An Arabic Mispronunciation Detection System by means of Automatic Speech Recognition Technology

Khaled Necibi¹, Halima Bahi²

¹, ²LabGED Laboratory, University of Annaba-BP.12. Annaba, Algeria
¹khaled.necibi@univ-annaba.org, ²bahi@labged.net

Abstract: The aim of the work presented in this paper is to help the Algerian (Arabic) young learners having problems with pronunciation of Arabic language to improve their pronunciation skills through the use of ASR technology (Automatic Speech Recognition). The proposed system is based on mispronunciation detection methods where different scores of pronunciation are calculated to decide how the word was badly pronounced in terms of quantitative measure. The obtained results have showed that the GLL (Global average Log Likelihood) score detects mispronunciation with 86.66% of correct rejection. To validate this finding, another experiment was carried out, on an expanding corpus of data, based on measuring the calculated mispronunciation scores in terms of CA (Correct Acceptance), CR (Correct Rejection), FA (False Acceptance) and FR (False Rejection). The obtained results have shown that the GLL score, comparing to the other calculated scores, had the higher CA+CR (76.66%) and the lower FA+FR (23.32%).

Keywords: CALL, CAPT, automatic speech recognition, pronunciation scoring, Arabic pronunciation.

1. Introduction

The recent advances in Automatic Speech Recognition (ASR) technology have promoted developments of Computer Assisted Pronunciation Training (CAPT) systems and automatic pronunciation assessment is becoming feasible. The CAPT systems is an important part of Computer Assisted Language Learning systems (CALL) where the benefit is that the learner with pronunciation problems may get unlimited amounts of practice at any time. In fact, in the area of pronunciation learning, little attention has been paid to error detection. ASR has been used in CAPT for two different purposes: teaching a correct pronunciation [6] and assessing the pronunciation quality of the speaker [13][9][4] to help improve learner’s pronunciation skills. An ASR-based CAPT system includes often two main modules: the speech recognition module and the evaluation module which is also called the scoring module. In this paper we present our experiments based on mispronunciation detection and the assessment of a pronunciation in CAPT system dedicated to teach Arabic pronunciation. Our aim is that the proposed system must be able to detect words errors pronunciation with high accuracy. This paper is organized as follow. In section 1 an introduction is given. Some related works are given in section two to review systems for assessing pronunciation and how we can use them to fit with Arabic language. In section 3 we review the most popular pronunciation scoring methods; we chose some of them to score the Arabic learner pronunciation. In section 4, we present our proposed system and the experiment results obtained. An evaluation decision performance is done in section 5. Finally we conclude the paper in section 6.

2. Related works

Various approaches have been proposed in the field of pronunciation errors detection and they can be found in many literatures. In the following we present some of these approaches and how we can exploits them to fit with Arabic language.

The experiments presented in [10] make part of the VILTS (Voice Interactive Language Training System) project where the goal was to generate phonetic segmentations that were used to produce pronunciation scores at the end of each learning session. To generate reliable scores at phoneme level; segment classification scores, segment duration scores and timing algorithms were used. The relative phoneme duration was a good predictor of pronunciation proficiency.

Based on approximate Bayesian adaptation methods, the main goal of the proposed system in [3] was the automatic evaluation of the pronunciation quality. The pronunciation scoring methods used was: spectral match score, phone duration scores and speech rate for the different phonetic segments. The obtaining
machine and human scores correlate well according to the proposed system.
In [8], the authors have presented the PLASER (Pronunciation Learning via Automatic Speech Recognition) multimedia tool, which aim to teach the correct English pronunciation. The system was based on the GOP score algorithm. Before calculating scores, two kinds of exercises were proposed: minimal pair-exercises and word exercises. At the end of each pronunciation assessment the system return to the learner a simple feedback, based on the GOP score, highlighting the phonemes that were mispronounced.
The aim of the work presented in [11] was to qualitatively and quantitatively distinguish between native and nonnative speakers of a language on the basis of a comparative study of two methods: the relative positions of vowels in formant space, while the second one exploit the sensitivity of trained phoneme models to accent variations [11], as captured by the log likelihood scores to distinguish between native and nonnative speakers. The study performed in [2] makes part of the Arabic Pronunciation improvement system for Malaysian teachers of Arabic language project. The system proposed in [2] was dedicated to teachers of Arabic language in order to help them to learn Arabic language quickly and to improve their pronunciation skills. The system consists on giving a mark for each pronunciation using the log likelihood probability. The proposed system achieved average accuracy of 89.63% comparing to human score.
Authors in [1] have presented the Versant Arabic Test (VAT) which consists on a fully automated test and score of spoken Modern Standard Arabic (MSA) aiming at facilitating the listening and speaking. This VAT provides four scores calculation over various dimension of facility: sentence mastery, vocabulary, flency and pronunciation scoring. Thos scores were obtained by applying an HMM automatic speech recognition.
Different languages were studied in works previously reviewed like English and Japanese. Little attention has been paid to the error detection of Arabic language pronunciation. As we can see in the above works, different scoring methods were proposed to detect errors pronunciation. Those methods have the benefit that they can be obtained in easily way with an ASR system, and can be calculated in similar ways for all language. So, we would like to apply these methods to fit with Arabic language pronunciation.

