interspeech 2008 incorporating SST 08

Verifying Pronunciation Accuracy from Speakers with Neuromuscular Disorders

Shou-Chun Yin[1], *Richard Rose*[1], *Oscar Saz*[2], *Eduardo Lleida*[2]

[1]Department of Electrical and Computer Engineering, McGill University, Montreal, Canada Montreal, Canada

[2]Communication Technologies Group (GTC) I3A University of Zaragoza, Spain

sss123ca@yahoo.com, rose@ece.mcgill.ca, oskarsaz@unizar.es, lleida@unizar.es

Abstract

This paper presents a study of confidence measure based techniques for detecting phoneme level mispronunciations in utterances from impaired children with neuromuscular disorders. Several different adaptation scenarios are investigated to determine the effects of mismatched speaker characteristics and mismatched task domain on the ability to verify the phoneme level pronunciations. These techniques are evaluated in the context of a speech corpus where utterances were elicited from children in interactive speech therapy sessions involving a multimodal game-like environment. Results are presented in terms of phone detection charateristics where, for example, equal error rates of as low as 16.2% were obtained for detecting instances where phonemes were deleted by impaired speakers.

1. Introduction

The value of speech-based interfaces for children with neuromuscular disorders has received a great deal of attention in the education and learning technology community [1]. These disorders can result in slurred and often unintelligible speech which limits the effectiveness of speech technology tools which are originally developed for an unimpaired user population [2, 3]. This condition is caused by a number of different medical conditions and is referred to in the speech pathology literature as dysarthria [4]. There are several applications that incorporate automatic speech recognition (ASR) technology that could potentially benefit from the use of acoustic models and pronunciation models which better represent the class of variabilities that are characteristic of dysarthric speech [5]. These include automated language learning for disabled individuals, automatic diagnosis of dysarthria and related conditions from speech utterances, and ASR systems that are robust with respect these varibilities.

This paper investigates the problem of detecting the occurrence of phone level mispronunciations that occur in utterances obtained from disabled children in a limited task domain. The task domain is defined by an interactive children's learning tool that accepts isolated word Spanish language speech input from users. The user population includes both disabled and non-disabled children whose first language is Spanish. The speaker population, task domain, and speech corpus that was collected from this domain are described in Section 2. It is assumed in the paper that all models used for pronunciation verification are trained from the non-disabled population and used to detect portions of utterances that have been mispronounced by disabled speakers.

It is well known that neuromuscular disorders affect speech at all levels including frame level spectral characteristics, segment level coarticulation, lexical level pronunciation rules, and super-segmental prosodic contours [4]. The development of precise models of these effects requires a description of how these disorders are reflected in the underlying articulatory dynamics of speech production. While a long-term goal is to develop techniques that exploit effects that occur in an articulatory space, the goal of this paper is more modest. The goal here is to determine the ability of techniques based on existing ASR formalisms to determine when dysarthria induced variability occurs in speech.

The techniques that are investigated here are similar to those applied to pronunciation verification for automated language learning and language skills evaluation applications [6]. Phone level measures of confidence are derived from the acoustic speech utterance and are used to define a decision rule for accepting or rejecting the hypothesis that a phone or syllable was mispronounced. The confidence measures used here are based on posterior probabilities derived from phone and syllable lattices and are described in Section 3. The application of acoustic model adaptation to limit the effects of other sources of variability is also described in Section 3. An experimental study was performed to evaluate the ability of these techniques to detect phone level mispronunciations in isolated word utterances from impaired children. The results of this study are presented in Section 4.

2. Task Domain and Speech Corpora

The pronunciation verification task described in Section 1 was evaluated using a speech corpus obtained from a population of children speakers enrolled in a special education program. Isolated utterances of words taken from a vocabulary used for speech therapy were recorded using an application designed to elicit utterances from children in a multimodal game-like environment. This section describes the speaker population, the data collection scenario, and the process of labeling the speech corpus.

