VERIFYING SESSION LEVEL PRONUNCIATION ACCURACY IN A SPEECH THERAPY APPLICATION

Shou-Chun Yin¹, Richard Rose¹, Yun Tang²

¹Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
²Nuance Communications Inc., Montreal, Canada

sssl23ca@yahoo.com, rose@ece.mcgill.ca, hottang@gmail.com

ABSTRACT

This paper investigates the problem of verifying the pronunciations of phonemes from continuous utterances collected from impaired children speakers engaged in a speech therapy session. A new pronunciation verification (PV) approach based on the subspace Gaussian mixture model (SGMM) is presented. A single SGMM is trained from test utterances collected from impaired and unimpaired speakers. PV measures are derived from model space based distances estimated from impaired and unimpaired speaker dependent SGMM parameters. The PV performance is presented in terms of the method’s ability to detect and report phone level pronunciation errors for an entire speech therapy session and for individual phoneme occurrences. Comparisons are made with respect to an approach using phone level confidence measures estimated from lattice posterior probabilities.

Index Terms— confidence measure, speech therapy

1. INTRODUCTION

The work described in this paper is part of a project involving the detection of phone level mispronunciations in utterances obtained from disabled children in a speech therapy task domain [4]. The pronunciation verification (PV) approach presented here addresses two possible applications. The first is to provide a single measure which might be used by speech therapists to determine the level of progress made by subjects across speech therapy sessions. This application is similar to that addressed in [1] where the quality of pronunciation was assessed in speech therapy sessions. This application is similar to that addressed in [1] where the quality of pronunciation was assessed for continuous utterances from impaired children speakers. The second application is to provide a measure that might be used to provide feedback to the subject concerning the quality of pronunciation of individual utterances.

The new approach is based on phone specific parameters associated with the subspace Gaussian mixture model (SGMM) [8, 7, 6]. The SGMM is based on forming state specific projections within multiple linear subspaces. This differs from the conventional continuous density hidden Markov model (CDHMM) whose observation densities are generally formed from state dependent Gaussians. In the SGMM, the state specific parameters are referred to as state projection vectors, which have been shown to correlate well with vowel classes of a language [7]. Based on this observation, direct comparison of the distance between two state projection vectors within the same subspace has been found to be a reasonable approach for comparing the pronunciation of phonemes associated with these state projections. The idea of comparing two state projection vectors for making a PV decision is similar to the comparison between i-vectors in a factor analysis based speaker verification system [2].

After introducing the task domain in Section 2, a brief review of the SGMM model and the application of the SGMM to pronunciation verification (PV) is presented in Section 3. Section 4 presents a PV approach based on a distance measure estimated between the state-level SGMM subspace projections. A scenario of applying linear discriminant analysis (LDA) to the state-level subspace projections is also presented. Finally, the session level and utterance level experimental studies are presented in Section 5.

2. SPEECH THERAPY TASK DOMAIN

This section gives a brief introduction to the speech therapy task domain. First, the speech corpus involved in this work is described. Then, the pronunciation labelling strategy is summarized.

The impaired and unimpaired children speech utterances used for this work are taken from the speech corpus used for the study presented in [4]. The impaired children speakers suffer from developmental disabilities of different origins and degrees that affect their language abilities, especially at the phonological level. All impaired and unimpaired children speech consists of utterances of isolated words taken from a vocabulary specified by the “Induced Phonological Register” (RFI) [5]. It contains a set of 57 words used for speech therapy in Spanish which is phonetically balanced and also balanced in terms of their pronunciation difficulty.

There are fourteen impaired children speakers in the impaired speech corpus, including 7 males and 7 females, ranging in age from 11 to 21 years old. Another unimpaired speech corpus was also collected from a population of 168 unimpaired children ranging in age from 10 to 18 years old. Four sets of recordings consisting of the 57 word utterances of each RFI word per set were collected from each impaired speaker and a single set was collected from each unimpaired speaker.

A simple manual system for phoneme level pronunciation labelling was devised for non-experts to measure the performance and developmental progress of patients. 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. There are a total of 16352 phoneme instances provided from the impaired children speech corpus. This scheme was evaluated by having three independent non-expert labellers label all phonemes in the speech corpus. Pairwise interlabeller agreement for the manual labelling task was 85.81%.  

