Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech

Kinfe Tadesse Mengistu and Frank Rudzicz

University of Toronto,
Department of Computer Science
6 King's College Road
Toronto, Ontario, Canada
{kinfe,frank}@cs.toronto.edu

Abstract. Speech is a complex process that requires control and coordination of articulation, breathing, voicing, and prosody. Dysarthria is a manifestation of an inability to control and coordinate one or more of these aspects, which results in poorly articulated and hardly intelligible speech. Hence individuals with dysarthria are rarely understood by human listeners. In this paper, we compare and evaluate how well dysarthric speech can be recognized by an automatic speech recognition system (ASR) and naïve adult human listeners. The results show that despite the encouraging performance of ASR systems, and contrary to the claims in other studies, on average human listeners perform better in recognizing single-word dysarthric speech. In particular, the mean word recognition accuracy of speaker-adapted monophone ASR systems on stimuli produced by six dysarthric speakers is 68.39% while the mean percentage correct response of 14 naïve human listeners on the same speech is 79.78% as evaluated using single-word multiple-choice intelligibility test.

Keywords: speech recognition, dysarthric speech, intelligibility

1 Introduction

Dysarthria is a neurogenic motor speech impairment which is characterized by slow, weak, imprecise, or uncoordinated movements of the speech musculature [1] resulting in unintelligible speech. This impairment results from damage to neural mechanisms that regulate the physical production of speech and is often accompanied by other physical handicaps that limit interaction with modalities such as standard keyboards. Automatic speech recognition (ASR) can, therefore, assist individuals with dysarthria to interact with computers and control their environments. However, the deviation of dysarthric speech from the assumed norm in most ASR systems makes the benefits of current speaker-independent (SI) speech recognition systems unavailable to this population of users.

Although reduced intelligibility is one of the distinguishing characteristics of dysarthric speech, it is also characterized by highly consistent articulatory errors [1]. The consistency of errors in dysarthric speech can, in principle, be exploited to build an ASR system specifically tailored to a particular dysarthric speaker since ASR models do not necessarily require intelligible speech as long as consistently articulated speech is available. However, building a speaker-dependent (SD) model trained of spoken data from an individual dysarthric speaker is practically infeasible due to the difficulty of collecting large enough amount of training data from a dysarthric subject. Therefore, a viable alternative is to adapt an existing SI model to the vocal characteristics of a given dysarthric individual.

The purpose of this study is to compare naïve human listeners and speaker-adapted automatic speech recognition (ASR) systems in recognizing dysarthric speech and to investigate the relationship between intelligibility and ASR performance. In earlier studies, it has been shown that ASR systems may outperform human listeners in recognizing impaired speech [2–4]. However, since intelligibility is typically a relative rather than an absolute measure [5], these results do not necessarily generalize. Intelligibility may vary depending on the size and type of vocabulary used, the familiarity of the listeners with the intended message or the speakers, the quality of recording (i.e. the signal-to-noise ratio), and the type of response format used.

Yorkston and Beukelman [6] compared three different types of response formats: transcription, sentence completion, and multiple choice. In transcription, listeners were asked to transcribe the word or words that have been spoken. In sentence completion, listeners were asked to complete sentences from which a single word had been deleted. In the multiple choice format, listeners selected the spoken word from a list of phonetically similar alternatives. Their results indicated that transcription was associated with lowest intelligibility scores, while multiple choice tasks were associated with the highest scores. This clearly shows that listeners' performance can vary considerably depending on the type of response format used. Therefore, when comparing human listeners and an ASR system, the comparison should be made on a level ground; i.e., both should be given the same set of alternative words (foils) from which to choose. In other words, it would be unfair to compare an ASR system and a human listener without having a common vocabulary, and since the innate vocabulary of our participants is unknown (but may exceed 17,000 base words [7]), we opt for a small common vocabulary. Hence, the multiple choice response format is chosen in this paper.

