ASR based pronunciation evaluation with automatically generated competing vocabulary

Carlos Molina¹, Nestor Becerra Yoma¹, Jorge Wuth¹, Hiram Vivanco²

¹Speech Processing and Transmission Laboratory
Department of Electrical Engineering, Universidad de Chile, Santiago, Chile
²Department of Linguistic, Universidad de Chile, Santiago, Chile

Abstract

In this paper the application of automatic speech recognition (ASR) technology in CAPT (Computer Aided Pronunciation Training) is addressed. A method to automatically generate the competitive lexicon, required by an ASR engine to compare the pronunciation of a target word with its correct and wrong phonetic realization, is presented. In order to enable the efficient deployment of CAPT applications, the generation of this competitive lexicon does not require any human assistance or a priori information of mother language dependent errors. The method presented here leads to averaged subjective-objective score correlation equal to 0.82 and 0.75 depending on the task. 

1. Introduction

The role of speech technology in education has gained attention in the last years. Particularly, computer aided pronunciation training (CAPT) may become an interesting framework to massively apply speech technology. CAPT systems provide several advantages over conventional teacher-class methods: teachers could prepare lessons ad-hoc to a class or a student; students could practice pronunciation with a PC at home, lab or at any other place that could be much less stressful and embarrassing than a conventional class; and, the problem of low penetration of properly trained second language instructors in some regions is softened. Moreover, speech technology could potentially identify a specific error in an utterance, and then provide a pertinent feedback to correct it without any human assistance. In contrast, a human teacher can hardly give the proper assistance to more than one student at a time. Nevertheless, speech technology has its own limitations and, despite of several decades of intense research and development efforts, the plug-and-play scheme has continuously failed in this field.

Addressing the problem of pronunciation quality evaluation in CAPT is not a new topic. The following indicators are usually found in the specialized literature as pronunciation quality scores: duration; syllabic-timing; and, hidden Markov model (HMM) log-likelihoods [1][2]. Initially, those features or confidence metrics attempted to compare the observed signal with native and non-native models by making use of the forced Viterbi algorithm [2]. Also in [2] phoneme log-posterior score, based on Bayes classification rule for a single feature, leads to a higher correlation between subjective and objective evaluations than the ordinary features or confidence metrics by themselves.

The explicit use of ASR technology as a 1:N classifier in CAPT has recently been explicitly mentioned in [3][4]. In [3] and [4] the language model is generated by rules that allow taking into consideration the error hypothesis, deletion and overlapping of probable mispronunciation. The confidence score for evaluation pronunciation is based on phoneme duration analysis. It is worth mentioning that the vocabulary in ASR for CAPT has mainly been based on the incorporation of empirical rules. For instance, in [5] the competitive vocabulary is generated by means of rules based on training data with non-native speaker mistakes. Knowledge about typical mispronunciation is highly dependent on the system domain and the target pronunciation word. This vocabulary forces the simultaneous competition of the correct and wrong pronunciation and is crucial to make ASR technology successful in CAPT. Surprisingly, the problem of how to generate this competitive vocabulary by employing an automated method has not been addressed.

In this paper, an automatic generation of the competitive vocabulary hypotheses for ASR to address the CAPT problem is presented. Accordingly, this competitive vocabulary should be composed of a lexicon that makes possible the contrast of a target word or sentence, uttered by a student, with the correct and wrong pronunciations. The method does not require a priori modeling or information about common non-native pronunciation errors. Consequently, the presented automatic generation of the competitive vocabulary does not need any non-native speaker database. Moreover, the technique is text-independent and enables the efficient integration of didactic text material with ASR technology. Also, the experiments reported here were achieved with inexpensive desktop microphones. Finally, the authors believe that the successful application of ASR technology in socially relevant and potentially massive services should significantly improve the impact of speech technology in economy.

2. Pronunciation assessment with ASR

As it was mentioned above, an ASR engine can provide a high number of confidence measures more efficiently than 1:1 matching methods. Moreover, ASR in a CAPT system could also be employed to simultaneously establish a user-machine dialogue and evaluate quality pronunciation. Actually, entering answers by voice instead of clicking a choice with a mouse should certainly improve the usability of a CALL (Computer Aided Language Learning) system. This section presents the method to generate the competitive vocabulary and describes the confidence measures employed to assess pronunciation evaluation [6]. As discussed above, a testing utterance can be com-
pared with the competitive lexicon and the correct pronunciation by making use of ASR technology.

