LANGUAGE LEARNING BASED ON NON-NATIVE SPEECH RECOGNITION

Silke Witt Steve Young
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ
United Kingdom
Email: {smw24,sjy}@eng.cam.ac.uk

ABSTRACT

This work presents methods of assessing non-native speech to aid computer-assisted pronunciation teaching. These methods are based on automatic speech recognition (ASR) techniques using Hidden Markov Models. Confidence scores at the phoneme level are calculated to provide detailed information about the pronunciation quality of a foreign language student. Experimental results are given based on both artificial data and a database of non-native speech, the latter being recorded specifically for this purpose. The presented results demonstrate the metrics' capability to locate and assess mispronunciations at the phoneme level.

1. INTRODUCTION

With the rapid progress of speech recognition, current research challenges are to improve non-standard speech recognition such as for dialects or non-native speech. This project applies non-native speech recognition to pronunciation teaching. So far computer-assisted language teaching was restricted to "record and play", because without speech processing and recognition technology no analysis of a student's speech was possible. Current commercially available systems are beginning to include ASR but are limited to calculating a score for an utterance or a single word and using this score for acceptance or rejection of the utterance. In contrast to this, the aim here is to be able to analyse a student's speech on a sub word level and to give feedback about the pronunciation mistakes the student has made.

1.1. State-of-the-art

Over the last decade several research groups have started to develop interactive language teaching systems which incorporate pronunciation teaching based on speech recognition techniques. The "SPELL" project from Hiller et al.,[6], concentrated on teaching pronunciation of individual words or short phrases. Another early approach, based on dynamic programming and vector quantisation, by Hamada et al., [5], was likewise limited to word level comparisons between recordings of native and non-native utterances of a word. SRI have reported a technique for scoring complete sentences but not smaller units of speech, (see [8] and [3],) and their main purpose was automatic pronunciation assessment rather than teaching. As in this project, the SRI approach is text-independent, hence, when changing the vocabulary of the language learning task, no retraining of models is required. The system used by Rogers et al. [9], was originally designed to improve the intelligibility of hearing or speech impaired people and also evaluates isolated word pronunciations only. Finally, the system described by Eskenazi, [2], uses the unnormalised log-likelihoods produced by a speaker independent recogniser in forced alignment mode.

In contrast to existing work, this paper focuses on the problem of producing pronunciation scores for each phone in a sentence. Two methods of calculating such scores will be presented and evaluated using a set of artificial data and a non-native speech corpus. The approach presented here is based on the following assumptions:

- The system will be text-independent so that no new training data is required for new teaching material.
- Each student response is known, either because the task consists of reading or because a dialog is designed so that the answer can be precisely determined.
- The target is to teach one single type of pronunciation, thus in the case of teaching British English, a native speaking US American English would be corrected as well.

2. GOODNESS OF PRONUNCIATION (GOP) SCORE

The goal is to have a score of correctness or confidence for each phone of a desired transcription. This can be interpreted as the posterior probability that a speaker uttered phone p given the acoustics, O and the set of all phones, Q:

$$GOP(p) \equiv P(p|O) = \frac{P(O|p)P(p)}{\sum_{q\in Q} P(O|q)P(q)} \tag{1}$$

2.1. Method 1: First Approximation

Assuming all phones are equally likely ($P(p) = P(q)$) and that the sum in the denominator can be approximated by its maximum, the above equation rewrites as:

$$GOP_1(p) = \frac{P(O|p)}{\max_{q\in Q} P(O|q)} \tag{2}$$

The numerator is found by using forced alignment (FA), the denominator by using an unconstrained phone loop (PL). Operating on a log scale and normalising the likelihoods by frame lengths yields:

$$GOP_1(p) = S_{FA} - S_{PL} \tag{3}$$

with:

$$S_{FA} = \frac{phone\ exit\ log\ prob - phone\ entry\ log\ prob}{NF} \tag{4}$$

$$S_{PL} = \frac{\sum_{i=1}^{m} \left(S_{(PL,i)} NF_{(PL,i)}\right)}{NF} \tag{5}$$

i.e. S_{PL} is calculated as the weighted sum of all phone segments occuring during the segment which is described by the number of frames it covers, NF. $S_{(PL,i)}$ represents the maximum likelihood of the ith phoneme model in the desired speech segment.