3. SUBSPACE GAUSSIAN MIXTURE MODEL

This section provides a brief introduction to both the subspace based model structure and model parameterization. Section 3.1 describes the SGMM as state specific projections within multiple...
3.1. SGMM Model Description

In the conventional continuous density hidden Markov model (CDHMM), for a given observation frame $x_t$ at time $t$, the observation density $P(x_t|s_t = j, \lambda_j)$ for a given state $j$ is formed from a mixture of state-dependent diagonal covariance Gaussians, $\lambda_j$. In the SGMM, the state-dependent diagonal covariance Gaussians become state-independent full covariance Gaussians. The corresponding observation densities of the $F$-dimensional features, $x_t$, for HMM state, $j = 1, \ldots, J$, $p(x_t|s_t = j)$, are formed from a set of state-dependent projection vectors, $v_j$, $w_i$, and a set of $I$ shared full covariance Gaussians $N(x_t; \mu_i, \Sigma_i)$. The shared $I$ full covariance Gaussians $N(x_t; \mu_i, \Sigma_i)$ yield a global Gaussian mixture model (GMM). This global GMM is similar to the universal background model (UBM) used in text-independent speaker verification. It is estimated from multiple speakers appearing in the speech training corpus.

In the simplest case, the SGMM state observation densities are described as follows,

$$ P(x_t|s_t = j) = \sum_{i=1}^{I} w_{ji} N(x_t; \mu_{ji}, \Sigma_i). \tag{1} $$

The state-dependent mean vector, $\mu_{ji}$, for state $j$ is a projection into the $i$-th subspace defined by linear subspace projection matrix $M_i$,

$$ \mu_{ji} = M_i v_j. \tag{2} $$

The term $v_j$ is the projection vector associated with the state $j$. The global mean projection matrices $M_i$, in Equation 2 are of dimension $F \times S$ where $S$ is the dimension of the subspace associated with the projection vectors $v_j$. The state specific weights in Equation 1 are given by

$$ w_{ji} = \frac{\exp w_i^T v_j}{\sum_{k=1}^{I} \exp w_k^T v_j}. \tag{3} $$

The state specific weights $w_{ji}$ are obtained from the state projection vector, $v_j$, using a log-linear model, as shown in Equation 3. The state specific weights computed from Equation 3 provides a convex auxiliary function, which facilitates the optimization of the SGMM parameters during the EM training iterations [8].

3.2. Application of SGMM to Pronunciation Verification

One important advantage of the SGMM is the fact that there is a very small number of parameters per state. The parameters, $\{M_i, \Sigma_i, w_i\}$, $i = 1, \ldots, I$, are shared across all states, and only the state projection vectors, $v_j$, are associated with each given state, $j$. This provides the potential for modeling phones associated with individual states using only a relatively small number of occurrences of that phone.

The SGMM configuration in this paper is shown in Figure 1. The shared parameters in the model are trained from the entire population of training speakers. However, a separate set of states is allocated for each speaker. Note in Figure 1, speech from both unimpaired and impaired children speakers is used for training this SGMM model.

The dimension, $F$, of feature vectors is 39, which includes 12 MFCC coefficients, normalized energy, and their first and second difference coefficients. The subspace dimension, $S$, is also set to 39. The SGMM structure is inherited from an initial CDHMM and speech UBM. Both CDHMM and speech GMM-UBM are trained using the unimpaired speech corpus described in Section 2. The CDHMM states are part of three state unclustered context-dependent triphones obtained from the 57 words from the RFI. The total number of context-dependent triphone units is 245. The speech GMM-UBM consists of 256 full covariance Gaussians.

There are several speaker-dependent sets of SGMM state projection vectors. Each set contains 735 state projection vectors, and is trained using the speech corpus for one unimpaired or impaired individual speaker. The 735 state projection vectors in each set are associated with the 245 three state context-dependent triphones. The state index, $j$, varies from 1 to 735. A subset of the unimpaired children speech corpus, which consists of 48 unimpaired children, is used for training 48 sets of unimpaired speaker-dependent state projection vectors, $\{v_j^{\text{un}}\}, k = 1, \ldots, 48$. Each of the impaired children, $s_{ik}$, $k = 1, \ldots, 14$, from the impaired children speech corpus is used for training one set of impaired speaker-dependent state projection vectors, $\{v_j^{\text{im}}\}$. This gives 14 sets of impaired speaker-dependent state projection vectors.

The motivation for configuring the model in Figure 1 is to define a measure of phonetic variation directly in the state projection vector space. State projection vectors associated with the states of individual phones from multiple unimpaired speakers have been found to form well behaved clusters. Section 4 defines a PV distance measurement that exploits this behavior.

4. A STATE PROJECTION BASED PV SCENARIO

Pronunciation verification (PV) 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. First, Section 4.1 describes a PV scenario based on the SGMM model structure. Second, Section 4.2 proposes a cosine distance approach to obtain the SGMM based PV scores. Third, Section 4.3 discusses the use of linear discriminant analysis (LDA) for suppressing inter-speaker variation in the phoneme-level PV scenario.