3. Existing Measures For Pronunciation Scoring

Different measures have been proposed to quantitatively assess the pronunciation quality of the learner (or to measure the fluency of the speech) can be found in the literature. As instance [7]:

- The phonation time ratio: this is calculated as the percentage of time spent speaking as a percentage proportion of the time taken to produce the speech sample.
- Mean length of runs: which is calculated as an average number of syllables produced in utterances between pauses of certain duration (ex. 0.25 seconds and above).
- Pace: the number of stressed words per minute.
- Space: the proportion of stressed words to the total number of words.

In the following, we present the most popular measures that were used in the field of error pronunciation detection.

3.1. Phoneme Duration Score

To calculate this score [3], the duration in frames of the i\textsuperscript{th} phoneme from the viterbi alignment is required. The silent phonemes in the calculation of this measure are excluded. To obtain the corresponding phoneme duration score, the log probability of the phoneme duration is computed using a duration distribution of that phoneme. Again, the corresponding word duration score is defined as the average of the phonemes scores over the word.

3.2. Rate Of Speech (ROS)

Speakers with correct pronunciation often speak faster than do beginning learner. Thus, the Rate of Speech (ROS) can be used as a predictor of the degree of pronunciation correctness. The ROS is calculated by dividing the total number of phonemes by the time taken to produce them including also the silent pauses [12].

3.3. Rate Of Articulation (ROA)

In calculating the rate of articulation [12], the total number of phonemes (or speech sounds) produced in a given speech sample was divided by the amount of time taken to produce them. Unlike in the calculation of the rate of speech, pause or silent units are excluded. As nonnative tends to have lower articulation rate comparing to native speaker, the articulation rate can be considered as good discriminative measure.

3.4. The Global Average Log Likelihood Score (GLL)

The logarithm of the likelihood of the speech data, computed by the viterbi algorithm using the HMM obtained from native speakers is a good measure for the similarity between the correct and the incorrect pronunciation [10]. However, for a given level of mismatch between speech and models, the log likelihood depends on the length of the word (if we consider the log likelihood of each phoneme of the
word). To normalize the effect of the word length, the global average log likelihood [10] is calculated using the following equation:

$$\text{GLL}(P_i) = \frac{\sum_{i=1}^{N} LL_i}{\sum_{i=1}^{N} d_i} \ldots (1)$$

Where $LL_i$ is the log likelihood corresponding to the $i^{th}$ phoneme and $d_i$ is its duration in frames, with sums over the number of phonemes $N$.

3.5. The Local Average Log Likelihood Score (LLL)

The degree of match during longer phonemes tends to dominate the global log likelihood measure. Although shorter phonemes may have an important perceptual effect, as their duration is smaller, the degree of mismatch along them may be swamped by that of longer phonemes. To attempt to compensate for this effect we use the local average log likelihood (as in [10]) given by the next formula:

$$\text{LLL}(P_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{LL_i}{d_i} \ldots (2)$$

Where $LL_i$ and $d_i$ are respectively the corresponding log likelihood and duration of the phoneme $P_i$, and $N$ is the total number of phonemes.

