eh], urged* → *[er jh]*
  clings → *[k l ih ng], tried* → *[t r ay], etc.*
  This is mainly observed in the utterances of F01, M01, and M04.

- Consonant cluster reduction: Omission of a more difficult consonant in a consonant cluster is observed in the speech of F01, M04 and M05. Examples include:
  bright → *[b ay t], explore* → *[ih p l ao r],*
  grow → *[g ow], play* → *[p ay], slip* → *[s ih p], etc.*

- Initial /s/ deletion: When /s/ is followed by a stop in a word initial syllable, F01, M01, and M04 often omit it. Examples include:
  spark → *[p ae r k], storm* → *[t ao r m],*
  spit → *[p ih t], snow* → *[n ow], etc.*

- Initial /h/ deletion: The voiceless glottal fricative sound /h/ is generally deleted when it occurs at the beginning of the target word in the utterances of M01. E.g.:
  hair → *[eh r], hitting* → *[ih t ih ng], hate* → *[ey t],*
  hat → *[ae t], house* → *[aw z], etc.*

- Devoicing: The voiceless counterpart of a voiced target is produced in the utterances of F03 and M02. E.g.:
  league → *[l iy k], bag* → *[b ae k],*
  deer → *[t ih r], ride* → *[r ay t], etc.*

- Prevocalic voicing: Voicing of voiceless target consonants is observed in F01, M01, and M04. E.g.:
  toe → *[d ow], feet* → *[v iy t], peer* → *[b ih r],*
  kitten → *[g ih d ah n], pile* → *[b ay l], etc.*

- Fronting: Consonants that are normally produced at the back of the alveolar ridge are substituted by consonants that are produced at or in front of the alveolar ridge. This is observed in F01, F03, M01, M04, and M05. E.g.:
  ship → *[s ih p], shoot* → *[s uw t],*
  share → *[s eh r], ring* → *[r iy n], etc.*

- Backing: In some cases, M02 substitutes /s/ which is produced further forward on the palate by /sh/ which is produced at the back of the palate. E.g.:
  spark → *[sh p ae r k], swarm* → *[sh w ao r m],*
  sip → *[sh ih p], suit* → *[sh uw t], etc.*

- Vocalization: Liquids (/l/ and /r/) are sometimes produced as vowels when they occur in word-final positions. This is predominantly observed in the speech of subjects F01, M02, and M04. E.g.:
  table → *[t ey b ow], double* → *[d ah b ow],*
  trouble → *[t r ah b ow], better* → *[b eh r ah] , etc.*

- Stopping: Substitution of a stop consonant for a fricative is observed in the utterances of subjects F01, F03, M01, and M04. E.g.:
  farm → *[p aa r m], single* → *[t ih ng g ah],*
  thorn → *[t ao r n], though* → *[d ow], etc.*

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

- Insertion of a short vowel in consonant clusters is also observed in the speech of subjects M01 and M02. E.g.: *slip* $\rightarrow$ *[s ih l ih p], floor* $\rightarrow$ *[f ih l ao r], blow* $\rightarrow$ *[b ih l ao w], bright*$\rightarrow$ *[b ih r ay t], etc.*

In many cases a combination of two or more of the above patterns are observed. Other articulatory errors include poor articulation of vowels and substitution of ejectives for some consonants (e.g. p' for p and k' for k).

4. EXPERIMENTS AND RESULTS

The acoustic features consist of 12 Mel-frequency cepstral coefficients (MFCCs), 1 energy term, and the corresponding delta (Δ) and acceleration ($\Delta\Delta$) coefficients generated every $15ms$ for severely dysarthric speech and every $10ms$ for moderately dysarthric speech. The use of $15ms$ frame rate for severely dysarthric speech has shown significant improvement in word recognition rates (by an average of 7.32% absolute). This may in part be due to the relatively slow speaking rates of severely dysarthric speakers. Cepstral mean subtraction (CMS) is also applied to account for a mismatch in channel conditions.

The baseline SI model consists of 40 left-to-right, three-state monophone hidden Markov models and one single-state short pause (sp) model with 16 Gaussian mixture components per state. We use a back-off bigram language model and the pronunciation lexicon is based on the CMU pronunciation dictionary. All the experiments in this paper are performed using the Hidden Markov Model Toolkit (HTK).