2.1. Competitive lexicon generation with distance measure between words

The idea is to generate a set of competitive words for each target word whose pronunciation is a student task. As a requirement, no previous analysis based on errors made by students could be employed in order to achieve an efficient integration of didactic material to ASR technology without human assistance. To comply with this requirement, a competitive vocabulary based on the distance between acoustic models is showed here (Fig. 1).

![Figure 1: Block diagram associated to the estimation of distance between words.](image)

This strategy allows establishing an objective metric that can define the competitive lexicon independently of the target pronunciation. First of all, the distances between a target word, whose pronunciation needs to be practiced, and words from a lexicon are estimated. The lexicon should be complete enough and representative of the target language in order to include a significant range of word distances. Second, the lexicon whose distance to the target word is within an interval defined by a minimum, $D_{\text{min}}$, and a maximum, $D_{\text{max}}$, thresholds is stored.

Then, the lexicon within the interval $[D_{\text{min}}, D_{\text{max}}]$ is sorted with respect to the distance to the target word and uniformly sampled to reduce the number of selected words to MNCW (Maximum Number of Competitive Words). This procedure attempts to find a trade-off between the accuracy of the pronunciation assessment and the limitation of the ASR technology: the higher the number of competing words, the more difficult the recognition task itself. Finally, a mother language pronunciation variant of the target word can be added. This pronunciation variant could be composed of: those acoustic models from the target language that are the most similar from the student mother language pronunciation variant; or, acoustic models trained with students mother language speech data. The metric between phoneme models is defined in the context of K-L (Kullback-Leibler) distance [7]. The alignment of models can be linear or non-linear.

2.1.1. Distance between two words

Consider that $w_x$ and $w_y$ are two words. Then, $w_x = \{\lambda_1^x, \lambda_2^x, ..., \lambda_{w_x}^x, ..., \lambda_{M_x}^x\}$ and $w_y = \{\lambda_1^y, \lambda_2^y, ..., \lambda_{w_y}^y, ..., \lambda_{M_y}^y\}$ correspond to the decomposition of $w_x$ and $w_y$, respectively, as sequences of triphone or monophone models: $M_x$ and $M_y$ are the number of models in $w_x$ and $w_y$, respectively; and, and denote triphone or monophone models $m_x$ and $m_y$ in $w_x$ and $w_y$, respectively.

Then, if words $w_x$ and $w_y$ are composed of different number of triphones or monophones, i.e. $M_x \neq M_y$, $D(w_x, w_y)$ is estimated by means of non-linear alignment of models. Assuming that $c(k) = \{m_x(k), m_y(k)\}$ is the alignment function determined by a standard DTW algorithm, then $D(w_x, w_y)$ can be written as:

$$D(w_x, w_y) = \frac{1}{K} \sum_{k=1}^{K} d\left(\lambda_{m_x(k)}^x, \lambda_{m_y(k)}^y\right)$$ (1)

where $d\left(\lambda_{m_x(k)}^x, \lambda_{m_y(k)}^y\right)$ is defined as the K-L distance between $\lambda_{m_x(k)}^x$ and $\lambda_{m_y(k)}^y$; and, $K$ is the length of the alignment path defined by $c(k)$.

2.1.2. Maximum number of competitive words

The presented CAPT method is based on using ASR to discriminate between correct and incorrect pronunciation of a given word or task. As mentioned above, finding a trade-off between the accuracy of the pronunciation assessment and the accuracy limitation of the ASR technology is required. This accuracy depends on the number of words that represents wrong acoustic versions of a given target word and that compete with the correct pronunciation in the Viterbi search.

Given a target word $w_t$, $W_t^{\text{comp}} = \{w_1^t, w_2^t, ..., w_T^t\}$ denotes the list of $T$ competitive words in the interval $[D_{\text{min}}, D_{\text{max}}]$, where $D_{\text{min}}$ and $D_{\text{max}}$ are distance thresholds. Also, components in $W_t^{\text{comp}}$ are sorting with respect to their distance to $w_t$. If $T$ is greater than MNCW, the words in $W_t^{\text{comp}}$ are uniformly sampled in order to pick MNCW words for the competitive lexicon. In contrast, if $T$ is lower than MNCW, no action is taken.

2.2. Including Spanish phonetic variant of target word

In order to improve the accuracy of the pronunciation quality evaluation, a variant of the phonetic realization of target word $w_t$ according to the students mother language (i.e. Spanish in this case) is included in competitive lexicon $W_t^{\text{comp}}$. This strategy attempts to incorporate information on users mother language without implementing a detailed study of pronunciation mistakes made by students. Two methods were tested to incorporate this phonetic variant: using the most similar phonetic decomposition using monophones or triphones in English, denoted as $w_{t-EP}$; and, employing the phonetic decomposition based on sub-phonetic units trained with speech material in the local language, denoted as $w_{t-SP}$.