2 Method

2.1 Speakers

The TORGO database consists of 15 subjects, of which eight are dysarthric (five males, three females), and seven are non-dysarthric control subjects (four males, three females) [8]. All dysarthric participants have been diagnosed by a

speech-language pathologist according to the Frenchay Dysarthria Assessment [9] to determine the severity of their deficits. According to this assessment, four speakers (i.e., F01, M01, M02, and M04) are severely dysarthric, one speaker (M05) is moderately-to-severely dysarthric, and one subject (F03) is moderately dysarthric. Two subjects (M03 and F04) have very mild dysarthria and are not considered as dysarthric in this paper as their measured intelligibility is not substantially different from the non-dysarthric speakers in the database.

2.2 Speech Stimuli

Three hours of speech are 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 phonetic contrasts [5]. 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 in his or her own words the contents of a few photographs that are selected from standardized tests of linguistic ability so as to include dictation-style speech in the database.

A total of 1004 single-word utterances were selected from the recordings of the dysarthric speakers and 808 from control speakers for this study. These consist of 607 unique words. Each listener is presented with 18% of the data (single-word utterances) from each dysarthric subject where 5% of randomly selected utterances are repeated for intra-listener agreement analysis resulting in a total of 180 utterances from the six dysarthric individuals. In addition, a total of 100 single-word utterances are selected from three male and three female control subjects comprising about 6% of utterances from each speaker. Altogether, each participant listens to a total of 280 speech files which are presented in a random order. Inter-listener agreement is measured by ensuring that each utterance is presented to at least two listeners.

2.3 Listeners

Fourteen native North American English speakers who had no previous familiarity with dysarthric speech and without hearing or vision impairment were recruited as listeners. The listening task consisted of a closed-set multiple-choice selection in which listeners were informed that they would be listening to a list of single-word utterances spoken by individuals with and without speech disorders in a random order. For every spoken word, a listener was required to select a word that best matched his/her interpretation from among a list of eight alternatives. Four of the seven foils were automatically selected from phonetically similar words in the pronunciation lexicon, differing from the true word in one or two phonemes. The other three foils were generated by an HMM-based speech recognizer trained on the entire data to produce an N-best list such that the first

three unique words different from the target word are selected. Listeners were allowed to replay prompts as many times as they want.

3 Intelligibility Test Results

For each listener, the percentages of correct responses out of the 180 dysarthric prompts and 100 non-dysarthric prompts were calculated separately. The correct percentages were then averaged across the 14 listeners to compute the mean recognition score of naïve human listeners on dysarthric and non-dysarthric speech. Accordingly, the mean recognition score of human listeners is 79.78% for stimuli produced by dysarthric speakers and 94.4% for stimuli produced by control speakers. Figure 1 depicts the recognition score of the 14 naïve listeners on stimuli produced by dysarthric and control speakers.