A subset of the TORGO database, described in Section 2, consisting of over 8400 utterances recorded from six dysarthric speakers, two subjects with mild dysarthria, and seven control (non-dysarthric) subjects is used in the following experiments. We considered three different training setups to train the SI models. First, we trained an SI model using data from the control speakers only, which resulted in a rather poor average word recognition rate of 30.41% on dysarthric speech. Second, we trained SI models using only dysarthric data from all the other dysarthric speakers except the test subject, which gave even worse word recognition rate (an average of 24.85%). The reason for the further degradation of performance in the latter setup is partly because the articulatory errors across dysarthric speakers vary widely although they tend to be consistent within a speaker. Finally, we used the merger of data from dysarthric and control speakers, excluding the test speaker, to train SI models. This resulted in a relatively better baseline word recognition rate (an average of 43.41%) and hence is used in the rest of the experiments. Word-internal triphone models show little improvement over the baseline monophone models for the dysarthric data. Therefore, monophone models are used as our baseline in the rest of the experiments. The evaluation-set consists of 75% of the utterances from the test dysarthric speaker (an average of 311 utterances) and the remaining 25% (an average of 104 utterances) is held as an adaptation-set.

Speaker-dependent (SD) models are also trained and evaluated using a 5-fold cross-validation procedure. For the dysarthric speakers, an average word recognition rate of 51.84% (ranging from an average of 12.1% to 82.7%) is obtained. The large performance discrepancy of the SD models is mainly due to the difference in the amount of training data we have for each dysarthric subject.

4.1. Acoustic Model Adaptation

Maximum Likelihood Linear Regression (MLLR) estimates linear transformations of model parameters to maximize the likelihood of the adaptation data. The transformations modify the component means and covariances in the initial model so as to reduce the mismatch between the model and the adaptation data.

Using the held-out 25% of data from each speaker as adaptation-set, we perform a two-pass MLLR adaptation. First a global adaptation is performed, which is then used as an input transformation to compute more specific transforms using a regression class tree with 42 terminals. MLLR, where both means and diagonal covariances are transformed, gave an average of 16.24% absolute (29.24% relative) WER reduction.

We then carried out two consecutive runs of Maximum *a Posteriori* (MAP) adaptation using the MLLR transformed models as the priors and maximizing the posterior probability using prior knowledge about the model parameter distribution. This further reduced the WER by an average of 3.9% absolute (10.75% relative).

4.2. Pronunciation Lexicon Adaptation (PLA)

The speaker-specific pronunciation lexicons consist of multiple pronunciations for some words that reflect the erroneous production of each dysarthric subject. The alternative pronunciations are added to all words that follow similar patterns as discussed in Section 3. For instance, for a speaker who omits initial /s/, the rule would be: for every word whose initial phoneme is /s/ and is followed by a stop consonant, add an alternative pronunciation where the initial /s/ is omitted.

Using speaker-dependent pronunciation lexicons during recognition reduces the WER further by an average of 2.73% absolute (8.29% relative) for the speaker-adapted (SA) models and by an average of 3% absolute (7.11% relative) for SD acoustic models. Significance tests are performed across all iterations of 5-fold cross validation, with pairing occurring between evaluations that are identical in every respect except for the type of lexicon used; i.e, generic or SD lexicon. In each case, the results are statistically significant at the 99% level of confidence according to the paired t-test.

Despite consistent improvements using PLA in both SD and SA models, errors remain. This can partially be explained

by the increased number of homophones in the speaker-dependent lexicons as a result of adapting the dictionary to pronunciation errors. For example, if a dysarthric speaker omits the voiceless glottal fricative /h/, pronunciations of the words *hate* and *hair* resemble that of *ate* and *air*, respectively, which are already words in the lexicon. This makes single-word recognition more difficult. In fact, 83% of the errors, on average, are single-word utterances. If homophones are considered equivalent, the relative gain due to PLA would be 11.72% (cf. 8.29%). Besides, of all the words correctly recognized due to PLA, 72.5% are in long utterances. This shows that PLA is more effective in relatively long utterances where context is available than in single-word utterances.