2.3. Feature extraction from the ASRs results

As was mentioned above, ASR brings the possibility of efficiently extracting a high number of features. In this section, four confidence measures delivered by the ASR procedure are employed: word density confidence measure; recognition flag; and, N-best position.

2.3.1. Word density confidence measure in the logarithmic domain

Word density confidence measure in the logarithmic domain, $\log WD\text{CM}_t$, of a target word $w_t$ is defined according as:

$$\log WD\text{CM}_t = \frac{\sum_{r \in E(w_t, H)} \log \left( Q\left(h_r\right) \right)}{\sum_{l=1}^{K} \log \left( Q\left(h_l\right) \right)}$$ (2)

where $Q\left(h_r\right) = P\left(h_r\right)^\gamma \cdot P\left(O\left|h_r\right\rangle\right)$, $h_r$ is the $r$th hypothesis in the N-Best Viterbi list; $Q\left(h_r\right)$ is the likelihood score
given by the Viterbi search; \( P(h_r) \) is the language model probability of \( h_r \); \( P(O|h_r) \) is the observation probability of \( h_r \); \( \gamma \) is the acoustic model scaling factor; \( E(w_t, H) \) corresponds to the indices of the hypotheses where word \( w_t \) is contained; and finally, \( H \) denotes all the N-best alignments or hypotheses obtained from Viterbi decoding.

2.3.2. Recognition Flag

This binary confidence measure, denoted by \( RE_{C_i} \), is defined as:

\[
RE_{C_i} = \begin{cases} 
1 & \text{if } w_t \subset h_i \\
0 & \text{if } w_t \not\subset h_i 
\end{cases}
\]

(3)

where \( h_i \) is the first hypothesis in the N-Best Viterbi list.

2.3.3. Position in the N-best

Position in the N-best of target word \( w_t \), \( POS_t \), corresponds to the index of the most likely hypothesis where \( w_t \) is recognized:

\[
POS_t = \arg \max_r \{ |Q(h_r)| \mid r \in E(w_t,H) \}
\]

(4)

2.3.4. Log-likelihood ratio

This confidence corresponds to the follow ratio:

\[
\text{LogRatio}_t = \frac{\sum_{r \in E(w_t,H)} \log (Q(h_r))}{\sum_{r \in E(w_t,H)} \log (Q(h_r)) + \sum_{t \in E(w_{-SP},H)} \log (Q(h_t))}
\]

(5)

where \( E(w_{-SP}, H) \) corresponds to the indices of the hypotheses where the Spanish phonetic variant of word \( w_t \), \( w_{-SP} \), is contained.

3. Bayes based objective-subjective score mapping

As described in section 2.3, four word features or confidence metrics are associated to an utterance: LogWDCM, \( RE_{C_i} \), \( POS_t \) and LogRatio. The problem in pronunciation quality evaluation is how to map objective measures (i.e. confidence metrics) to subjective scores that emulate the opinion of a human instructor. Suppose that the subjective score is quantized in several competitive lexicons were tested by varying parameter MNCC that determines the number of words in each lexicon given a target word. Thresholds \( D_{\text{min}} \) and \( D_{\text{max}} \) were made equal to 8 and 25, respectively. There is also the addition of the Spanish pronunciation variant of the target word to the competitive language model as described in section 2.2: NON-SPA denotes a language model that does not include any Spanish pronunciation version of the target word; and, SPA-EP and SPA-SP indicate that the competitive lexicon incorporates the Spanish pronunciation version by making use of English and Spanish phonetic units, respectively. Four word features or confidence metrics are delivered by the ASR.

The main testing database is composed of 20 target words: Against, Behave, Boyfriend, Chocolate, College, Doesn’t, Example, Handsome, Hospital, Mouth, Scientist, Should, Special, Student, Television, Thirty two, Tourism, Vegetable, Vibration and Yesterday. These words were selected by experts in English language and phonetics in order to achieve a phonetically balanced evaluation data set. Then, examples of wrong and correct pronunciation of each word were recorded by the nine experts in English language using two different low-cost desktop microphones. The sampling frequency was equal to 16 kHz. Then, all the recorded data (1878 utterances) was re-labeled by the same experts in English language by applying two score pronunciation scale: acceptable (5) or unacceptable (1).