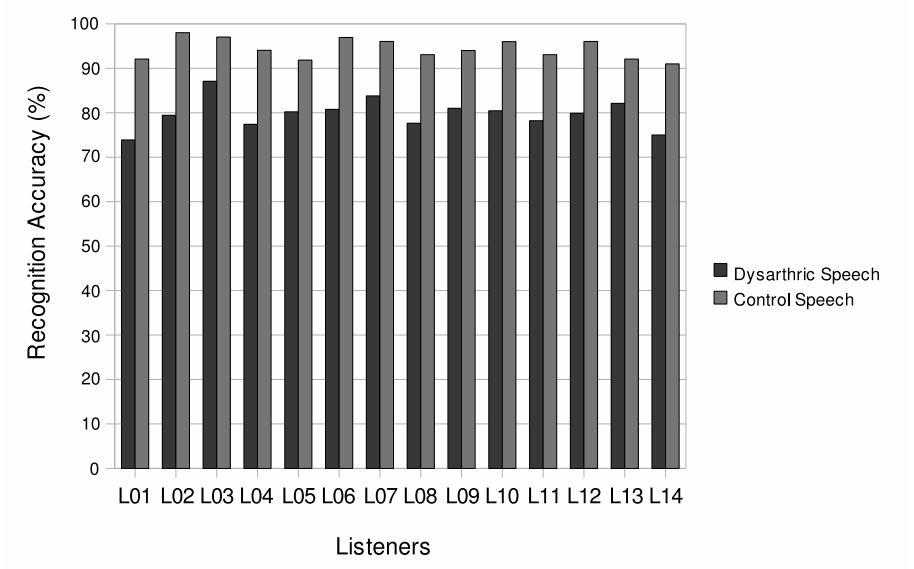

Fig. 1. Word recognition score of 14 naïve human listeners

To measure the intelligibility of stimuli produced by a speaker, the responses of all listeners for the stimuli produced by that speaker are collected together and the percentage of correct identifications is computed. Accordingly, for severely dysarthric speakers, the intelligibility score ranged from 69.05% – 81.88% with the mean score being 75.2%. Speaker M05, who is moderately-to-severely dysarthric, had 87.88% of his words correctly recognized, and the moderately dysarthric speaker F03 had 90% of her words recognized correctly. These results are presented in Figure 2.

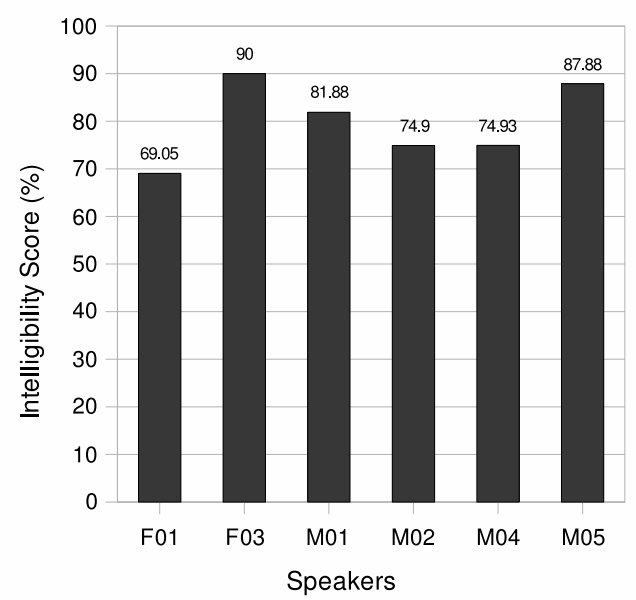

Fig. 2. Intelligibility score of six dysarthric speakers as rated by 14 naïve human listeners

On average, listeners agreed on common utterances between 72.2% and 81.6% of the time with the mean inter-listener agreement being 77.2%. The probability of chance agreement here is 12.5% since there are 8 choices per utterance.

Intra-listener reliability is measured as the proportion of times that a listener identifies the same word across two presentations of the same audio prompt. The mean intra-listener agreement across all listeners is 88.5%, with the lowest being 79.6% and the highest being 96.3% (listeners 7 and 10).

4 ASR Experiments and Results

4.1 Data Description

The speaker-independent (SI) acoustic models are built using a subset of the TORGO database consisting of over 8400 utterances recorded from six dysarthric speakers, two speakers with very mild dysarthria, and seven control subjects. The SI models are trained and evaluated using the leave-one-out method; i.e., data from one speaker are held out for evaluation while all the remaining data from the other speakers are used for training. The held-out data from the test speaker is divided into an evaluation-set and an adaptation-set. The evaluation-set consists of all unique single-word stimuli spoken by the test dysarthric speaker (described in Section 2.2) while the remaining data are later used as adaptation-set to adapt a SI acoustic model to the vocal characteristics of a particular dysarthric speaker.

