USING LIKELIHOOD RATIOS TO PERFORM UTTERANCE VERIFICATION IN AUTOMATIC PRONUNCIATION ASSESSMENT

Febe de Wet, Catia Cucchiarini, Helmer Strik and Lou Boves
A²RT, Dept. of Language and Speech, University of Nijmegen, The Netherlands
{F.de.Wet, C.Cucchiarini, W.Strik, L.Boves}@let.kun.nl; http://lands.let.kun.nl/

ABSTRACT

The aim of our current research is to investigate the possibility of using likelihood ratios to perform utterance verification within the context of automatic oral proficiency assessment. The likelihood ratios under investigation have the appealing feature that they may be computed simply by using an off-the-shelf automatic speech recognition system in two different recognition modes (forced and free phone) instead of using a system with specifically trained anti-models. We achieved 93% correct classification for 10 phonetically rich sentences uttered by 60 non-native language students.

1. INTRODUCTION

The long-term goal of our research is to employ ASR technology in an automatic pronunciation test for Dutch as a second language. As a consequence of this aim we are not concerned with learners of Dutch with a specific mother tongue, but rather with a group of speakers who are highly varied in this respect. In this sense our situation is different from that of many studies on the use of ASR in automatic pronunciation assessment, in which fixed language pairs (L1 & L2 fixed) are involved [e.g. 1, 6]. In our case L2 is always Dutch, but the L1 of the language students are extremely diverse.

In [3] we showed that human ratings of pronunciation quality can be predicted very well by automatically obtained temporal measures. In this study, read speech of natives and non-natives was scored for pronunciation quality by different groups of experienced raters. Subsequently, the data was processed by means of an ASR-system using forced Viterbi alignment to obtain a number of temporal measures. The expert ratings and the machine scores were then submitted to statistical analyses which revealed a strong relationship between the two sets of scores, e.g. correlations between the human scores and rate of speech (*ros*) varied between 0.81 and 0.93 (see Table 1, Section 4.1). On the basis of these findings we could conclude that automatically calculated temporal measures can be employed successfully in pronunciation assessment.

However, even though these experiments revealed very high correlations between *ros* and expert human ratings, some issues remain unresolved. For example, students who know that the automatic system completely relies on temporal measures can achieve high scores simply by speaking fast, despite a poor pronunciation quality. As a worst case example, students who produce an arbitrary utterance fast enough might even obtain high grades. In more general terms this means that using only temporal measures to evaluate pronunciation quality introduces two problematic issues, i.e. (1) subjects who produce a target prompt fast but with poor pronunciation and (2) subjects who utter an incorrect utterance fast (where an incorrect utterance is any utterance other than the prompted one) may obtain high scores - in both instances unjustly so.

In [3] we used read speech. Even though in read speech one should know beforehand what a speaker is going to say, one can never be sure that test subjects will utter the prompted sentences exactly as they are represented on paper. For this reason, in [3], we used specific verbatim transcriptions of the speech material, including phenomena such as hesitations, false starts, repetitions, repairs, etc. This introduces a third problem, i.e. that making specific transcriptions is both costly and time consuming.

In [4] we addressed the first problem. In the present paper we will focus on the solutions of the second and third problem. First, we will also introduce likelihood ratios (LRs) that appear to be very successful in performing utterance verification. We will also show that the correlation between automatically calculated temporal measures based on prompts and human expert ratings are just as high as the correlation between automatic measures calculated from specific orthographic transcriptions and human expert ratings.

In order to get a better understanding of the LRs we investigated to what extent they vary as a function of the duration and the spectral content of the input speech. To this end utterances of different duration were synthesized for both a female and a male voice.

This paper is organized as follows. In Section 2 we give an overview of the speech material used and in Section 3 we describe how the experiments were conducted. Section 4 reports on the results obtained during experimentation. The conclusions are presented in Section 5.

2. MATERIAL

2.1 Training Material

The material that was used to train the ASR-system consisted of the phonetically rich sentences of 4019 speakers from the Dutch Polyphone database [5]. 38 monophone models were trained. The phonetic transcriptions used during training were obtained by concatenating the canonical transcriptions of the words, taken from a lexicon (For further details, see [2,3]).

2.2 Test Material

2.2.1 Read Speech

The speakers involved in this experiment are 60 non-native speakers (NNS), 16 native speakers with strong regional accents (NS) and 4 Standard Dutch speakers (SDS). The speakers in the three groups were selected according to different sets of variables, such as language background, proficiency level and sex, for the NNS group, and region of origin and sex for the NS and SDS groups. Each speaker read two sets of five phonetically rich sentences (about one minute of speech per speaker) over the telephone [2].