Figure 1 shows the word recognition accuracy of the baseline SI monophone models, the models after acoustic adaptation, the adapted models using SD lexicons, baseline SD acoustic models, and SD acoustic models using SD lexicons.

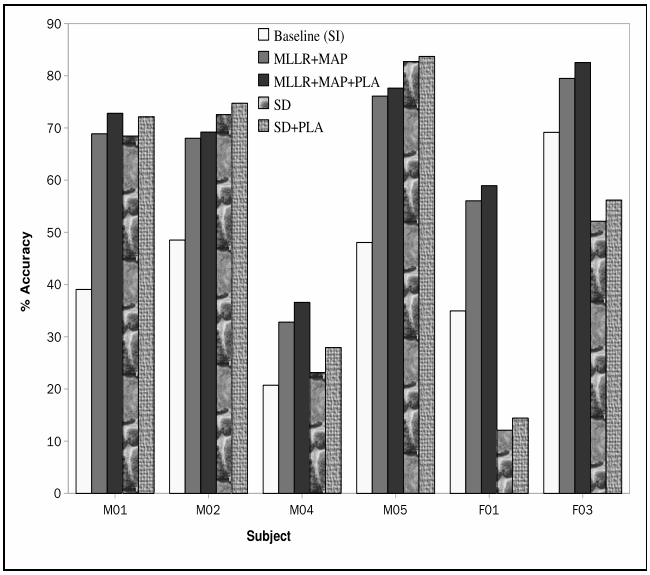

Fig. 1. Acoustic and Lexical Model Adaptation to Dysarthric Speech.

It can be seen in Figure 1 that PLA improves performance consistently for both SA and SD models. In most cases, SA models outperform SD models where the amount of data from a particular subject is relatively small and when the intra-speaker variation is too wide to be modeled by the available training data (e.g., subject M04).

5. CONCLUDING REMARKS

In this paper, we described the use of acoustic and lexical adaptation techniques to compensate for the articulatory errors made by speakers with dysarthria. We have shown that the consistent articulatory deviations in dysarthric speech can be exploited through speaker and pronunciation lexicon adaptation which resulted in an average of 22.87% absolute (42.11% relative) WER reduction, which is significant for the relatively large vocabulary size used in these experiments. PLA has shown statistically significant improvement on SA and SD models. While the results obtained are encouraging, phonetic articulatory errors are only part of the problem in dysarthric speech. In addition to the articulatory errors discussed in this paper, severely dysarthric speech consists of involuntary breathing, irregular articulatory breakdowns, prosodic disruptions, stuttering, and accidental pauses which make the task more complex. We are currently investigating approaches to deal with these challenges.

6. REFERENCES

[1] C. Coleman and L. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria," *Augmentative and Alternative Communication*, vol. 7, pp. 34–42, 1991.

[2] 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*, vol. 17, no. 4, pp. 265–275, 2001.

[3] N. J. Manasse, K. Hux, and J. L. Rankin-Erickson, "Speech recognition training for enhancing written language generation by a traumatic brain injury survivor," *Brain Injury*, vol. 14, no. 11, pp. 1015–1034, 2000.

[4] K. Hux, J. Rankin-Erickson, N. Manasse, and E. Lauritzen, "Accuracy of three speech recognition systems: Case study of dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 3, pp. 186–196, 2000.

[5] M. Parker, S. Cunninghm, P. Enderby, M. Hawley, and P. Green, "Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project," *Clinical Linguistics and Phonetics*, vol. 20, no. 2-3, pp. 149–156, 2006.

[6] K. M. Yorkston, D. R. Beukelman, and K. R. Bell, *Clinical management of dysarthric speakers*, San Diego, CA: College-Hill Press, 1988.

[7] F. Rudzicz, A. Namasivayam, and T Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation, in press*, 2010.

[8] P. Enderby, "Frenchay dysarthria assessment," *International Journal of Language & Communication Disorders*, vol. 15, no. 3, pp. 165–173, 1980.

[9] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, 1989.

[10] K. M. Yorkston and D. R. Beukelman, *Assesment of intelligibility of dysarthric speech*, Tigard, Oregon: C.C. Publications Inc., 1981.

[11] V. Zue, S. Seneff, and J. R. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.