A STUDY OF PRONUNCIATION VERIFICATION IN A SPEECH THERAPY APPLICATION

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

Techniques are presented for detecting phoneme level mispronunciations in utterances obtained from a population of impaired children speakers. The intended application of these approaches is to use the resulting confidence measures to provide feedback to patients concerning the quality of pronunciations in utterances arising within interactive speech therapy sessions. The pronunciation verification scenario involves presenting utterances of known words to a phonetic decoder and generating confusion networks from the resulting phone lattices. Confidence measures are derived from the posterior probabilities obtained from the confusion networks. Phoneme level mispronunciation detection performance was significantly improved with respect to a baseline system by optimizing acoustic models and pronunciation models in the phonetic decoder and applying a non-linear mapping to the confusion network posteriors.

Index Terms— confidence measure, speech therapy

1. INTRODUCTION

The techniques developed in this paper are intended to be used as part of a semi-automated system for providing interactive speech therapy to a potentially large population of impaired individuals. While there are several areas of diagnosis and treatment for patients with speech and language disorders, the interest here is in the acquisition of the phonetic systems of a language. User interaction involves the patient receiving feedback evaluating the quality of pronunciation of words presented in a therapy session. Automatic procedures for verifying the quality of phoneme level pronunciations are proposed and evaluated in this paper.

This is part of a larger effort to evaluate the feasibility of a more efficient, lower cost methodology for diagnosis and treatment of patients with neuromuscular disorders [1]. This methodology includes, at its lowest level, an interactive dialog between the patient and an automated system for performing training, collecting speech from the student, and providing performance feedback. At the next level, a mechanism exists for a non-expert to measure the performance of the patient. This is provided through a simple, easily reproducible scheme for labeling utterances at the phonemic level according to the accuracy of pronunciation. At the highest level, the speech therapist can assimilate the evaluations obtained from the interactive sessions, review sample utterances, provide performance assessment, and prescribe additional therapy.

The phoneme level pronunciation verification (PV) techniques presented here are based on phone level measures of confidence that are derived from the acoustic speech utterance. Utterances of known words are presented to a phonetic decoder and confusion networks are generated from the resulting phone lattices. Confidence measures derived from the confusion networks are used to define a decision rule for accepting or rejecting the hypothesis that a phoneme was mispronounced. This decision can then be used to help provide the speech therapy patients with feedback concerning pronunciation quality.

It is well known, however, that dysarthria induced variation in pronunciation is one of many sources of variability in the speech utterance. Physiological and dialect variability exists among unimpaired speakers and coarticulation is a fundamental source of acoustic variability in all speaker populations. It can be difficult for a localized measure of confidence to distinguish between some preconceived notion of mispronunciation and naturally occuring variability. Section 3 describes several approaches for reducing the effects of other sources of variability on the decision rule described above. 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. SPEECH THERAPY TASK DOMAIN

2.1. Utterances of the Speech Therapy Corpus

Utterances were elicited from impaired and unimpaired children speakers from 11 to 21 years old enrolled in a special education program. The children were interacting with a multimodal computer-aided speech therapy application called “Vocaliza” [1]. The Vocaliza system provides a user interface that is designed for speech therapy sessions with children and facilitates natural human-computer interaction for children. 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) [2]. 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 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 abnormality or pathology in the articulatory or phonatory organs.

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

2.2. Pronunciation Labelling by Non-Expert Human Labellers

One very important aspect of a semi-automated system for interactive speech therapy discussed in Section 3 is a mechanism for non-experts to measure the performance and developmental progress of patients. A simple manual system for phoneme level pronunciation labelling was devised for this purpose. Phonemes in isolated word utterances produced by impaired children speakers were labelled as having been either deleted by the speaker, mispronounced and therefore substituted with another phoneme, or correctly pronounced.

This scheme was evaluated by having three independent non-expert labellers label all phonemes in the speech corpus. Pairwise inter-labeller agreement for the manual labelling task was 85.81%. This use of a less-descriptive but highly repeatable labelling system represents a trade-off that has enabled the study described in Section 3. It is important because it addresses the trade-off between the need for a consistent, repeatable, and easily implemented labeling strategy against the need for an accurate characterization of the quality of pronunciation of a given phoneme. Analysis of the manually derived labels shows that 7.3% of the phonemes were deleted by the impaired speakers, and 10.3% of the phonemes were mispronounced. These mispronunciations affect 47.7% of the words in the database. The final label assigned to a given phoneme was chosen by consensus among the labellers.