3.6. The Duration Normalized Log Likelihood (DNLL)

This score is obtained by dividing the log likelihood score of each phoneme obtained from the viterbi decoding by its corresponding duration (see the next equation).

$$\text{DNLL}(P_i) = \frac{LL_i}{d_i} \ldots (3)$$

3.7. The Goodness Of Pronunciation Score (GOP)

The Goodness Of Pronunciation (GOP) [5] is the most popular score used in the field of pronunciation teaching. This algorithm calculates the log likelihood ratio that a phoneme realization corresponds to the phoneme that should have been spoken (the GOP score). The learner’s speech is subjected to both a forced and free speech recognition phase. A GOP score of a specific phone realization is then calculated by taking the absolute difference of the log probability of the forced and the log probability of the free recognition phase.

$$\text{Sigmoid}(P) = \frac{\alpha}{e^{-\beta LL_P}} \ldots (5)$$

where $P$ represents the phoneme, $LL_P$ is the log likelihood of the corresponding phoneme and $\alpha$ and $\beta$ are two parameters empirically found. Once the log

4. An ASR System to Teach Arabic Pronunciation

4.1. System Architecture

It is assumed that with HMM models trained on native speech data (the good pronunciation), the log likelihood of the input speech data, computed by the viterbi decoding, seems to be a good measure of similarity between good (the correct pronunciation) and poor pronunciation [11].

![Figure 1. The system overview](image-url)
likelihood is normalized, the proposed system is then able to calculate the four measures GLL, LLL, ROS, ROA as illustrated in the figure 1. Once those scores are obtained, the system is then able (based on the most decision score result) to decide how the word is badly pronounced, comparing to its correct pronunciation, in term of quantitative measure.

4.2. System Configuration and Measured Scores Results Obtained

4.2.1. System Configuration

The system proposed works on isolated Arabic word recognition mode and uses acoustic models (19 phonemes Hidden Markov Model HMM) and a lexicon. The lexicon contains orthographic and phonemic transcription of words to be recognized. The corpus used in this work contains speech from 6 young speakers. The distribution of speakers was balanced in terms of age (in the range of 8 years to 12 years old) and gender (4 boys and 2 girls). One from them (boy) which has a good pronunciation was used as reference for the quantitative evaluation of the other five learner pronunciation. The speaker with good pronunciation was prompted to pronounce the words “Sun”: “Chamsoun in Arabic”, “write”: “Kataba in Arabic”, “Chair”: “Koursi in Arabic” previously stored with their phonetic transcription in the lexicon. Table 1 shows the words orthographic annotation and their corresponding phonetic transcription.

The other five speaker where choosing with precision and they have some pronunciation problems. As all young learners around the world, the Arabic one makes errors in reading words such as inversion of phonemes, confusion between letters and sounds and insertion/deletion of phonemes in/from words. It was easy found that, in the Algerian young disabled learner community, the pronunciation errors often made were in the words: “Chamsoun” where the phoneme /s/ is substituted by the phoneme /ch/ and then the word will be spoken as “Chamchoun”, the same thing for the word “Koursi” where the phoneme /k/ is substituted by /t/ (badly pronounced as “Tourssi”) and finally the word “Kataba” where the phonemes /t/ and /b/ were confused by replacing the /b/ position before /t/ and then the word will be spoken as “Kabata”. Those five speakers with pronunciation problems were asked in a second time to pronounce the three words “Chamsoun”, “Koursi” and “Kataba” where errors pronunciation were made as explained above. This makes at all three pronounced words considered as reference for the evaluation and fifteen pronounced words to be evaluated.

The acoustic models used in our experiments were a set of 19 phoneme Hidden Markov Model (HMM) trained from native Arabic speaker and representing 6 phonemes corresponding to the word Chamsoun /ch/ah/m/s/uh/n/, 6 phonemes for the word Kata

4.2.2. Results Obtained

In the following, speaker_ref denotes the speaker with good pronunciation and speaker1-5 denotes the speakers with pronunciation problems. The words pronounced by the speaker_ref were all correctly recognized by the ASR engine. This makes our evaluation process more precise. The next 4 figures show the proposed system results obtained.