4.2 Acoustic Features

We compare the performance of acoustic models based on Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding-based Cepstral Coefficients (LPCCs), and Perceptual Linear Prediction (PLP) coefficients with various feature parameters, including the use of Cepstral Mean Subtraction (CMS) and, the use of the 0^{th} order cepstral coefficient as the energy term instead of the log of the signal energy. The use of CMS was found to be counterproductive in all cases. This is because single-word utterances are very short and CMS is only useful for utterances longer than 2–4 seconds [12]. The recognition performance of the baseline SI monophone models based on MFCC and PLP coefficients with the 0^{th} order cepstral coefficient are comparable (39.94% and 39.5%) while LPCC-based models gave the worst baseline recognition performance of 34.33%. Further comparison on PLP and MFCC features on speaker-adapted systems showed that PLP-based acoustic models outperformed MFCC-based systems by 2.5% absolute. As described in [13], PLP features are more suitable in noisy conditions due to the use of different non-linearity compression; i.e., the cube root instead of the logarithm on the filter-bank output. The data used in these experiments consist of considerable background noise and other type of noise produced by the speakers due to hyper-nasality and breathy voices. These aspects may explain why PLP performed better than MFCCs and LPCCs in these experiments. The rest of the experiments presented in this paper are based on PLP acoustic features. PLP incorporates the known perceptual properties of human hearing, namely critical band frequency resolution, pre-emphasis with an equal loudness curve, and the power law model of hearing.

A feature vector containing 13 cepstral components, including the 0^{th} order cepstral coefficient and the corresponding delta and delta-delta coefficients comprising 39 dimensions, is generated every 15 ms for dysarthric speech and every 10 ms for non-dysarthric speech.

4.3 Speaker-Independent Baseline Models

The baseline SI systems consist of 40 left-to-right, 3-state monophone hidden Markov models and one single-state short pause (sp) model with 16 Gaussian mixture components per state. During recognition, the eight words that are used as alternatives for every spoken test utterance during the listening experiments are formulated as an eight-word finite-state grammar which is automatically parsed into the format required by the speech recognizer. The pronunciation lexicon is based on the CMU pronunciation dictionary[1]. All ASR experiments are performed using the Hidden Markov Model Toolkit (HTK) [14].

The mean recognition accuracy of the baseline SI monophone models using PLP acoustic features on single-word recognition where eight alternatives are provided for each utterance is 39.5%. The poor performance of the SI models in recognizing dysarthric speech is not surprising since data from each dysarthric

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

speaker deviates considerably from the training data. Word-internal triphone models show little improvement over the baseline monophone models for the dysarthric data in our database. Hence, we use the monophone models as our baseline in the rest of the experiments.

4.4 Acoustic and Lexical Model Adaptation

To improve recognition accuracy, the SI models are tailored to the vocal characteristics of each dysarthric subject. Here 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 each SI model to the vocal characteristics of a particular dysarthric subject. We then analyze the pronunciation deviations of each dysarthric subject from the canonical form and build an associated speaker-specific pronunciation lexicon that incorporates their particular behavior of pronunciation.

Using the adaptation data from a particular speaker, 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. We then carry out 2 to 5 consecutive iterations of Maximum *a Posteriori* (MAP) adaptation using the models that have been transformed by MLLR as the priors and maximizing the posterior probability using prior knowledge about the model parameter distribution. This process resulted in 25.81% absolute (43.07% relative) improvement.

Using speaker-dependent (SD) pronunciation lexicons, constructed as described in [15], during recognition improved the word recognition rate further by an average of 3.18% absolute (8.64% relative). The SD pronunciation lexicons consist of multiple pronunciations for some words that reflect the particular pronunciation pattern of each dysarthric subject. In particular, we listened to 25% of speech data from each dysarthric subject and carefully analyzed the pronunciation deviations of each subject from the norm; i.e., the desired phoneme sequence as determined by the CMU pronunciation dictionary 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 existing pronunciations [15]. Figure 3 depicts the performance of the baseline and speaker-adapted (SA) models on dysarthric speech.