3. VERIFYING PHONEME PRONUNCIATIONS

This section describes techniques for detecting phoneme level mispronunciations in utterances from the impaired children population described in Section 2.1. There are three parts. First, a phoneme level confidence measure is defined based on posterior probabilities derived from a confusion network (CN). Second, acoustic model adaptation approaches are presented for reducing the effects of speaker and task variability on PV performance. Third, a nonlinear mapping is described that incorporates a variety of additional information to map the CN derived confidence measures into measures that can better predict the manually derived pronunciation labels.

3.1. Phoneme Level Confidence Measure

In the phoneme pronunciation verification (PV) scenario described in Section 4, it is assumed that the “target” word and its baseform lexical expansion $\mathbf{q} = q_1, \ldots, q_N$ are known. PV 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 phoneme was correctly pronounced. The process is performed in two steps. First, phonetic decoding is performed on the given isolated word utterance where search is constrained using a network that describes the pronunciations that might be expected from an unimpaired speaker.

Two simple approaches have been used to model this set of expected pronunciations. First, a bigram phonotactic model was trained from baseform phonetic expansions obtained from an 8 million word subset of the Spanish language section of the Europarl speech corpus [3]. This phonotactic bigram model was used to constrain search in phoneme recognition. Second, a network was trained from observed pronunciations decoded from approximately 9600 utterances taken from a population of unimpaired children speakers. Using a phonotactic bigram language model was found to provide the best performance of the two methods partly because of the superior size of the training corpus. As a result, only the performance of the bigram phonotactic pronunciation model is considered in Section 4.

A phone lattice containing phone labels and their associated acoustic and language probabilities is generated by an automatic speech recognizer (ASR) acting as a phonetic decoder. A confusion network is created from the phone lattice using a lattice compression algorithm. The confusion network is a linear network where all arcs that emanate from the same start node terminate in the same end node. 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.

The last step associated with obtaining a phoneme level confidence estimate involves identifying the confusion network transition that most likely corresponds to the given target phoneme from the baseform transcription. This posterior phone probability is used as the phone-dependent confidence score. This is done by obtaining the best alignment of the target baseform transcription phone string with the original phone lattice. A decision criterion for verifying whether a given target phoneme has been correctly pronounced can be implemented by comparing these scores with a decision threshold.

3.2. Reducing variability through model adaptation

Acoustic model adaptation scenarios are presented here for reducing the effects of sources of variability outside of those introduced by the speech disorders existing among the disabled speaker population. This section describes the baseline task independent acoustic model training, task dependent model adaptation, and speaker dependent model adaptation.

Baseline hidden Markov models (HMMs) are trained from the Spanish language Albayzín speech corpus [4], 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 Gaussian distributions per state. MFCC observation vectors along with their first and second difference coefficients are used as acoustic features.

Task dependent acoustic model adaptation is performed using isolated word adaptation utterances from the same vocabulary described in Section 2. The utterances are obtained from a population of 120 unimpaired children speakers resulting in a total of 6,840 utterances and 4.5 hours of speech. Combined maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) [5] based adaptation is used to adapt the means of the distributions of the baseline model listed above.

Supervised speaker-dependent adaptation for each of 14 impaired test speakers is performed using an MLLR based transform applied to the Gaussian means of the task-dependent HMM. For each speaker, a single MLLR transform matrix is estimated from 2.2 minutes of speech. The supervised speaker-dependent MLLR transformation is then applied prior to verifying the phoneme level pronunciation of the impaired children speech utterances.

Even a supervised speaker adaptation paradigm is problematic for the impaired children population since the utterances contain many phonemes that are known to be mispronounced or deleted. It is possible, however, to modify the adaptation procedure to incorporate the pronunciation labels obtained from the human labellers. This was done for MLLR adaptation to the impaired speakers by creating two regression matrices. One regression matrix was estimated from occurrences of phonemes in the adaptation data that were labelled as being correctly pronounced and another matrix was estimated from occurrences of phonemes that were labelled as being incorrectly pronounced. During recognition, only the first matrix was applied to transforming the mean vectors of all model distributions. Phonemes in the adaptation data that were labeled by the human labellers as having been deleted by the speaker were simply deleted from the reference transcription during adaptation. This procedure, referred to later as "Label Supervised MLLR", is similar in spirit to unsupervised adaptation procedures that rely on acoustic confidence measures [6]. These procedures apply varying weight to regions of an adaptation utterance to reflect the relevance of the region to the distributions being adapted. It is shown in Section 4 that significant performance improvement can be obtained by exploiting the supervision provided by the human labellers.