2.2. Method 2: Including a Phoneme Language Model

In order to extend the above measure to include information about the likelihood of phone sequences a second GOP score is given by:

$$GOP_2(p) = \frac{P(O|p)P(p)}{\max_{q \in Q} P(O|q)P(q)}$$

where bigram statistics have been used for the phone level language model of the target language, $P(p)$.

These GOP scores can be used in several ways. Firstly, they serve to assess pronunciation quality. Secondly, they may help to correct mispronunciations in that a poor score indicates an error and the most likely phone determined by the phone loop indicates the error type. Finally, the application of these scores is not limited to pronunciation teaching since using them as confidence scores can also provide statistics on the occurrence of poor scores. These can be used to determine systematic pronunciation errors for a particular speaker.

2.3. Speaker Adaptation

The quality of the GOP scoring depends on the quality of the models used. Therefore, in order to improve baseline recognition performance we used speaker adaptation to adapt to the specific spectral characteristics of a speaker. Non-native speech is characterised by different formant structures compared to those from a native speaker for the same phonemes; for a more detailed description of these characteristics see for instance [1]. However, since the aim is to detect pronunciation errors, adaptation to these errors has to be avoided. Given this reasoning, we applied the Maximum Likelihood Linear Regression (MLLR) algorithm, see [7], to adapt the HMM mixture means with one global transform, in attempt to provide speaker normalisation without adapting to specific phoneme error patterns.

2.4. Duration Modelling

Non-native speech is further characterised by a considerably slower speaking rate. This causes a large number of insertions in the outputs of the phoneme loop S_{PL}. A simple ad-hoc solution to this problem is to increase the insertion penalty at the transition from one phoneme model to another. This has been done in the evaluation shown in the following sections. Future work will include more detailed duration modelling, such as adapting the duration of the HMMs.

3. DATABASES USED FOR EVALUATION

The evaluation of the GOP scoring method uses two different sets of data. Firstly, we tested the algorithm's performance on a set of artificial data. Secondly, we tested it on non-native speech data manually annotated by phoneticians.

3.1. Artificial Data

In order to have a baseline measurement where both the locations of pronunciation errors are known and the effects of bad modelling of non-native speech can be ignored, a set of artificial data was generated. The speech data comprised the DARPA Resource Management task containing utterances spoken by North Americans. For this data set a pronunciation dictionary based on a dictionary from Carnegie Mellon University using the ARPABET set of 48 phonemes was manipulated such that the pronunciations were changed to contain different phonemes. For instance, all occurrences of the vowel [ih] where changed to [ow]. Thus, from the dictionary's point of view, speech by a native speaker appears to contain errors.

Model	SA in % for $FA = 8\%$
MFCC-E-D-A Monophones	90
MFCC-E-D-A Triphones	83
MFCC	80
MFCC-D-A	88

Table 1. Performance of artificial data for different model sets

3.2. Non-native Database

The second data set consists of a newly collected database of non-native speech in which phone errors have been manually annotated by linguists. The recording guidelines for this database collection were based on the procedures used for the WJSCAM0 corpus [4]. Students of English as a second language read prompting texts composed of a limited vocabulary of 1500 words. Each prompting session consisted of 40 phonetically balanced sentences read by each subject and an additional 80 sentences which vary for each prompting text, thus providing a larger range of training data. There were 17 students, of which 12 were female and 5 were male. Their mother-tongues included Latin-American Spanish, Italian, Japanese, Korean and Chinese.

With the help of an interactive interface, this data was phonetically labelled and assessed by trained phoneticians, all of them native speakers. Firstly, each sentence was scored on a scale of 1-4, then each word was scored on the same range. Secondly, a transcription based on a standard southern English pronunciation dictionary was annotated with the substitution, deletion and insertion errors made by the subject.

The target pronunciation is Standard Southern British English, therefore for this data the British English Example Pronunciation (BEEP) Dictionary, see [4] has been used.