2.1. Speaker Population

The impaired children speakers suffer developmental disabilities of different origins and degrees that affect their language abilities, especially at the phonological level. It is believed that all speakers suffer from a neuromuscular disorder so that all of them can be characterized as having dysarthria. None of the speakers are known to be hearing impaired or suffer from any

This work was supported under NSERC Program Number 307188-2004, and supported by the national project TIN-2005-08660-C04-01 from MEC of the Spanish government

abnormality or pathology in the articulatory or phonatory organs. All the impaired children are students at the Public School for Special Education "Alborada" in Zaragoza, Spain [7]. There are fourteen impaired children speakers in the corpus, including 7 males and 7 females, ranging in age from 11 to 21 years old.

A speech corpus was also collected from a population of 168 unimpaired children ranging in age from 10 to 18 years old. This corpus is intended to allow the development of task domain systems adapted to the speech of children with the same age range as in the impaired children population.

All speech collected from both impaired and unimpaired speakers consists of utterances of isolated words taken from a vocabulary specified by the "Induced Phonological Register" (RFI) [8]. It contains a set of 57 words used for speech therapy in Spanish which are phonetically balanced and also balanced in terms of their pronunciation difficulty. The set of 57 words contains 129 syllables and 292 phonemes. Four sessions consisting of 57 utterances of each vocabulary word per session were collected from each impaired speaker and a single session was collected from each unimpaired speaker.

2.2. Data Collection Scenario

Isolated utterances were elicited from the children speakers using a multimodal computer-aided speech therapy application called "Vocaliza" [7]. All impaired speakers recorded 4 sessions which involved uttering the 57 isolated words of the RFI. Speech was recorded using a wireless close-talking microphone and audio was sampled at 16 kHz and stored in 16 bits format. The average signal to noise ratio obtained for the recordings was 26.35 dBs. The Vocaliza system provides a user interface that is designed for speech therapy sessions with children and facilitates natural human-computer interaction for children. Both Vocaliza based speaker-independent and speaker-dependent word recognition performance are presented in [7].

2.3. Labelling Mispronunciations

Manual phoneme level labelling of pronunciation errors from the impaired children was performed according to the following procedure. All phonemes in the database were labelled by three independent labellers as having been either deleted by the speaker, mispronounced and therefore substituted with another phoneme, or correctly pronounced. The final label for the phoneme was chosen by consensus among the labellers.

Analysis of the manually derived labels shows that 7.3% of the phonemes are deleted by the impaired speakers, and 10.3% of the phonemes are mispronounced. Hence, only 82.4% of the phonemes were labelled as phonetically correct. These mispronunciations affect 47.7% of the words in the database. Pairwise interlabeller agreement for the manual labelling task was 85.81%.

3. Pronunciation Verification

This section describes the approach for verifying the correctness of phoneme level pronunciations in utterances from the disabled children speakers described in Section 2. It is assumed that each speaker generates an utterance of a known word from the 57 word vocabulary described above. Pronunciation verification in this context involves verifying the claim that a given phoneme is pronounced correctly. The section has two parts. First, the procedure for obtaining phoneme level confidence scores will be presented. Second, adaptation scenarios will be presented for reducing the effects of other sources of variability. These may include all sources of variability, outside of those introduced by the speech disorders existing among the disabled speaker population, that influence the ability to detect mispronunciations that can be attributed to speech disorders.

3.1. Computing Confidence Scores

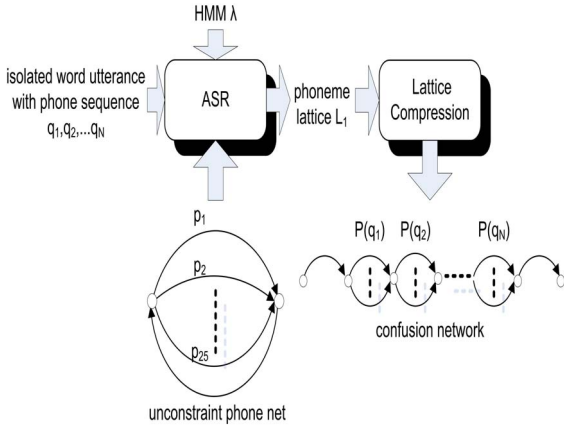

Figure 1: Confusion network based posterior probability based score

In the phoneme pronunciation verification scenario discussed in this section, it is assumed that the "target" word sequence and its baseform lexical expansion is known. Pronunciation verification in this context simply refers to obtaining confidence measures for each phoneme in the baseform expansion and applying a decision rule for accepting or rejecting the hypothesis that a given phone was correctly pronounced.

