3P – Portuguese Pronunciation Professor (October 2014)

M. Lopes, MEEC, 65433

Abstract— The research presented investigates a solution for creating a computer assisted language learning (CALL) system for EP (European Portuguese) using as its base the work of Witt (1999)[1]. The algorithm is based on Goodness of Pronunciation (GOP), a measure that uses confidence scores drawn from automatic recognition and alignments results at phone-level. The GOP is computed for native data, in order to select a threshold that separates good and bad pronunciation. The method is then tested with a non-native corpus, and the results are analyzed and adjusted using several performance measures. The GOP module has been integrated in an existing web interface.

Index Terms— CAPT, GOP, normalization, pronunciation, European Portuguese.

I. INTRODUCTION

Learning to pronounce written words means learning the intricate relations between a language writing system and its speech sounds [2]. When children learn to read and write in primary school, they face an analogous learning task, as do students when mastering the writing system, the speech sounds, and the vocabulary of a language different from their mother tongue.

Learning to pronounce words can also be modeled on computers. In contrast with humans, machines can be modeled in such specific ways that, for instance, a machine can be set up to accommodate a data base of representations of word-pronunciation knowledge, without having learned any of those representations by itself: it is hardwired in memory by the system’s designer.

Computer assisted language learning (CALL) applications and, more specifically, computer assisted pronunciation training (CAPT) applications for language learning that make use of automatic speech recognition (ASR) have received considerable attention in recent years. Most of the literature on pronunciation assessment in language learning has focused on pronunciation scoring, while less attention has been paid to error detection.

But before CAPT methods can be devised, it is important to recognize the specific difficulties encountered in pronunciation teaching. First and foremost, explicit pronunciation teaching requires the sole attention of the teacher to a single student in order to analyse his speech and give some notes on how to improve. This in a normal classroom environment poses a problem. Then learning a new language involves a large repetition of the words that requires not only a mental task but demands coordination and control over many muscles to achieve proficiency. These may cause social implications to students that are afraid to perform in the presence of others. This costly time consuming approach and therefore its automatization is highly desirable for self-study.

On the other hand, these technologies do not take into account the variations due to speaker accent, demanding a strict distinction among the different sounds unlike what would happen with human teachers. So it can be said that there are two strands in the area of pronunciation learning: teaching correct pronunciation of a foreign language to students, which requires a precise phoneme recognition and is more objective and easily computed, and assessing the pronunciation quality of a speaker speaking a foreign language that can tolerate more mispronunciations, but is also more correlated with what human teachers perceive as the correct pronunciation.

II. 3P–SYSTEM DESIGN

A. State of the art

One of the first functioning projects was the SPELL project [1] which concentrated on specific phonemes. The standard method, and the method used in this research was firstly analyzed by Witt (1999) in “Use of Speech Recognition in computer assisted language training”. This method uses a measure denominated GOP to score the pronunciation. There are other possible processes to measure, such as, computing MFCC [3] or the likelihood [4] [5] but to this date there are not register of a method with better efficiency than the described by Witt or some of its modifications [6] [7] [12].

As for available tools/software, free software or applications available in the market mostly focus on the hearing and repetition of several words or sentences without giving any feedback on how well they were pronounced. We have not found any paid interactive applications for EP, but there are several for more practiced languages such as English.

B. Overall System

For measuring the quality of the pronunciation, the process requires an audio file and its transcription. The audio speech is digitalized, then, using Audimus, the in-house recognizer, posterior probabilities on 20 ms frames are calculated from the
extracted features. Subsequently, a GOP score is calculated for each frame, and given a score on each phoneme and word by averaging the GOP from each frame. Afterwards the GOP is normalized and using a pre-established threshold, from native speakers’ data, the threshold is adapted in order to obtain the maximum efficiency in scoring the phonemes as a correct or incorrect utterance, having the concern not to augment it so much that all phonemes are considered correct. Finally, given the subjective nature of this threshold, the scores of the system are compared with three human judges in order to compute its correlation.

1) Audimus

Audimus is an Automatic Speech Recognition System customized to the European Portuguese language and developed by Spoken Language Systems Laboratory (L2F) of INESC-ID [8]. The system is based in a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs) [9]. As an output, Audimus gives the posterior probabilities of each one of the SAMPA phonemes and the identification of the phoneme of every frame.