3.3. Non-linear Mapping of Posterior Probabilities

A nonlinear transformation is performed to map the lattice posterior probabilities to phone level confidence measures. There are two motivations for this. The first motivation stems from the fact that all of the PV techniques presented here are evaluated in terms of their ability to predict the labels defined by the labeling scheme defined in Section 2.2. The decision made by an expert as to whether a given occurrence of a phone is classified as being "mispronounced" rather than as a "pronunciation variant" will always have a subjective component. There is no guarantee that the posterior probabilities estimated as described in Section 3.1 will always be accurate predictors of these labels.

The second motivation is the fact that there is a great deal of prior information available in this PV scenario. This includes knowledge of the target word, the target phone, and the position of the phone within the word. This prior information can be combined with the phone level posterior probability using one of many possible fusion strategies to better predict the human derived labels.

In the experimental study described in Section 4, the parameters of a single layer multilayer perceptron with the above parameters as input are trained to implement a non-linear transformation. Backpropagation training is performed for a network with input activations which include the phone level posterior probabilities, indicator variables corresponding to each of the phone labels, and optional indicator variables corresponding to speaker identity. The network is trained with the human derived pronunciation labels serving as targets. PV is performed using the output activations obtained from this network on test utterances.

4. STUDY OF PV PERFORMANCE

This Section evaluates the performance of the pronunciation verification techniques presented in Section 3. For each isolated word test utterance, the task is to verify the claim that the pronunciation of phonemes in the baseform expansion of the word is correct according to the human labels assigned using the labeling scheme described in Section 2.2. This is thought to be a reasonable predictor of the performance of a system for providing feedback to patients concerning the quality of word pronunciations during an interactive therapy session.

In all of the mispronunciation detection experiments, the performance is presented using the equal error rate (EER) measures that can be obtained from the detection error trade off (DET) curves. The EER is computed by applying a threshold to the phone level confidence scores and identifying the threshold setting where the probability of false acceptance is equal to the probability of false rejection. All the results reported in this section are obtained using a test set consisting of 2,394 utterances from 14 impaired children speakers resulting in a total of 12,264 monophone test trials. These include 10,083 phonemes labeled by human labellers as being correctly pronounced and 2,128 labelled as incorrectly pronounced. The 2,128 'incorrect' test trials correspond to phoneme instances that have been either mispronounced by the test speaker (substituted for another phoneme) or deleted altogether.

Phoneme level Verification Performance (EER)		
Acoustic Model	zerogram	bigram
TIND - Baseline	25.3%	22.2%
TDEP - MAP/MLLR Adaptation	19.7%	18.4%
SDEP - MLLR Adaptation	18.3%	17.1%
SDEP - Label Supervised MLLR	17.2%	16.2%

Table 1. Phoneme detection performance measured using EER

Table 1 displays the PV performance as percent EER obtained from the confusion network derived posterior probabilities as described in Section 3.1. Results are presented using four different acoustic HMM's and two different pronunciation networks in the phonetic decoder. The column labelled "bigram" in Table 1 corresponds to the case where the bigram phonotactic network described in Section 3.1 was used for decoding. The column labelled "zerogram" corresponds to the case where an unconstrained phonotactic network was used. The first row of the table displays the performance for the baseline HMM model described in Section 3.1. While a baseline EER of 25 percent is relatively poor, it is interesting to note that the bigram network results in 12.4 percent reduction in EER relative to the unconstrained phone decoder. The bigram network results in smaller but consistent reductions in EER for all conditions.

The second row of Table 1 shows that combined MAP/MLLR task dependent (TDEP) adaptation to the unimpaired children corpus results in approximately twenty percent decrease in EER. This rather significant improvement is due largely to the significant mismatch in speaker characteristics that exists between the largely adult speaker population in the Albayzín corpus and the unimpaired children speaker population in the adaptation corpus.

The third row of Table 1 displays the EER for speaker dependent (SDEP) MLLR adaptation of the TDEP HMM models using 2.2 minutes of speech from each test speaker. This results in a decrease in EER of approximately 7 percent with respect to the TDEP performance in the second row. Note that the speaker dependent adaptation data includes both correctly pronounced phonemes and phonemes that were mispronounced by the impaired speakers. Including the mispronounced phonemes in the adaptation data may limit the potential performance improvements that are achievable in this scenario. The fourth row of Table 1 displays the result after performing SDEP adaptation using the "label supervised" MLLR adaptation described in Section 3.2. The corresponding results show that when the MLLR regression matrix is trained only from phoneme segments that have been labelled as being correctly pronounced, the relative reduction in EER increases from 7 to 12 percent with respect to the TDEP EER.