The process, as depicted in Fig.1, is performed in two steps. Assume an isolated word test utterance and the corresponding phone string q_n, $n = 1, \ldots, N$, are given. First, phoneme recognition is performed on the given isolated word utterance where search is constrained using a network that describes the potential pronunciations that might be expected from an unimpaired speaker. This network could potentially be created from the syllabification rules of the language or be trained from observed pronunciations decoded from the population of unimpaired speakers. While both the rule based and statistical constraints are currently being investigated, a simple unconstrained phone network is used in the experimental study described in this paper. A phoneme lattice L_1 which contains only phone labels and their associated acoustic probabilities is generated by the decoder.

Second, a confusion network, as depicted in Figure 1, is created using a lattice compression algorithm. This is a linear network where all arcs that eminate from a node terminate in the same node and the ordering properties of the original lattice are maintained in the confusion network. The posterior phone probabilities $P(q_n)$, $n = 1, \ldots, N$, appear on the transitions of the confusion network. These posterior phone probabilities are used as phone-dependent confidence scores. A decision criterion for verifying whether a given target phoneme has been correctly pronounced can be implemented by comparing these scores with a fixed decision threshold.

3.2. Baseline System and Adaptation Scenarios

In this section, the baseline system and the adaptation scenarios included in our experimental study described in Section 4 are introduced. The baseline ASR system is trained from the Spanish language Albayzin speech corpus [9], which includes 6,800 sentences with 63,193 words. This corpus contains 6 hours of speech including silence; however, only 700 unique sentences are contained in the corpus. Because of this lack of phonetic diversity, it is difficult to train context dependent models that will generalize across task domains. For this reason and because of the simplicity of this small vocabulary task, context independent monophone models are used here. In all experiments, 25 monophone based context independent HMMs are used which consist of 3 states per phone and 16 Gaussians per state. Mel Frequency Cepstral Coefficient (MFCC) observation vectors along with their first and second difference coefficients are used as acoustic features.

Phoneme level pronunciation verification is performed on isolated word utterances from a 57 word vocabulary were each utterance is an average of only 2.3 seconds in length including silence. In order to obtain a more robust task-dependent acoustic model, a subset of the unimpaired children corpus described in Section 2 is used to perform a maximum a posteriori (MAP) based adaptation of the Gaussian mean vectors. This adaptation corpus includes 6840 adaptation utterances spoken by 120 unimpaired children. Each unimpaired speaker provides 57 isolated test word utterances where all the words are assumed to be accurately pronounced. The adaptation corpus contains 4.5 hours of speech including silence.

Supervised speaker-dependent adaptation for each test speaker is performed using a maximum likelihood linear regression (MLLR) based transform applied to the Gaussian means of the task-dependent HMM. For each speaker, a single MLLR transform matrix is estimated and applied for speaker adaptation. The speaker-dependent MLLR adaptation data consists of 57 isolated word utterances, or 2.2 minutes of speech for each of the test speaker. The remaining 2394 impaired children utterances, three sessions of 57 isolated word utterances for each impaired speaker, are used for evaluation. The supervised speaker-dependent MLLR transformation is then applied prior to verifying the phoneme level pronunciation of the impaired children speech utterances.

4. Experimental Study

This section presents an experimental study performed to evaluate the ability of confidence scores derived from lattice posterior probabilities to detect mispronunciations in utterances obtained from disabled speakers. The pronunciation verification results are presented for a variety of adaptation scenarios.