In total, the cascaded approach of acoustic and lexical adaptation improved the recognition accuracy significantly by 28.99% absolute (47.94% relative) over the baseline yielding a mean word recognition accuracy of 68.39%.

For non-dysarthric speech, the mean word recognition accuracy of the SI baseline monophone models is 71.13%. After acoustic model adaptation, the mean word recognition accuracy rises to 88.55%.

5 Discussion of Results

When we compare the performance of the speaker-adapted ASR systems with the intelligibility rating of the human listeners on dysarthric speech, we observe

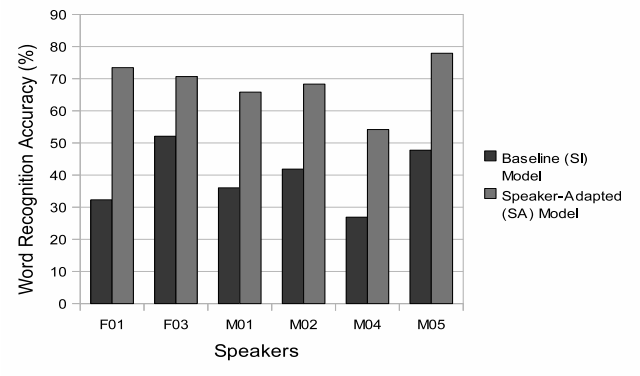

Fig. 3. ASR performance on dysarthric speech

that in most cases human listeners are more effective at recognizing dysarthric speech. However, an ASR system recognized more stimuli produced by speaker F01 than the human listeners. Figure 4 summarizes the results.

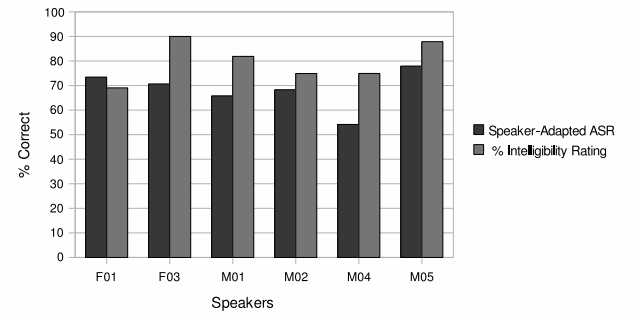

Fig. 4. Human listeners vs. ASR system recognition scores on dysarthric speech

Humans are typically robust at speech recognition in the presence of even very low signal-to-noise ratios [16]. This may partially explain their relatively high performance here. Dysarthric speech contains not only distorted acoustic information due to imprecise articulation but also undesirable acoustic noise due to improper breathing that severely degrades ASR performance. Due to the remarkable ability of human listeners to separate and pay selective attention to the different sound sources in a noisy environment [17], the acoustic noise due to improper breathing has less impact on human listeners than in ASR systems. For instance, the audible noise produced by breathy voices and hyper-nasality is strong enough to confuse ASR systems while human listeners can easily ignore it. This suggests that noise resilience is an area that should further be investigated

to improve ASR performance to dysarthric speech. Furthermore, approaches to deal with other features of dysarthric speech such as stuttering, prosodic disruptions, and inappropriate intra-word pauses are areas for further investigation in order to build an ASR system that possesses comparable performance with human-listeners in recognizing dysarthric speech.

Although there appears to exist some relationship between intelligibility ratings and ASR performance, the latter is especially affected by the level of background noise, and the involuntary noise produced by the dysarthric speakers. The impact of hyper-nasality and breathy voice appears to be more severe in ASR systems than in the intelligibility rating among human listeners on single-word utterances. F01, for instance, is severely dysarthric but the ASR performs better than the human listeners because most of the errors in her speech could be offset by acoustic and lexical adaptation. M04, on the other hand, who is also severely dysarthric, was relatively more intelligible but was the least well understood by the corresponding speaker-adapted ASR system since this speaker is characterized by breathy voice, prosodic disruptions, and stuttering.