Table 2 provides a comparison between PV performance obtained from the CN derived posterior probabilities and from the NN based nonlinear mapping (NLM) described in Section 3.3. All of the results in Table 2 were obtained from a phone decoder with TDEP

Comparison of CN and NLM Confidence Scores (EER)	
Confidence Score Definition	zerogram
CN Posteriors - TDEP (MAP/MLLR)	19.7%
NLM - Context Input	18.1%
NLM - Speaker and Context Input	14.9%

Table 2. Comparison of confidence scores derived from CN posteriors and from NLM of CN posteriors

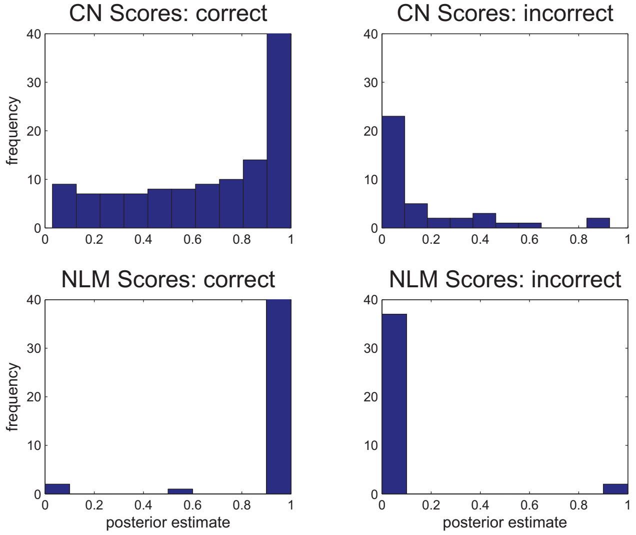

Fig. 1. Histogram plots for the phone context "sil-k+a" dependent scores obtained from the SDEP Label Supervised MLLR adaptation and the SDEP NLM, respectively. Zerogram network used in ASR.

adapted acoustic models and an unconstrained phonotactic network. The first row of the table is taken from the corresponding result in Table 1. The second and third rows of Table 2 display the EER obtained using NLM's trained from the same 2.2 minutes of speech from each test speaker that was used for training the SDEP models in Table 1. The second row of the table corresponds to the case where, in addition to the CN derived posteriors, only the context attributes were input to the NN. This corresponds to an 8% relative reduction in EER with respect to the performance obtained using CN posterior probabilities. The third row of Table 2 corresponds to the case where speaker attributes (indicator variables specifying speaker identity) were also input to the NN. This corresponds to a more substantial 23% relative reduction in EER with respect to the performance obtained using CN posterior probabilities.

To provide an illustration of how the NLM enhances the distribution of the phone level PV scores, score distributions for an example phone in context are plotted before and after the NLM in Figure 1. The histogram plots in Figure 1 were obtained from 139 correctly pronounced and 39 incorrectly pronounced samples of the phone context "sil-k+a", where "sil" represents an initial silence. The CN based confidence scores, along the top row of Figure 1, are obtained using the best SDEP acoustic model with performance given in the fourth row of Table 1. The NLM based confidence scores, along the bottom row of the figure, are obtained from the NLM that provided the best PV performance given in the third row of Table 2. For each case, the distributions of scores for examples labelled as being correctly pronounced and incorrectly pronounced are shown in the figure. It is clear that the NLM does indeed significantly reduce the overlap of the correct and incorrect distributions resulting in a better detection characteristic for the Spanish phoneme "k" in this context.

5. CONCLUSION

Simple phoneme level confidence measures based on confusion network posterior probabilities were found to provide reasonable performance in detecting mispronunciations in utterances taken from the impaired children corpus described in Section 2. However, after adapting acoustic models and performing nonlinear mapping of the CN posteriors as described in Section 3, a relative forty percent improvement in detection performance was obtained. This corresponds to an improvement from 25 percent equal error rate for the baseline system to 14.9 percent equal error rate for best system presented in Section 4. The results obtained here demonstrate that the ability to detect mispronunciations resulting from neuromuscular disorders can be significantly improved by reducing the effects of other sources of variability in speech. It is believed that 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 [7]. All CNC system combination experiments were performed with the help of the SRI LM Toolkit [8].

7. REFERENCES

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

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

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the 10th Machine Translation Summit*, Phuket,Thailand, September 2005.

[4] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. Mari no, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proceedings of the 3th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Berlin, Germany, September 1993.

[5] C.-J. Legetter and P.-C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[6] Gyucheol Jang, Sooyoung Woo, Minho Jin, and C.D. Yoo, "Improvements in speaker adaptation using weighted training," in *Proc. ICASSP*, Hong Kong, China, Apr. 2003.

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

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