Adaptation Scenario	EER
Task Indep. HMM (Baseline)	25.3%
Task Dep. MAP Adaptation	19.9%
Speaker Dep. MLLR Adaptation	18.6%

Table 1: Phone detection performance measured as the equal error rate (EER) for baseline, task dependenet MAP adaptation, and speaker dependent MLLR adaptation.

Table 1 summarizes the performance of a system for verifying the hypothesis that a phoneme has been mispronounced in terms of the equal error rate (EER). The EER is computed by applying a threshold to the phoneme level scores obtained from the posterior confusion network probabilities and identifying the threshold setting where the probability of false acceptance is equal to the probability of false rejection. There are 12,264 monophone test trials including 10,083 phonemes labelled as being correctly pronounced and 2,128 labeleled as incorrectly pronounced. The 2128 'incorrect' test trials correspond to phoneme instances that have been either mispronounced by the test speaker (substituted for another phoneme) or deleted altogether. Comparing the first and second rows of Table 1, there is 20% relative reduction in EER for task-dependent MAP adaptation relative to the baseline ASR system. This rather significant improvement is due largely to the significant mismatch in speaker characteristics that exists between the largely adult speaker population in the Albyzin corpus and the unimpaired children speaker population in the adaptation corpus.

Comparing the second and third rows of Table 1, there is an additional 7.5% relative reduction in EER obtained for speaker-dependent MLLR adaptation. Note that the speaker dependent adaptation data includes both correctly pronounced phonemes and phonemes that were mispronounced by the impaired speakers. This may limit the potential performance improvements that are achievable in this scenario.

These same results are presented as detection error tradeoff (DET) curves in Figure 2. Note the performance characteristics are well behaved in that the same relative performance is achieved by the three systems at all operating points.

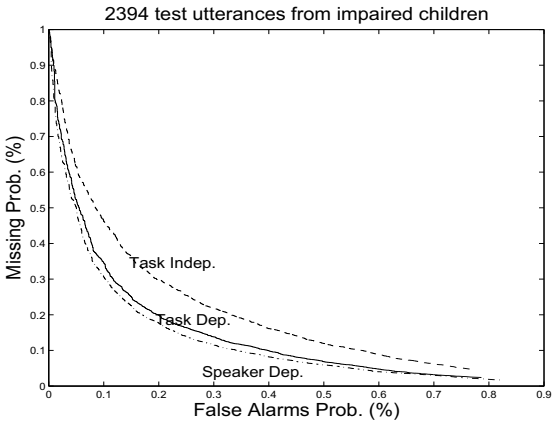

Figure 2: DET curve comparison for different Adapt. Scenario

The performance of phoneme level pronunciation verification was measured for two different subsets of the incorrectly pronounced test trials. These included instances where, first, the target phoneme was deleted by the impaired speaker and, second, where the target phoneme was mispronounced by the target speaker and substituted with another phoneme. The performance measured for these two subsets of the incorrect utterances are shown in Table 2 and Figure 3. These results show that for all conditions, verifying the hypothesis that a phoneme was deleted by the speaker is easier than verifying that a phoneme was mispronounced. There are many potential explanations for this behavior. One explanation may relate to the strategy followed by the human labellers in assigning correct and incorrect pronunciation labels to the phonemic expansions of words in the

test corpus. Rather forgiving subjective judgements were made when deciding whether a given utterance contained a "pronunciation variant" of a phoneme as opposed to a mispronunciation. This may result in many cases where the decision threshold defines a phoneme instance to be mispronounced when the reference label indicates the phoneme was correctly pronounced.

Comparison in Equal Error Rate (EER)			
Adaptation Scenario	All Error	Deletion Errors	Mispron. Errors
Task Indep. (Baseline)	25.3%	23.4%	26.8%
Task Dep. MAP Adapt.	19.9%	17.4%	21.7%
Speaker Dep. MLLR Adapt.	18.6%	16.2%	20.2%

Table 2: Phone detection performance comparison for baseline, task dependent MAP adaptation, and speaker dependent MLLR adaptation performed on different subsets of the test data. The 'All' case includes all the 12264 test trials (10083 correct, 2181 incorrect). 'Deletion Errors' includes only the 943 incorrect test trials that correspond to deleted phonemes. 'Mispronunciation Errors' includes only the 1238 incorrect test trials corresponding to mispronounced phonemes.

