Expert Systems with Applications 41 (2014) 841–852

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition

Santiago-Omar Caballero-Morales *, Felipe Trujillo-Romero

Technological University of the Mixteca, Road to Acatlima K.m. 2.5, Huajuapan de León, Oaxaca 69000, Mexico

ARTICLE INFO

Keywords:
Evolutionary algorithms
Dysarthric speech recognition
Metamodels

ABSTRACT

Dysarthria is a motor speech disorder caused by neurological injury of the motor component of the motor-speech system. Because it affects respiration, phonation, and articulation, it leads to different types of impairments in intelligibility, audibility, and efficiency of vocal communication. Speech Assistive Technology (SAT) has been developed with different approaches for dysarthric speech and in this paper we focus on the approach that is based on modeling of pronunciation patterns. We present an approach that integrates multiple pronunciation patterns for enhancement of dysarthric speech recognition. This integration is performed by weighting the responses of an Automatic Speech Recognition (ASR) system when different language model restrictions are set. The weight for each response is estimated by a Genetic Algorithm (GA) that also optimizes the structure of the implementation technique (Metamodels) which is based on discrete Hidden Markov Models (HMMs). The GA makes use of dynamic uniform mutation/crossover to further diversify the candidate sets of weights and structures to improve the performance of the Metamodels. To test the approach with a larger vocabulary than in previous works, we orthographically and phonetically labeled extended acoustic resources from the Nemours database of dysarthric speech. ASR tests on these resources with the proposed approach showed recognition accuracies over those obtained with standard Metamodels and a well used speaker adaptation technique. These results were statistically significant.

1. Introduction

Dysarthria is a motor speech disorder that is often associated with irregular phonation and amplitude, in-coordination of articulators, and restricted movement of articulators (Kain, Niu, Hosom, Miao, & Santen, 2004). This condition can be caused by neurological injury of the motor component of the motor-speech system, affecting articulation, resonance, and respiration. Hence, a stroke, cerebral palsy, traumatic brain injury (TBI), or degenerative neurological diseases such as Parkinson or Alzheimer can lead to this disorder.

The muscles affected by this condition may include the lungs, larynx, oropharynx and nasopharynx, soft palate and articulators (lips, tongue, teeth and jaw), and the degree to which these muscle groups are compromised determines the particular pattern of speech impairment (Kain et al., 2004). Phonatory dysfunction is a frequent impairment which is a primary feature noted in clinical assessment of dysarthric speech (Darley, Aronson, & Brown, 1969; Kent, Kent, Duffy, & Weismer, 1998, 2003).

As a consequence of phonatory dysfunction, dysarthric speech is typically characterized by strained phonation, imprecise placement of the articulators, and incomplete consonants closure, leading to reduced intelligibility and phonemic repertoire. Because of these articulatory deficits, the pronunciation of dysarthric speakers often deviates from that of non-dysarthric speakers in several aspects: rate of speech is lower, segments are pronounced differently, pronunciation is less consistent, and for longer stretches of speech, pronunciation can be even more varying due to fatigue (Strik, Sanders, Ruiter, & Beijer, 2002).

Speech Assistive Technology (SAT), which incorporates Automatic Speech Recognition (ASR), has been developed to improve interaction and communication of people with dysarthria. This technology presents advantages for these tasks when compared to interfaces such as switches or keyboards which may be more physically demanding and tiring (Green, Carmichael, Hatzis, Enderby, & Hawley, 2003; Kotler & Tam, 2002; Kotler, Thomas-Stonell, Leeper, & Doyle, 1998; Poock, Lee, & Blackstone, 1987). Also, ASR can be superior to that of humans as intelligibility declines into the moderate to severe range as measured by intelligibility tests (Ferrier, Shane, Ballard, Carpenter, & Benoit, 1995).

However the development of ASR systems is very challenging because phonatory dysfunction increases deletion, insertion, and

* Corresponding author.
E-mail addresses: scaballero@mixteco.utm.mx (S.-O. Caballero-Morales), ftrujillo@mixteco.utm.mx (F. Trujillo-Romero).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.014

substitution of phonemes in the recognition process (Hamidi, Baljko, Livingston, & Spaltecholz, 2010; Hasegawa-Johnson, Gunderson, Perlman, & Huang, 2006; Kent, Wang, Duffy, & Thomas, 2005; Polur & Miller, 2005; Niu & Santen, 2003; Raghavendra, Rosengren, & Hunnicutt, 2001; Rosen & Yampolsky, 2000; Strik et al., 2002). Tests with commercial Speaker Independent (SI) ASR systems for non-disordered speech have shown varying levels of performance.[1] In Raghavendra et al. (2001) recognition accuracies within the range of 22–97% were obtained for speakers with severe to moderate levels of dysarthria using the Prototype Swedish Dragon Dictate (PSDD) and Infovox RA. In Manasse, Hux, and Rankin-Erickson (2000), Manasse, Hux, Rankin-Erickson, and Lauritzen (2000) Microsoft Dictation and Dragon, NaturallySpeaking, and VoicePad Platinum were tested with a speaker with moderate dysarthria, obtaining accuracies of 50–95%. However in these works high performance usually was obtained after extensive training sessions (up to eight hours), assistance, and adjustements of the ASR system's vocabulary.