6 Concluding remarks

In this paper we compared naïve human listeners and speaker-adapted automatic speech recognition systems in recognizing dysarthric speech. Since intelligibility may vary widely depending on the type of stimuli and response format used, our basis of comparison is designed so that both the human listeners and the ASR systems are compared on a level ground. Here, we use multiple choice format from a closed set of eight alternatives, where the same set of alternatives are provided for every single-word utterance to both the ASR systems and to the human listeners. Although, there is one case in which a speaker-adapted ASR system performed better than the aggregate of human listeners, in most cases the human listeners are more effective in recognizing dysarthric speech than ASR systems. However, the mean word recognition accuracy of the speaker-adapted ASR systems (68.39%) relative to the baseline of 39.5% is encouraging. Future work ought to concentrate on an improved method to deal with breathy voice, stuttering, prosodic disruptions, and inappropriate pauses in dysarthric speech to further improve ASR performance.

Acknowledgments. This research project is funded by the Natural Sciences and Engineering Research Council of Canada and the University of Toronto.

References

1. Yorkston, K.M., Beukelman, D.R., Bell, K.R.: Clinical Management of Dysarthric Speakers. Little, Brown and Company (Inc.), Boston (1988)
2. Carlson, G.S., Bernstein, J.: Speech recognition of impaired speech. In: Proceedings of RESNA 10th Annual Conference, pp. 103–105 (1987)

3. Stevens, G., Bernstein, J.: Intelligibility and machine recognition of deaf speech. In: Proceedings of RESNA 8th Annual Conference, pp. 308–310 (1985)
4. Sharma, H.V., Hasegawa-Johnson, M., Gunderson,J., Perlman, A.: Universal access: speech recognition for talkers with spastic dysarthria. In: Proceedings of INTERSPEECH-2009. 1451–1454 (2009)
5. Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C.: Toward phonetic intelligibility testing in dysarthria. In: Journal of Speech and Hearing Disorders 54, 482–499 (1989)
6. Yorkston, K.M., Beukelman, D.R.: A comparison of techniques for measuring intelligibility of dysarthric speech. In: Journal of Communication Disorders 11, 499–512 (1978)
7. Goulden, R., Nation, P., Read, J.: How large can a receptive vocabulary be? In: Applied Linguistics 11 341–363 (1990)
8. Rudzicz, F., Namasivayam, A., Wolff, T: The TORGO database of acoustic and articulatory speech from speakers with dysarthria. In: Language Resources and Evaluation, *in press* (2011)
9. Enderby, P.: Frenchay Dysarthria Assessment. In: International Journal of Language & Communication Disorders 15 (3), 165–173 (1980)
10. Yorkston, K.M., Beukelman, D.R.: Assessment of Intelligibility of Dysarthric Speech. Tigard, Oregon: C.C. Publications Inc. (1981)
11. Zue, V., Seneff, S., Glass, J.R.: Speech database development at MIT: TIMIT and beyond. In: Speech Communication 9(4), 351–356 (1990)
12. Alsteris, L. D., Paliwal, K. K.: Evaluation of the Modified Group Delay Feature for Isolated Word Recognition. In: Proceedings of International Symposium on Signal Processing and Applications, pp. 715–718 (2005)
13. Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. In: Journal of the Acoustical Society of America 87(4), 1738-1752 (1990)
14. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book. Revised for HTK Version 3.4, Cambridge University Engineering Department (2006)
15. Mengistu, Kinfe T., Rudzicz, F.: Adapting Acoustic and Lexical Models to Dysarthric Speech. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, *in press* (2011)
16. Lippmann, R.: Speech recognition by machines and humans. In: Speech Communication 22(1), 1–15 (1997)
17. Bregman, A.S.: Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, Massachusetts: MIT Press (1990)