This system starts by dividing the desired audio file into 20 ms frames and in each frame it extracts three types of features thus sectioning them into three different branches. The first branch extracts 26 PLP (Perceptual Linear Prediction) features, the second 26 Log-RASTA (log-RelAtive SpecTrAl) features and the 3rd uses 28 MSG (ModulationSpectrogram) coefficients. Then each branch incorporates an MLP classifier that is used to estimate the probability based on the distinctive extracted features. Each MLP has the same basic structure, which is an input layer with 9 on text frames, a non-linear hidden layer with over 1000 sigmoidal units and 40 softmax outputs. Lastly the MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches using an average in the logprobability domain [8] [9].

![Waveform](image)

Figure 1 - Visualization of the phoneme division in Wavform.

This tool was employed in this project not only to calculate the posterior probabilities, i.e. the probability of a frame of an introduced audio file being one of each phoneme described previously but also as a mean for identification of when a phoneme is uttered given an audio transcription.

2) GOP

The GOP method (goodness of pronunciation) was introduced by Witt and Young (1999) and is one of the most used methods to score the articulation of words. Its popularity is due to its reduced computational complexity and indistinctness of the language applied. This means that the same method can be used for different dialects, as long it has the analysis of the posterior probabilities and the sectioning of the phonemes in the utterance. Although it has been shown that the method can yield satisfactory results [10], it requires the determination of a threshold to define the boundary between a good and a bad pronunciation. Thus the quality of the GOP scoring depends on the models utilized and on the native speakers employed. Nonetheless the GOP is calculated equally for both accurate and inaccurate utterances.

The GOP algorithm calculates the likelihood ratio that the recognized phoneme corresponds to the phoneme that should have been spoken for each phoneme in an utterance. The GOP score of phoneme p is defined as the frame-normalized logarithm of the posterior probability P(p|O), where O(p) refers to the acoustic segment uttered by the speaker. NF(p) corresponds to the number of frames in the acoustic segment O(p) [1].

\[ PP(p) = \log \left( \frac{P(p|O)}{NF(p)} \right) \]

This normalization takes the nearer it is to the 0 the greater the value, that the sum can be approximated by its maximum value, then the derived GOP can be described as Eq.2.

\[ GOP(p) = \log \left( \frac{p(p|O)}{\max_{q\in Q} p(p|q)} \right)/NF(p) \]  

The GOP(p) value is always equal or greater than 0. The greater the value, the more likely there is a mispronunciation. In contrast, the nearer it is to the 0, the more probable that the pronunciation is as a native [38].

3) NGOP

In order to reduce the influence of extreme values or outliers of the data set without having to remove them, a Sigmoidal normalization was applied. This way, all the data is included and since this normalization is almost linear near the mean value, the standard deviation of the mean is preserved. The normalized data is in the range between 0.0 and 1.0 [11].

This normalization takes the raw GOP score and concatenates it to a GOP score, denominated NGOP (normalized GOP), into the former range. That is,

\[ NGOP = sigmoid(s_u) = \frac{1}{1+\exp(-u)} \]  

where the parameters alpha and beta are empirically found according how rapidly it is wanted to reach the maximum
values and at what values of the abcises the scale starts, respectively. This way it is also easier to visualize the boundaries between a good and bad score.[11] [12].

4) Threshold

In order to distinguish when a GOP begins to define an incorrect utterance there is a need to establish a threshold. The threshold is different for each phoneme. Given a native speakers training corpus, the threshold value can be calculated using not only the mean but also the standard deviation of the NGOP by this expression:

\[ T_{p1} = \mu_p + astd_p + \beta \]  \hspace{1cm} (4)

where \( \alpha \) and \( \beta \) are empirically determined scaling constants[1].

5) Performance measure

To analyse the performance of the NGOP classification algorithm for a given threshold, four decision types can be defined: correctly accepted (CA) phoneme realizations, when phonemes that were pronounced correctly are also judged as correct; correctly rejected (CR), when phonemes that were pronounced incorrectly are judged as incorrect; false accepted (FA), when phonemes that were mispronounced are erroneously judged as correct; and false rejected (FR), when phonemes that were pronounced correctly are judged as incorrect [1] [39].