2.2.2 Prompts vs Specific Transcriptions

In some cases the subjects produced utterances which deviated from the prompts. Therefore, the recorded speech material was orthographically transcribed. We will refer to these detailed verbatim transcriptions as the *specific transcriptions*, while the prompts will simply be referred to as the *prompts*.

2.2.3 Synthetic Speech

The time and spectral dependency of the LRs were also investigated. For this purpose synthesized speech data was created using a diphone speech synthesis system. Because the synthesis is not formant-based, there is no direct way to manipulate the spectral content of the signals. As an approximation of a change in spectrum, we used a female and a male voice. The average duration of each utterance (as produced by the 4 SDS speakers) was taken as a starting point and then two faster and two slower versions of each utterance were synthesized by varying the duration of the vowels and consonants in each utterance. In total 10 different versions (5 male and 5 female) of each of the 10 phonetically rich sentences were synthesized.

3. METHOD

It is well-known that LRs can be used for utterance verification [e.g. 7]. According to the likelihood ratio test, the null hypothesis H_0 (X is a target utterance) is accepted if the likelihood ratio statistic, T(X), exceeds a certain threshold, ω. T(X) is determined in terms of the null hypothesis and the alternative hypothesis, H_1 (X is not a target utterance), as follows:

$$T(X) = \frac{likelihood\ score\ H_0}{likelihood\ score\ H_1}$$

H_0 is accepted if $T(X) \geq \omega$, where ω is a threshold value determined from training data. In utterance verification problems, the likelihood score for H_0 is obtained by determining the acoustic likelihood of the target utterance. The corresponding score for H_1 is evaluated as the acoustic likelihood of a so-called *anti-model* or *world model*.

However, the likelihood ratio test is by no means trivial to implement, if only because it requires a clear definition of exactly what anti-models should represent. In this regard we were faced with two problems. First, it is difficult to determine exactly what an anti-model should represent if the target utterance is known to be produced by someone learning Dutch as a second language. Other than in most other studies reported on in this field [1, 6], there is an enormous diversity in the language backgrounds of the subjects whose Dutch oral proficiency needs to be evaluated by our system. Secondly, even if it were possible to clearly define such an anti-model, the availability of a sufficient amount of applicable training material would still remain an unresolved issue.

Given that we could not train specific anti-models, we looked for a less complex approach in which a standard off-the-shelf ASR could be used to calculate LRs. In this approach the ASR is used in two different modes, e.g. forced and free phone recognition mode, and the likelihoods calculated for each mode are divided to obtain a LR. The resulting LR was then used to classify an utterance as correct or incorrect. In contrast with previous experiments, we did not use specific orthographic transcriptions during these calculations, and the transcriptions of the utterances were taken to be the prompts instead.

Different likelihoods were calculated by means of different versions of a standard HMM-based automatic speech recognition (ASR) system (for further details about the ASR-system, see [8]). For instance, we experimented with forced Viterbi alignment and free phone recognition, phone models and broad-phonetic class models, context independent and dependent HMMs, etc. Due to space limitations we will limit the scope of the present discussion to two sets of likelihood (LH) scores, i.e. LH_{forced} and $LH_{freephone}$.

LH_{forced} was evaluated by using a forced Viterbi alignment to align an acoustic signal with its prompt. To perform the alignment, the ASR-system based on 38 monophone HMMs was used together with a lexicon containing all the words occurring in the set of phonetically rich sentences.

$LH_{freephone}$ was determined with the same monophone-based ASR-system, but this time operating in free phone recognition mode, i.e. the lexicon consisted of phones only and all the phones in the resulting language model had an equal probability. We used the LH-values corresponding to the path through the word graph with the highest acoustic score in calculating the following likelihood ratio (LR):

$$LR = \frac{LH_{forced}}{LH_{freephone}}$$

4. RESULTS

4.1 Prompts vs Specific Transcriptions

Previous work [2,3] has shown that human expert ratings of pronunciation quality can accurately be predicted by automatic measures based on temporal information alone, e.g. *ros*. These measures were calculated from segmentational information that was obtained using a forced Viterbi alignment together with the specific transcription of the utterances. In a realistic application it would not be feasible to create a specific transcription for each utterance that is to be evaluated. We therefore needed to establish whether meaningful automatic pronunciation measures could also be calculated using prompts instead of specific transcriptions. To this end we calculated, the correlation coefficients between the automatic measure, *ros*, and the human expert ratings based both on the prompt and the specific transcription of each utterance. Table 1 shows the correlation coefficients between *ros* and the average values of the three sets of human expert ratings for both instances (see [2,3] for further details).