4.1. SGMM based PV Scenario

It is assumed that the pronunciation of a context-dependent phoneme, $p$, is characterized by the state projection vector associated with its center state, $j$. Given the state index, $j$, as the center state of the phoneme $p$, the PV decision of rejecting or accepting that phoneme $p$ has been correctly pronounced by an impaired speaker $s_{ik}$ can be made based on the distance between two state projection vectors within the same SGMM. One state projection vector $v_j^{\text{im}}$ is trained from an impaired speaker $s_{ik}$. A reference state projection vector $v_j^{\text{un}}$ is obtained using ut-
where \( v_{i}^{ref} \) is obtained by clustering state projection vectors, \( v_{i}^{u_{k}} \), \( k = 1, \ldots, 48 \), where \( v_{i}^{u_{k}} \) is the state projection vector for the \( k \)-th unimpaired speaker.

4.2. Cosine Distance based PV Scores

The cosine distance was investigated for making the phoneme level PV decision. The cosine distance between two state projection vectors is given by

\[
D(v_{j}^{ref}, v_{j}^{si}) = \frac{v_{j}^{ref} \cdot v_{j}^{si}}{|v_{j}^{ref}||v_{j}^{si}|},
\]

where \( |v_{j}| \) indicates the magnitude of the vector \( v_{j} \).

One can also assume that the pronunciation of a given phoneme \( p \) is not just characterized by its center state projection vector, but is instead characterized by the three state projection vectors associated with states \( \{j-1, j, j+1\} \). A state projection supervector associated with the phoneme \( p \) with dimension of \( 3S \), where \( S = 39 \), can be constructed by concatenating the three state projection vectors as follows,

\[
V_{p} = (v_{j-1}, v_{j}, v_{j+1}).
\]

The cosine distance in Equation 4 can then be computed in the supervisor domain, \( D(V_{p}^{ref}, V_{p}^{si}) \). The reference vector, \( V_{p}^{ref} \), in the supervisor domain is given by clustering the 48 unimpaired speaker specific state projection supervectors, \( V_{p}^{u_{k}} \). This is identical to the clustering procedure for obtaining \( v_{i}^{ref} \) given in Section 4.1.

4.3. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a well known technique in pattern recognition and machine learning [9]. The goal in this work is to apply LDA to reduce the impact of speaker variability in the PV task by performing a dimensionality reducing linear transformation on the super vector \( V_{p} \) in Equation 5. The new state projection supervector \( V_{p} \) for a given phoneme, \( p \), can be obtained by applying the LDA transform as follows,

\[
V_{p} = L^{T}V_{p},
\]

where the LDA transform \( L \) is of dimension \( 3S \times S' \) and \( S' \leq 3S \). There is a single LDA transformation estimated using unimpaired speaker-specific state projection supervectors. It will be shown in Section 5 that this process yields LDA transformed supervectors that provide better phone class discrimination and improve PV performance.

5. EXPERIMENTAL STUDY

The PV evaluation shown in this experimental study is based on the impaired children’s speech corpus. First, the baseline system is described in Section 5.1. Second, the session level PV evaluation is presented in Section 5.2. This is presented as an average equal error rate (EER) which describes detection performance across an entire session. Third, the utterance level PV evaluation is presented in Section 5.3. This provides a measure of performance for verifying the occurrence of individual instances of phoneme level mispronunciation.

5.1. Baseline System

The baseline PV scenario is based on the confusion network (CN) based confidence scores and has been presented in [4].

Simply stated, given a well trained monophone based CDHMM, the phonetic decoding can be performed on the given testing isolated word utterance. A phone lattice containing phone labels and their associated acoustic probabilities is then generated through the phonetic decoder. Finally, a confusion network is created from the phone lattice using a lattice compression algorithm. The posterior phone probability corresponding to the given target phoneme from the baseform expansion of the given testing word will appear on the arcs of the confusion network.

This posterior phone probability is used as the utterance specific phoneme based confidence score. For the baseline system presented in this work, the acoustic CDHMM training is initialized using Albayzin corpus [3]. Then it is adapted to the unimpaired children speech corpus using both maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) adaptation. An unconstrained phonotactic network is used for decoding.

5.2. Session Level PV Results

The session level PV performance is reported as the average mispronunciation detection performance across all context-dependent phonemes in the test corpus. There are four utterances of each phoneme in an impaired speaker’s session. If a majority of the individual instances of a phoneme in a speaker’s session are labeled as being correctly pronounced, then that phoneme is labeled as correctly pronounced for the session. There are 245 context-dependent phonemes from the 57 RFI words and provided by each of 14 impaired speakers. This gives total of 3430 session level test trials. This involves 2756 test trials labeled as ‘correctly pronounced’, 543 test trials labeled as ‘incorrectly pronounced’, and 131 test trials which are excluded from the evaluation.