To achieve a good performance the algorithm has to be able to not only detect mispronunciations, but also to not classify them as a correct articulation. As a result, the performance of the scoring can be defined by:

\[ SA = \left( \frac{(CA + CR)}{(CA + CR + FA + FR)} \right) \times 100 \]  \hspace{1cm} (5)

where the objective is to achieve optimal performance by maximizing the scoring accuracy while minimizing the false acceptances. Other useful performance measures include the calculation of the precision (number of correct results divided by the number of all returned results), recall (the number of correct results divided by the number of results that should have been returned) and F-measure (the weighted average of the precision and recall) of correctly accepted or rejected phonemes realizations [1]:

Precision of CA = \( \left( \frac{CA}{CA + FA} \right) \times 100 \)  \hspace{1cm} (6)
Precision of CR = \( \left( \frac{CR}{CR + FR} \right) \times 100 \)  \hspace{1cm} (7)
Recall of CA = \( \left( \frac{CA}{CA + FR} \right) \times 100 \)  \hspace{1cm} (8)
Recall of CR = \( \left( \frac{CR}{CR + FA} \right) \times 100 \)  \hspace{1cm} (9)

\[ F\text{measure} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \]  \hspace{1cm} (11)

III. DATABASE

This research will focus on comparing EP with two different languages, Spanish and Bulgarian. European Portuguese is a West Iberian Indo-European language composed phonetically by 39 phonemes, including the pause. Comparing phonetically with Spanish, in the latter there are no nasal vowels/semi-vowels as well as open-mid and mid vowels. The voiced fricatives /ʃ/, /ʒ/, /ʒ/ and the unvoiced /S/ are nonexistent. However, there are additional fricatives (/T/, such as in cinco, and /Iter, as in mujer), and affricates (ʼhS/, as in mucho and /fj/, as in hielo). As for Bulgarian, EP is phonetically similar. Most consonants of EP are present in Bulgarian with the exception of /hl/, /n~/ and /R/. The only semi-vowel is /i/ and there are no nasalized vowels. As for the other vowels, only /E/, /e/ and /o/ are not present.

The native corpus is composed by 15 people from the Lisbon area, 7 males and 8 females, with age between 22 and 24 years old. The non-native is composed by one 23 year old female Venezuelan, which has Spanish as the native language, and a group of 11 Bulgarians, 6 males and 5 females, with age between 27 and 42 years old. The group was asked to read several sentences, having this over 14000 phonemes by the native and 9600 by non-native speakers.

The sentences were recorded using a high-quality head-mounted microphone with Mono 16-bit resolution and 16 kHz sampling rate.

The native speakers corpus was used to compute the threshold. The non-native as group, separated by native languages, tested and adjusted the former threshold. It should be noted that the quality of the corpus is not the best for the non-natives, having several pauses and gasping in many sentences due to the degree of difficulty of these sentences was not in accordance with the level of the students. In many cases, the alignment was difficult, especially from middle to the end of each track. To improve the alignment, the tracks were divided into simple sentences which provided a slight improvement, nonetheless the disparities still caused several misalignments.

IV. RESULTS

Using Audimus the posterior probabilities are computed in forced alignment mode for the native speakers speech. These are used in the approximated expression Eq. 2 to calculate the GOP score. Then a normalization is calculated with alpha=10 and beta=-10 giving the NGOP.

The first part of the implementation is repeated for the non-native corpus, but this time each phoneme resulting from the alignment is classified by a human jury as either a good or bad pronunciation.

For the first tests, the mean of the score of the natives was established as the threshold. Testing the score in the non-native data, in the phonemes with SA inferior to 70% the threshold was augmented using the expression Eq. 4 and alpha=0.5 and beta=0.1. Since the corpus is limited the results can be inexact for certain phonemes

A) Native Results

The values of NGOP for the natives vary between 0.33 for w and 0.005 for the interword_pause. There are problems with vowel reduction with @, u, and 6 and problems with word co-articulation, which result from the training of Audimus not taking into account this phenomena.

To better examine the results and since a perfect pronunciation implies that the score is 0, i.e. the PP of the phoneme is the maximum PP, a confusion matrix was created...
in other to count which phoneme has the maximum PP. It is noticeable that for natives in most cases the maxima are the phoneme itself, with the exception of some /z/ being pronounced as /S/, and some /u/ and /@/ being deleted, not only in intra-word position, but also in word boundaries.

B) Spanish Results

For the phoneme /e/, /@/, /o/, /u/, /u~/, /l~/and /w/, the threshold was modified. But despite improving the SA in the majority of cases it did not surpass the 70% accuracy. Also, despite some phonemes having an accuracy of 1, perfect accuracy, this does not mean necessarily that the pronunciation is perfect, but can also mean that there are not enough occurrences of the phoneme. The average SA is 82.53% and the average F measure 88.45%.