Parameter	Specific	Prompts
Overall Pronunciation	0.82	0.83
Segmental Quality	0.81	0.81
Fluency	0.93	0.93
Speech Rate	0.91	0.91

Table 1 Correlation coefficients between human expert ratings and ros evaluated with specific transcriptions and prompts.

The values given in the Table 1 show that there is only a marginal difference between the correlation coefficients based on the specific transcription and those based on the prompts. The small discrepancy between the two sets of values may be explained by the fact that the test subjects were cooperative in that they did their best to complete the reading task to the best of their abilities. One would therefore not expect substantial differences between the verbatim transcriptions and the

prompts, certainly not at segmental level. This expectation was confirmed by the observation that, for our data, the difference between the two sets of transcriptions was limited to phenomena such as hesitations, repetitions, false starts, repairs, etc.

In other instances where subjects may attempt to "fool" the system by producing random utterances with a high speech rate, one would expect larger differences between the two sets of results. It is likely that the Viterbi alignment process will not yield meaningful segmentational information if there is absolutely no relation between the speech signals and the acoustic models corresponding to the prompt. This makes it all the more imperative that automatically calculated temporal measures should be supported by some form of utterance verification if it is to be used in automatic pronunciation assessment applications.

4.2 LRs & Utterance Verification : Read Speech

As was mentioned in Section 2.2.1, each subject produced 10 utterances. In turn, each of the 10 utterances was treated as a correct utterance, and the other 9 as incorrect utterances. The goal is to determine whether an utterance is correct or incorrect based on the LR between its forced and freephone LH-scores. To this end, each utterance was subjected to 1 freephone and 10 forced recognitions. The 10 forced recognitions were performed using the prompts of the 10 utterances where 1 of the prompts was the correct transcription for the utterance at hand and the other 9 were incorrect transcriptions.

Figures 1 and 2 show the percentage classification error that is made as a function of the LR-values. For instance, Figure 1 shows the results for the 20 native speakers (4 SDS + 16 NS). The lefthand curve is based on the LRs of the 200 correct utterances (`) and the righthand curve on the 1800 incorrect utterances (.). These curves may be used to set an LR-threshold that determines the error level that is allowed in the classification. Values above the threshold that correspond to a correct utterance will unjustly be classified as an incorrect utterance (false reject) while LR-values of incorrect utterances that fall below the threshold will be classified as correct (false accept).

Figure 1 illustrates the results of the utterance verification experiment based on the read speech material of the 20 native subjects (4 SDS + 16 NS). It shows that, if an LR threshold value of ±12 is chosen, it is possible to achieve almost 100% correct classification.

The results in Figure 2 correspond to the experiments performed for the 60 non-native (NNS) subjects. The point of

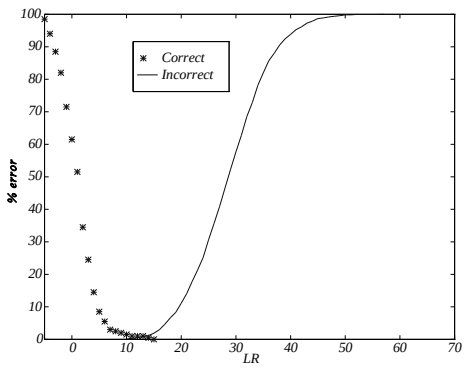

Figure 1 False reject and false accept curves for LRs calculated from native data.

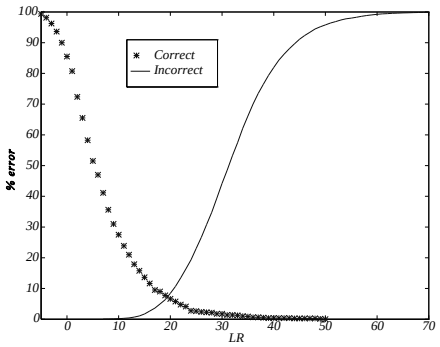

Figure 2 False reject and false accept curves for LRs calculated from non-native data.

equal error (the LR-value for which the number of false accepts is equal to the number of false rejects), is close to 20. At this point a classification error of ±7% is made. This means that it is possible to distinguish between correct and incorrect utterances using LRs, i.e. it is possible to determine whether a speaker had actually produced the utterance that he/she had been prompted to (correct utterance) or not.

4.3 LRs & Utterance Verification : Synthetic Speech

LRs similar to those described in the previous section were calculated for the synthetic speech data. Figure 3 illustrates the results of this experiment. The LR-values of the whole set (all durations) of both the female and male utterances are incorporated into this figure. It shows that the synthesized speech can be classified as correct or incorrect utterances with zero error, the false accept and false reject curves do not even intersect. From this observation it may be concluded that, to the extent that the range of these variables has been explored in the current experiment, the discriminative ability of the LRs to perform utterance verification is not affected by changes in the duration and/or spectral content of an utterance.