In addition, these systems usually require some speech samples from the speaker to get adapted to his/her voice and increase recognition performance. Because these systems are trained with normal (non-disordered) speech, they are not expected to work well on severely dysarthric speech as adaptation techniques (Leggetter & Woodland, 1995; Young & Woodland, 2006) may be insufficient to deal with gross abnormalities (Green et al., 2003). ASR systems specifically developed for dysarthric speech have shown more stable performance. In Hawley et al. (2007) Speaker Dependent (SD) ASR systems were built for individual speakers.[2] These were based on Hidden Markov Models (HMMs) (Rabiner, 1989) and achieved recognition accuracies of 88.5–95.4% for speakers with severe dysarthria on a limited vocabulary (10 command words). In Hasegawa-Johnson et al. (2006) SD ASR systems based on HMMs and SVMs (Support Vector Machines) were built for speakers with low intelligibility. A conclusion of this work was that HMMs have some degree of robustness against large-scale word-length fluctuations, while SVM have some robustness against the reduction and deletion of consonants. In Hamidi et al. (2010) a HMM-based SD ASR system was developed for the task of activating different parts of a web browser with a vocabulary of 47 pre-selected words obtaining accuracies of 37.5–83.3%. However developing SD ASR systems for individual speakers also requires a significant investment of time and assistance.

Thus, alternative approaches have been focused to better adapt SI and SD ASR systems to dysarthric speakers to reduce the development time and improve recognition accuracy (Ben-Mosbah, 2006). Articulatory and pronunciation knowledge have been proposed to perform more specific adaptation of ASR systems (Christina, Vijayalakshmi, & Nagarajan, 2009; Mengistu & Rudzicz, 2011; Rudzicz, 2011). In Levit, Alshawi, Gorin, and Nöth (2003) pronunciation patterns were used to correct the phoneme output of an ASR system. These patterns were integrated in the ASR process with a set of Weighted Finite State Transducers (WFSTs). The work in Levit et al. (2003) was extended in Caballero and Cox (2009a) where the phoneme pronunciation patterns were modeled to improve the word output of adapted SI ASR systems. These patterns were integrated in the ASR process by using WFSTs and the technique of Metamodels (Dasmahapatra & Cox, 2000) which consists of a set of discrete HMMs with a specific structure.

In Caballero and Cox (2009a) the information estimated from a speaker's phoneme confusion - matrix was used for the modeling of the pronunciation patterns with Metamodels and WFSTs. A phoneme confusion-matrix is an important resource to identify phonetic errors in the response of the ASR system. With this approach, statistically significant gains of 5% and 3% in word recognition accuracy were obtained respectively on severe and moderate dysarthric speech when compared to other techniques. Other works as those presented in Matsumasa, Takiguchi, Ariki, LI, and Nakabayash (2009), Wu, Su, and Shen (2011), Seong, Park, and Kim (2012) followed this approach and also obtained significant improvements on baseline ASR performances.

An important issue of the approaches in Levit et al. (2003), Caballero and Cox (2009a), Matsumasa et al. (2009), Wu et al. (2011) and Seong et al. (2012) is that the pronunciation patterns were estimated from a single output (response) of the ASR system. In Levit et al. (2003) the phoneme sequences for confusion-matrix estimation were obtained from the phoneme response of an SI ASR system. In Caballero and Cox (2009b) the phoneme sequences were obtained from the phonetic transcriptions of the word output of an adapted SI ASR system. An ASR system's output is estimated by the following mathematical model:

$$\widehat{W} = \max_{W \in L} P(O|W)P(W) \tag{1}$$

where $\widehat{W}$ is the most likely ASR system's output (vocabulary units); O is the acoustic observation (speech signal); $P(W)$ is the *prior probability* or statistical knowledge about the structure and vocabulary units in the used language (Language Model, L); and $P(O|W)$ is the *observation likelihood* or the probability that an acoustic observation O is associated to a vocabulary unit W (Acoustic Model) (Jurafsky & Martin, 2009; Young & Woodland, 2006). $P(W)$ is usually estimated by using N-gram grammars, and $P(O|W)$ is usually modeled by using Hidden Markov Models (HMMs) (Rabiner, 1989), or Artificial Neural Networks (ANN). For word recognition tasks, the ASR system estimates the sequence of the most likely phoneme models (HMMs) that represents the speech O. Then, this sequence is restricted to form words by incorporating a lexicon. Finally, these words are restricted by the information of the language model $P(W)$ to form valid sequences of words. The influence of $P(W)$ is very important for the response of the ASR system and for the estimation of the phoneme confusion-matrix (Caballero & Cox, 2009b).

In this paper we present an evolutionary approach to obtain more accurate estimates of a phoneme confusion-matrix to increase ASR performance for dysarthric speech. Instead of using a single ASR system's response as source for pronunciation estimation, we change the influence of $P(W)$ in the ASR process to obtain multiple responses with pronunciation patterns not present in the single source approach. The patterns from the multiple responses then are weighted and integrated into a single phoneme confusion-matrix by means of a Genetic Algorithm (GA). The implementation technique for the incorporation of the integrated confusion-matrix into the ASR process is the extended version of the Metamodels presented in Caballero (2011). The GA considers dynamic mutation/crossover to further diversify the candidate sets of confusion-matrix weights and Metamodel structures to improve ASR performance.

One problem in ASR research for dysarthric speech is the limited availability of speech data for development and experimentation. The use of specialized speech corpora has been used for this purpose. In this work we consider the Nemours database of dysarthric speech (Bunnel, Polikoff, Menéndez-Pidal, Peters, & Leonzio, 1996) which consists of speech data from 10 american english speakers with varying levels of dysarthria. In contrasts to other works (Christina et al., 2009; Niu & Santen, 2003; Rudzicz, 2011; Seong et al., 2012) and previous research (Caballero & Cox, 2009a), we performed orthographic and phonetic labeling of additional acoustic resources from the Nemours database to perform tests with a larger vocabulary.

[1] The SI ASR system is trained with the speech of hundreds or thousands of speakers different from the one that will use it.

[2] The SD ASR system is trained with the speech of the single speaker that will use it.