4. EXPERIMENTAL RESULTS

Given these two data sets we have studied two different methods of evaluating the performance of the GOP scoring. Firstly, we analysed the individual scores of each phoneme in a sentence, secondly, we looked at averaged scores per phoneme for each subject. The latter allows a general assessment of the main strengths and weaknesses of a student, since those phones which are consistently incorrectly pronounced over a test set, will have a poor average score and vice versa. The results for both of these metrics were composed for each of these two proposed GOP scores, i.e. GOP_1 and GOP_2. The recognition passes for all results were generated using the HTK Toolkit [10].

4.1. Individual Scores

4.1.1. Individual Scores for Artificial Data

Firstly, the artificial data test was used to test the scoring of individual phonemes in an utterance. The output of the scoring procedure consists of one score per phoneme, higher scores representing worse pronunciation. Given these scores, a decision threshold was used to decide whether the GOP of a phone is good enough to accept as correct. The choice of this threshold can be based on a receiver operation curve. The scoring accuracy (SA) — defined as the sum of those phone instances which were correctly accepted and rejected — was plotted versus the rate of false acceptances (FA), i.e. the number of phones incorrectly accepted as correct.

Scoring accuracy was measured over false acceptance rate for several sets of speaker independent hidden Markov models in order to find an optimal model set.

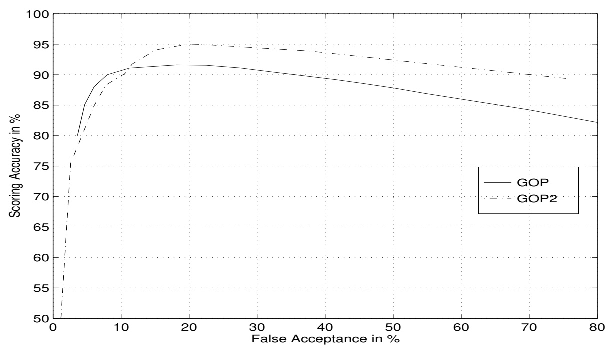

Figure 1. Contrasting the performance of GOP_1 and GOP_2 for artificial data

As can be seen from Table 1, multiple mixture component monophones using mel cepstral frequency coefficients (MFCC), with first order, (D), and second order (A) derivatives, plus energy values (E) yield the best performance.

The relative performance of the two GOP measures is shown in Figure 1. As can be seen, using a language model worsens the performance for low rates of false acceptances, but otherwise improves it.

4.1.2. Individual Scores for Non-native Speech

Unlike the artificial data set, the measurement of SA vs FA is not possible for non-native data because the exact error locations are not known. Also the labelling of the phoneticians is subjective and error prone. An alternative method of measuring the performance of the GOP scoring on non-native data is to compute the overlap (i.e. correspondence) between the rejection marked transcriptions of the human judge and the aligned GOP scored transcriptions for various thresholds. The alignment was computed using dynamic programming in order to be able to handle insertions and deletions. The inter-judge alignment was 81% and this can be regarded as an upper bound as to how high an agreement between the human scores and the computer scores can be expected. This overlap measure will be referred to as "Total Overlap" (TO), because with this measure the combined performance of rejections and acceptance is measured, in contrast to the two following measures.

Additionally, in order to measure the overlap of those phonemes which were rejected by both the judge and the GOP score for a given threshold, the phonetic transcriptions were converted into strings of symbols of "accept" and "reject". Then all corresponding rejections between these two strings were counted and normalised by the total number of rejections (R) in both strings, yielding the error overlap measure (O).

This procedure is complicated by the existence of a number of instances where a "reject" is shifted by one phoneme, i.e. in the GOP scoring a vowel was rejected, whereas in the human scoring the preceding consonant had been marked. However in both cases the same phoneme segment was indicated to be badly pronounced. To take this effect into account, we also measured the rejection string overlap using a window of 3 phonemes. This measure was called "soft overlap" (SO).

These three different ways of aligning the results yield three parameters permitting performance analysis. The overlap between the rejection sets (O and SO) indicates whether the agreement between pronunciation error judgements is low or high. The number of marked rejections (R), indicates whether a comparable amount of errors were rejected (For instance, choosing a low threshold yields higher rejection overlaps but also many rejections.). Finally, the total overlap between the rejection marked transcriptions indicates the overall agreement of judgements for rejected and accepted phonemes.