C) Bulgarian Results

Here, 21 phonemes had SA lower than 70% and the threshold was also modified. Likewise, in this case despite improving the SA in the majority, the accuracy did not surpass 70%.. Moreover, since there were more phoneme samples in this case the SA decreased, not having any phoneme with accuracy equal to 1. The average SA is 73.23% and the average F measure 76.66%. The lower scores can also be justified by the quality of the audio tracks provided.

D) Performance measures

With the threshold established, and considering that this is a subjective technique, three performance measures were also computed to compare the scoring between the transcription by two judges or one judge and automatic NGOP. This allows a cross-validation between judges in the number of the errors that each can find.

The first one is strictness and measures how strict a judge is. This also allows seeing how subjective judgment interferes with the border line cases between correct and incorrect. And to compare to judges it is simply means to compute the difference between the two judges, J1 and J2 (AA).

The second measure is the agreement (CC) and takes into account if the phonemes are considered mispronounced or not by two different judges.

The last one, the cross correlation (PC) measures the overall agreement between the reference and the detected error, i.e. the similarity between all segments which contain rejections in the transcriptions.

With the performance measures described above, a small study was conducted to compare the ratings between human judges. The inter-judge correlation was measured for 20 calibration sentences (8 of natives and 12 of non-natives), for 3 natives judges. The results were calculated by averaging A, CC and PC for each judge in relation to all others.

As for comparison with the NGOP, after adjusting the threshold to the result of best score possible, each judge obtained the AA, CC and PC below.

<table>
<thead>
<tr>
<th>Judge</th>
<th>AA</th>
<th>CC</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>J1 vs NGOP</td>
<td>0.86</td>
<td>0.52</td>
<td>0.79</td>
</tr>
<tr>
<td>J2 vs NGOP</td>
<td>0.88</td>
<td>0.40</td>
<td>0.67</td>
</tr>
<tr>
<td>J3 vs NGOP</td>
<td>0.87</td>
<td>0.43</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table 2 - Judges vs GOP scoring

Evaluating these results we find that there was not a very discrepant view in comparison to the NGOP, it all depended each subjective interpretation for each judge. A stricter judge would scale down the results.

V. INTERFACE

The VITHEA project interface was adapted to make a 3P Interface, however due to time restrictions it was not possible to test the Interface as a whole. Nevertheless, the VITHEA interface was profoundly tested with a good performance. It uses JSP/Servlet Server, connected to Audimus, a Database Management System and the internet, to lodge a Flash application available in a Web browser.

The interface has on the left side what the student is supposed to repeat and on the right side the sound recording and feedback instructions player. [13]

VI. CONCLUSIONS

The SA for Spanish and Bulgarian are satisfactory, despite the bad recording conditions of the corpora, but can be influenced by how many times a phoneme occurs as well. A phoneme that occurs more frequently is better evaluated. It was also noticed that despite the good phoneme based results, the fact that some phonemes were not uttered with the same duration as a native makes the complete word/sentence sound unnatural. This implies that further improvements should take duration into account. This was not implemented for two reasons, firstly the Audimus already measure the duration of each phoneme and if phoneme is not at least the time interval, computed by averaging the time of each phoneme in the corpus that Audimus was trained, it doesn’t count as the
said phoneme. And secondly, even if the interval of Audimus is incorrect if we wanted a time variable we would have to scrutinize every single phoneme in text and save the average time this phoneme in this particular place in the sentence, making the interface not so adaptable to further updates.

As for Future work, the principal flaw in this project was the lack of good non-native audio files, so a major improvement can be made by recording simpler sentences in accordance to the level of the students. There were only two languages analyzed, Spanish and Bulgarian, so a wider and more differentiated corpus would help to accommodate a more varied number of native languages. Another enhancement can be made by retraining the Audimus taking into consideration the vocalized reductions and the co-articulations. Finally, despite the VITHEA interface being heavily tested, it would be good to test the 3P interface with native and non-native subjects.

VII. ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisor Professor Isabel Trancoso and the L2F staff, especially to Professor Hugo Meinedo, Professor Alberto Gareta, Professor Thomas Pelligrini and Anna Pompili for the support, and provision of the source materials essential to helping me completing this project.

REFERENCES

[1] WITT, S., Use of Speech Recognition in Computer-assisted Language Learning, November 1999
[2] VILLALOBOS, O., Reflections on the connection between computer-assisted language learning and second language acquisition, February 2013