For the baseline PV scenario, the overall session level CN confidence score for a phoneme \( p \) is obtained by averaging all the CN based confidence scores corresponding to \( p \), among all the four instances of that phoneme. This gives an equal error rate (EER) of 15.83%.

In Section 4.2, a cosine distance measure between two state projection vectors sharing the same linear subspaces in a SGMM is defined. This provides a model based approach for measuring the similarity between phoneme models trained from an impaired speaker and a cluster of unimpaired speakers. It gives a session level EER of 21.73%. When the cosine distance is measured between two state projection supervectors, as described in Equation 5, the EER reduces to 19.85%. If LDA transformed supervectors described in Equation 6 are used for computing the cosine distances, the EER can be further reduced to 18.44%. These results can be shown in Figure 2 with a fusion weight equal to zero.

It is reasonable to assume that there would be some advantage to combining the scores from the two systems. The SGMM cosine distances can provide context information which is potentially complementary to the baseline CN scores, which are obtained from a context-independent phonetic decoder. A simple session level combined score, \( S_{ui}(p, si_{k}) \), for a given impaired speaker \( si_{k} \) and context-dependent phoneme \( p \) can be expressed as follows,

\[
S_{ui}(p, si_{k}) = \alpha S_{CN}(p, si_{k}) + (1 - \alpha)D(v_{j}^{ref}, v_{j}^{si_{k}}).
\]

The first additive term in Equation 7, \( S_{CN}(p, si_{k}) \), represents the overall session level CN posterior score obtained from the baseline system. The second additive term is the SGMM based cosine distance measurement between two state projection vectors, as defined in Equation 4. The state index \( j \) represents the center state of the phoneme \( p \). This cosine distance can also be computed between two state projection supervectors, as defined in Equation 6. The fusion weight is controlled by the factor \( \alpha \), varying from zero to one. The session
level PV performance from the combined scores computed from Equation 7 is shown in Figure 2. These results show that the CN baseline PV performance can be improved by incorporating any of three kinds of SGMM cosine distance scores, if a proper fusion weight is selected. Comparing the best results from the combined scores with the baseline results, the EER drops from 15.83% to 13.40%.

Fig. 2. EER comparison, session level PV evaluation

5.3. Utterance Level PV Results

The utterance level PV is evaluated directly based on the transcription provided by human labellers, as described in Section 2. There are 16352 phoneme level test trials, which involve 13472 ‘correctly pronounced’ trials and 2880 ‘incorrectly pronounced’ trials. The ‘incorrectly pronounced’ trials include all the ‘mispronounced’ and ‘deleted’ trials.

The baseline CN confidence scores provides an EER of 19.86%, which is shown in Figure 3 with a fusion weight equal to one. This result is slightly different from the result reported in the early paper [4]. It is because the evaluation shown in this experimental study involves four instead of three recordings for each of 57 RFI words for each of the impaired children.

For a given impaired test speaker, SGMM state projection vectors have been trained using all occurrences of each phone from the test speaker utterances. So the model based distance given in Equation 4 is a score that implicitly incorporates all of these occurrences. However, it is still believed that the context information captured by SGMM cosine distances can be complementary to the utterance level CN posterior scores. The combined score \( S_{ud}(p, si_k, u) \) for a given impaired speaker \( si_k \) and phoneme \( p \) in the baseform expansion of the word in the utterance \( u \) is proposed as follows,

\[
S_{ud}(p, si_k, u) = \alpha S^{CN}(p, si_k, u) + (1 - \alpha)D(v^{ref}_p, v^{si_k}_p).
\]

(8)

The first additive term in Equation 8, \( S^{CN}(p, si_k, u) \), represents the utterance specific CN posterior score obtained from the baseline system. The second additive term is the SGMM based cosine distance measurement between two utterance-independent state projection vectors, which has exactly the same form as the second additive term in Equation 7. Similarly, the cosine distance involved in the second additive term can also be computed between two state projection supervectors, or between two LDA transformed state projection supervectors.

The fusion weight is controlled by the factor \( \alpha \), varying from zero to one. The utterance level PV performance obtained from the combined scores given by Equation 8 is shown in Figure 3. These results show that, even in the utterance level PV task, the context information captured by the SGMM cosine distances can still be useful for improving the CN baseline performance. Comparing the best results from the combined scores with the baseline results, there is a reduction in EER from 19.86% to 17.29%.

6. CONCLUSIONS

A SGMM based measure of performance accuracy has been presented and evaluated on a pronunciation verification task for impaired children speakers. It was that, when combined with a lattice based method for deriving phone level confidence measures, a PV EER of as low as 13.4% was obtained. The best performance was obtained by forming supervectors by concatenating the SGMM state projection vectors and performing discriminative dimensionality reduction in this space. These performance improvements are believed to result from an efficient characterization of context information for each phoneme by SGMM parameterization.

7. REFERENCES