T	J	NAO	R	O	SO	TO
none	J1 J2	0.32	188	0.32	0.36	0.81
0.5	J1	0.17	557	0.17	0.32	0.38
2.0	J1	0.20	389	0.21	0.33	0.57
5.5	J1	0.22	203	0.25	0.33	0.78
8.0	J1	0.20	156	0.19	0.28	0.82
0.5	J2	0.31	564	0.32	0.58	0.48
2.0	J2	0.36	410	0.38	0.59	0.63
5.5	J2	0.31	261	0.32	0.40	0.74
8.0	J2	0.23	229	0.19	0.25	0.74

Table 2. Overlap between GOP scoring and judge: T=threshold, J=Judge R=number of rejections, O=Overlap of rejections, NAO= Overlap with No Adaptation, SO=Soft Overlap, TO= Total Overlap

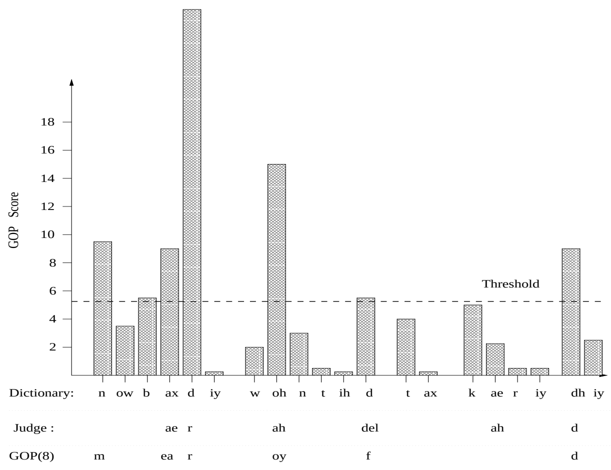

Figure 2. Scoring example for a single sentence

Table 2 shows these measures for different thresholds (T) and judges (J1 and J2). In order to demonstrate the (albeit slight) improvement gained by using speaker adaptation the column "NAO" shows the rejection overlap based on speaker independent models alone. The results for GOP_2 were very similar. Given the rejection overlap between the judges of 0.33 for 188 rejections on either side, a GOP threshold should be chosen with a similar amount of rejections, which suggests a threshold of 5.5. For this the comparison with judge 1 yields 203 rejections, an error overlap of 0.25, a soft error overlap of 0.33 and a total overlap of 0.78. These results are close to the human performance which shows that using the computer-based method scoring yields meaningful output.

An example for the individual scores can be seen in Figure 2. High scores indicates a bad pronunciation, therefore all boxes above the threshold indicate those phonemes whose pronunciation would be rejected. The transcriptions along the x-axis are the target transcription according to the pronunciation dictionary. In the line below are the phonemes which have been corrected by the human judge. The bottom line contains those phonemes which were rejected by the GOP scoring at the given threshold. It can be seen that there exists a good correspondence between the corrections of the human judge and the GOP_1 scores.

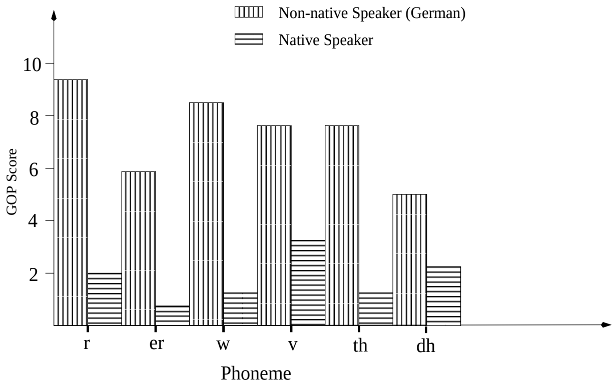

Figure 3. Difference in global GOP scores between native and non-native (German) speaker for typical error sounds