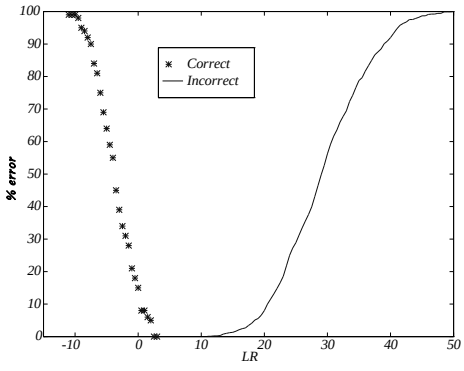

Figure 3 False reject and false accept curves for LRs based on synthesized speech.

5. CONCLUSIONS

We have shown that meaningful automatic pronunciation assessment measures can be calculated from speech data using prompts only. This means that the enormously time-consuming

task of creating specific transcriptions for all the material that needs to be evaluated is no longer a requirement for reliable evaluation.

Furthermore, it was established that LRs that are calculated from acoustic scores based on prompts can be used successfully to perform utterance verification. First of all, experiments performed on synthesized speech data revealed that the discriminative ability of LRs to perform utterance verification is not affected by changes in the duration and/or spectral content of an utterance, at least to the extent that such changes could be modeled by our data.

Target and incorrect utterances were correctly classified in almost 100% of the utterance verification tests performed on native speech data. For non-native data, correct classification was achieved in 93% of the cases. There are two possible explanations for the lower classification rate of the non-natives. Firstly, we did not have ample data to train acoustic models based on the non-native material. Models trained only on native speech may not be optimal to perform utterance verification for non-native speakers. Secondly, utterance verification may be inherently more difficult for non-native subjects because there is probably much more variation in their articulation, given the diversity in their L1 language backgrounds.

Based on our results we conclude that LRs that can be computed without training any specific anti-models can be used to perform utterance verification successfully within the context of automatic oral proficiency assessment. It may very well be remarked that distinguishing between 10 phonetically rich sentences is by no means an intricate utterance verification task, but within the context of using off-the-shelf ASR technology in application software that is meant to support second/foreign language learning and testing, this is indeed an encouraging result.

6. ACKNOWLEDGEMENTS

This research was supported by the Ministry of Economic Affairs through SENTER, the Dutch National Institute for Educational Measurement (CITO), Swets and Zeitlinger and KPN telecom. The research of Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. The research done by Febe de Wet was financially supported by a scholarship from de Stichting Studiefonds voor Zuidafrikaanse Studenten, Nuffic and the Van Ewijck Stichting. We would like to thank Joop Kerkhoff for preparing the synthesized speech data.

7. REFERENCES

[1] H. Bratt, L. Neumeyer, E. Shriberg and H. Franco (1998), Collection and detailed transcription of a speech database for for development of language learning technologies. *Proceedings ICSLP '98*, Sydney, Australia, pp.926-929.

[2] C. Cucchiarini, H. Strik and L. Boves (1997), Using speech recognition technology to assess foreign speakers pronunciation of Dutch. *Proceedings New Sounds '97*, Klagenfurt, Austria, pp.61-68.

[3] C. Cucchiarini, H. Strik and L. Boves (1998), Automatic pronunciation grading for Dutch. *Proceedings STiLL '98*, Marholmen, Sweden, pp.95-98.

[4] C. Cucchiarini, F. de Wet, H. Strik and L. Boves (1998), Assessment of Dutch pronunciation by means of automatic speech recognition technology. *Proceedings ICSLP '98*, Sydney, Australia, pp.751-754.

[5] E. A. den Os, T. I. Boogaart, L. Boves and E. Klabbers (1995), The Dutch Polyphone corpus. *Proceedings Eurospeech '95*, Madrid, Spain, pp.825-828.

[6] G. Kawai and K. Hirose (1998), A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. *Proceedings ICSLP '98*, Sydney, Australia, pp.782-785.

[7] C.H. Lee (1997), A unified statistical hypothesis testing approach to speaker verification and verbal information verification. *Proceedings COST Workshop*, Rhodes, Greece, pp.63-72.

[8] H. Strik, A. Russel, H. Van den Heuvel, C. Cucchiarini and L. Boves (1997), A spoken dialogue system for the Dutch public transport information service. *International Journal of Speech Technology*, vol.2, pp.121-131.