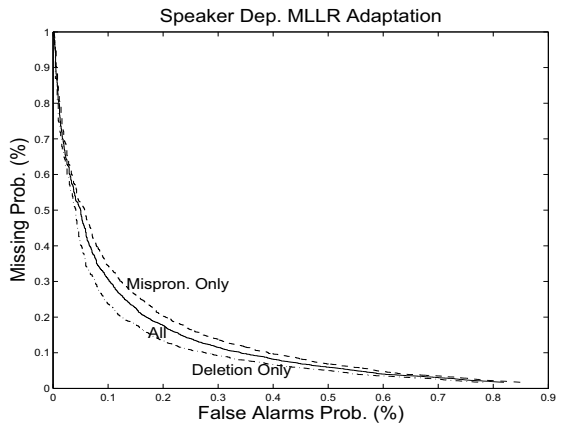

Figure 3: DET curve comparison for different Test cases

5. Summary and Conclusions

This paper addresses the problem of neuromuscular disorders affecting in speech. A phone level pronunciation verification scenario is proposed to detect the influence of mispronunciations induced by neuromuscular disorders in dysarthric speech. MAP based task domain adaptation and MLLR based speaker adaptation were applied to reduce the influence of extraneous sources of variability on verification performance. When both task-dependent MAP adaptation and speaker-dependent MLLR adaptation are applied, an improvement in EER of approximately 25% relative the performance obtained using a baseline ASR system is obtained, resulting in an EER of 18.6%. It is believed that the performance of the confidence measures used in this system achieve a performance that is close to that necessary to provide useful feedback to impaired speakers in language learning and speech therapy applications.

6. Acknowledgements

The authors would like to thank Yun Tang at McGill University for many helpful discussions and his generous help with simulations performed for this work. All HMM training and recognition simulations were based on the HTK HMM Toolkit [10]. All CNC system combination experiments were performed with the help of the SRI LM Toolkit [11].

7. References

[1] A. Hatzis, P.-D. Green, and S.-J. Howard, "Optical Logo-Therapy (OLT) : A computer-based real time visual feedback application for speech training," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Rhodes, Greece, October 1997.

[2] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M.Parker, "Automatic Speech Recognition with sparse training data for dysarthric speakers," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Geneva, Switzerland, September 2003.

[3] M.-S. Hawley, P. Green, P. Enderby, S. Cunningham, and R.-K. Moore, "Speech technology for E-inclusion of people with physical disabilities and disordered speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Lisbon, Portugal, September 2005.

[4] R. Patel, "Phonatory control in adults with cerebral palsy and severe dysarthria," *AAC Augmentative and Alternative Communication*, vol. 18, pp. 2–11, March 2002.

[5] J.-R. Deller, D. Hsu, and L.-J. Ferrier, "On the use of Hidden Markov Modelling for recognition of dysarthric speech," *Computer Methods and Programs in Biomedicine*, vol. 35, pp. 125–139, 1991.

[6] F. Zhang, C. Huang, F. K. Soong, M. Chu, and R. Wang, "Automatic mispronunciation detection for Mandarin," in *Proc. ICASSP*, Las Vegas, USA, Apr. 2008.

[7] C. Vaquero, O. Saz, E. Lleida, and W.-R. Rodrǵuez, "E-inclusion technologies for the speech handicapped," in *Proc. ICASSP*, Las Vegas, USA, Apr. 2008.

[8] M. Monfort and A. Juárez-Sánchez, "Registro fonológico inducido (tarjetas gráficas)," *Ed. Cepe, Madrid*, 1989.

[9] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. M. no, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proc. Eurospeech*, Berlin, Germany, Sept. 1993.

[10] S. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Cambridge University Engineering Department, Speech Group, Cambridge, Tech. Rep., 1993.

[11] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.