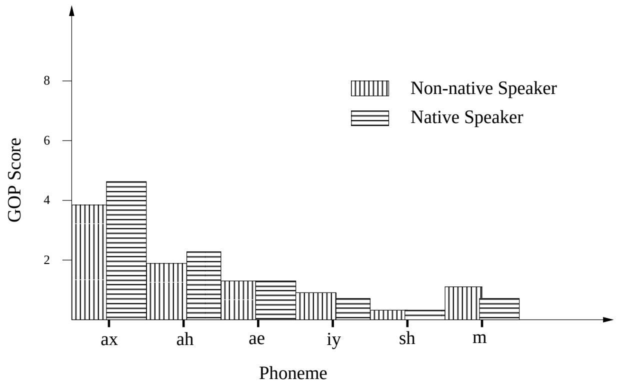

Figure 4. Similarity in global GOP scores between native and non-native (German) speaker for common sounds in both languages

So far the discussion of error detection has been restricted to substitutions. However, the mistakes of a language student also include insertions and deletions. If a deletion occurs then the forced alignment recognition pass generally is characterised by a very short duration of the respective phone and the phone itself or the neighbouring phone have a poor GOP score. Insertions are more difficult to handle, because the algorithm as presented here is not capable of detecting them, finding solutions for this problem will be part of future work.

4.2. Assessing overall pronunciation quality

Whereas the previous evaluation metric applies a score for each individual phone in sequence, the second metric is based on averaging all scores for each phone to allow a general assessment of the main strengths and weaknesses of a student. The assumption here is that phones which are consistently pronounced incorrectly over a test set, will have a poor average score and vice versa.

In order to have an assessment measure of the overall pronunciation quality for a sample set of sentences, we applied the global GOP scoring to both a non-native speaker and to a native speaker. The mother tongue of the non-native speaker was German and typical error sounds for German speakers are for instance [r], [w], [v] , [th] and [dh]. This corresponds with the results in Figure 3 where the average score for the non-native speaker is much poorer than that for the native speaker. As a contrast, see Figure 4, where the average score for sounds is shown which are common to both language and thus not difficult for the language student.

5. CONCLUSIONS

The results obtained to date suggest that confidence based GOP scoring methods can be effective in detecting mispronunciations in non-native speech. Unlike previous work on pronunciation teaching based on speech recognition, the methods studied here are focused at the phoneme level. The results show that the method can be used on both individual analysis level and on assessing overall pronunciation quality. When incorporated in a teaching system, this will allow more detailed advice to be given to the students on pronunciation errors. Future work will incorporate more knowledge about the student's native language and the transition between source and target language, such as modelling different degrees of fluency with differently adapted model sets.

6. ACKNOWLEDGEMENTS

Silke Witt is funded by an EPSRC advanced studentship and a Marie Curie Research Fellowship of the European Union.

REFERENCES

[1] L. Arslan and J.H.L. Hansen. Frequency Characteristics of Foreign Accented Speech. In *ICASSP '97*, München, Germany, April 1997.

[2] M. Eskenazi. Detection of foreign speakers' pronunciation errors for second language training- preliminary results. In *ICSLP '96*, Philadelphia, PA, USA, Oct 1996.

[3] H. Franco, L. Neumeyer, Kim Y., and Ronen O. Automatic Pronunciation Scoring for Foreign Instruction. In *ICASSP '97*, München, Germany, April 1997.

[4] J. Fransen, Dave Pye, Tony Robinson, Phil Woodland, and Steve Young. WSJCAM0 corpus and recording description. Technical Report 192, Cambridge University Engineering Department, 1994.

[5] H. Hamada, S. Miki, and R. Nakatsu. Automatic evaluation of english pronunciation based on speech recognition techniques. *IEICE Trans. Inf and Sys.*, E76-D(3):352–359, March 1993.

[6] S. Hiller, E. Rooney, J. Laver, and M. Jack. SPELL: An automated system for computer-aided pronunciation teaching. *Speech Communication*, 13:463–473, 1993.

[7] C.J. Leggetter and P.C. Woodland. Speaker adaptation of hmms using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Department, Trumpington Street,Cambridge, June 1994.

[8] L. Neumeyer, H. Franco, M. Weintraub, and P. Price. Pronunciation Scoring of Foreign Language Student Speech. In *ICSLP '96*, Philadelphia, PA, USA, Oct 1996.

[9] C.L. Rogers, J.M. Dalby, and G. DeVane. Intelligibility training for foreign-accented speech: A preliminary study. *J. Acoust. Soc. Am.*, Vol. 96:no. 4, pt. 2, 1994.

[10] S.J. Young, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory, 1996.