The global average log likelihood (GLL)

![Figure 2. The global average log likelihood score results](image2)

The local average log likelihood (LLL)

![Figure 3. The local average log likelihood score results](image3)

<table>
<thead>
<tr>
<th>Words</th>
<th>Phonetic transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chamsoun</td>
<td>/ch/ah/m/s/uh/n/</td>
</tr>
<tr>
<td>Kataba</td>
<td>/k/ah/ah/b/ah/</td>
</tr>
<tr>
<td>Koursi</td>
<td>/k/uh/r/s/iy/</td>
</tr>
</tbody>
</table>

Table 1. The phonetic transcription of words to be evaluated
The scores obtained from the measures calculated indicate how certain the recognizer for a target word was pronounced correctly; the lower the confidence the higher the chance that the sound (or utterance) was mispronounced [12]. When looking at figures 2, 3, 4, and 5, corresponding to GLL, LLL, ROS and ROA respectively, it is easily to check if the word pronounced by the five speakers (speaker 1-5) was badly pronounced by the learner or not when comparing to speaker_ref.

5. Evaluation Decision Performance

To be able to decide whether the score calculated by the system can detect mispronunciation, we calculate an evaluation decision type defined as the correct rejection (Figure 6) which consists of a word that has been pronounced incorrectly and was detected to be incorrect since our test data contain only words that are badly pronounced. This score was calculated because in pronunciation learning context the system must encourage the learner and avoid discarding pronunciations which are correct.

The GLL score decision detects mispronunciation with 86.66 % which is the higher percentage comparing to the other decision scores. Thus, it is considered as good measure to detect mispronunciation of Arabic young learner taking into account our system configuration.

Since our corpus of data seems to be little small and the words pronunciation evaluated by the system were uttered by speakers with pronunciation problems we can not confirm at all that the GLL score decision is the most decision measure for mispronunciation detection. In such case, the learner (even has a pronunciation problems) may correctly pronounce the word and then this word is rejected by the system (i.e. falsely rejected). Thus, we have performed another experiment which was based on the same ASR system configuration explained in the section IV (subsection A). The only modification is that we have used another corpus of data which consists of four added speakers with good pronunciation. Those four speakers was asked to pronounce the three words “kataba”, “kourssi” and “chamsoun” as for the speakers who have pronunciation problems. In this experiment a specific threshold had been used to determine what is the performance of the system when using the GLL, LLL, ROS and ROA as scores for the evaluation of the pronunciation. For a more detailed analysis of performance four decisions types can be calculated (over the three pronounced words) for each score: correct acceptance (CA), correct rejection (CR), false acceptance (FA) and false rejection (FR). The threshold used must maximize the acceptance score accuracy (SA) defined as: \( SA = CA + CR \) and minimizing the rejection score (RS) defined as \( RS = FA + FR \). The next table shows the obtained results and figure 7 shows the plotted results.

<table>
<thead>
<tr>
<th></th>
<th>GLL</th>
<th>LLL</th>
<th>ROS</th>
<th>ROA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>43.33%</td>
<td>46.66%</td>
<td>16.66%</td>
<td>50%</td>
</tr>
<tr>
<td>CR</td>
<td>33.33%</td>
<td>20%</td>
<td>26.66%</td>
<td>3.66%</td>
</tr>
<tr>
<td>FA</td>
<td>16.66%</td>
<td>30%</td>
<td>23.33%</td>
<td>43.33%</td>
</tr>
<tr>
<td>FR</td>
<td>6.66%</td>
<td>3.33%</td>
<td>33.33%</td>
<td>3%</td>
</tr>
</tbody>
</table>

Table 2 the performance results of GLL, LLL, ROS and ROA in term of CA, CR, FA and FR
As we can see from the figure 7 the GLL score has the higher SA and the lower RS comparing to the other scores; which allow us to confirm that it is the most decision measure for mispronunciation detection.

6. Conclusion

We have presented in this paper our experiments where the goal was to provide system based on the use of automatic speech recognition technology that detects mispronunciation of Arabic young learner. The system was based on different score methods calculated over Arabic words spoken by learners with pronunciation problems. The results have shown that the system was able to detect mispronunciation, using the Global average Log Likelihood score method (GLL), with 86.66% of correct rejection. Another experiment was done on an expanding corpus of data which contain correct and mispronounced words. The results obtained have demonstrated that the GLL score, comparing to the other calculated scores, had the higher CA+CR (76.66%) and the lower FA+FR (23.32%) which can be considered as good predictor for Arabic mispronunciation detection.

References