Moreover, further results with one additional testing database with learner speakers and non-seen vocabulary are provided to show the generalization ability of the presented approach. Thus, a group of five learner speakers (children aged 11 or 12 years old) recorded a testing database composed of four words: Gently, Counter, Healthy and Race. Each speaker uttered each word three or four times (72 utterances). Then, these utterances were also classified by the experts in English language with the same two levels of pronunciation quality described above. It is worth emphasizing that in both testing data the recording was done in less controlled environment from the additive noise and speaker-microphone relative position points of view. The same low-cost microphones mentioned above were employed to record this additional testing databases.

4. Experiments

The native American English acoustic models were trained with CSR-I WSJ0 corpus [8]. In CSR-I WSJ0 speech data was recorded with a high-quality microphone and the sample rate was equal to 16 kHz. All the training signals (20,055 utterances) were used to train CDHMMs. Also, LATINO 40 [9] was employed to train the Spanish phonetic units used in wt-SP described in section 2.2. This database is composed of continuous speech from 40 Latin American native speakers, with each speaker reading 125 sentences from newspapers in Spanish. The training utterances were 4500 uncoded sentences provided by 36 speakers and context-dependent phoneme HMMs were employed. The vocabulary is composed of almost 6000 words. Thirty-three MFCC parameters per frame were computed: the frame energy plus ten static coefficients and their first and second time derivatives. Cepstral Mean Normalization (CMN) was also employed. Each monophoneme and triphoneme was modelled with a three-state left-to-right topology without skip-state transition, with eight multivariate Gaussian densities per state with diagonal covariance matrices. Flat language model is adopted: it corresponds to the competitive vocabulary estimated as mentioned in section 2.1. The competitive lexicon for each target word evaluated here was chosen from the vocabulary that composes the CSR-I WSJ0 corpus. Moreover, several competitive lexicons were tested by varying parameter MNCC that determines the number of words in each lexicon given a target word. Thresholds \( D_{\text{min}} \) and \( D_{\text{max}} \) were made equal to 8 and 25, respectively. There is also the addition of the Spanish pronunciation variant of the target word to the competitive language model as described in section 2.2: NON-SPA denotes a language model that does not include any Spanish pronunciation version of the target word; and, SPA-EP and SPA-SP indicate that the competitive lexicon incorporates the Spanish pronunciation version by making use of English and Spanish phonetic units, respectively. Four word features or confidence metrics are delivered by the ASR.

The main testing database is composed of 20 target words: Against, Behave, Boyfriend, Chocolate, College, Doesn’t, Example, Handsome, Hospital, Mouth, Scientist, Should, Special, Student, Television, Thirty two, Tourism, Vegetable, Vibration and Yesterday. These words were selected by experts in English language and phonetics in order to achieve a phonetically balanced evaluation data set. Then, examples of wrong and correct pronunciation of each word were recorded by the nine experts in English language using two different low-cost desktop microphones. The sampling frequency was equal to 16 kHz. Then, all the recorded data (1878 utterances) was re-labeled by the same experts in English language by applying two score pronunciation scale: acceptable (5) or unacceptable (1).

Moreover, further results with one additional testing database with learner speakers and non-seen vocabulary are provided to show the generalization ability of the presented approach. Thus, a group of five learner speakers (children aged 11 or 12 years old) recorded a testing database composed of four words: Gently, Counter, Healthy and Race. Each speaker uttered each word three or four times (72 utterances). Then, these utterances were also classified by the experts in English language with the same two levels of pronunciation quality described above. It is worth emphasizing that in both testing data the recording was done in less controlled environment from the additive noise and speaker-microphone relative position points of view. The same low-cost microphones mentioned above were employed to record this additional testing databases.
4.1. Experiment data set 1
The evaluation data, EDB-1, was composed of utterances read by seven out of nine experts in English language. Transcription in score 5, correct pronunciation, was pronounced three times per each speaker. The whole procedure was repeated with the two low-cost desktop microphones. Altogether, 1458 utterances were recorded in EDB-1.

The testing database, TDB-1, was composed of 420 utterances pronounced by two out of nine experts in English language by basically repeating the procedure adopted for EDB-1. EDB-1 was employed to estimate the a priori p.d.f.s in (6) and TDB-1 was used to test the approach presented in this paper.

4.2. Experiment data set 2
As a second experiment, the presented method is tested with the recorded data provided by the five children, aged 11 or 12 years old, mentioned above. This database is denoted by TDB-2. The same a priori p.d.f.s in (6) estimated in sub-section 4.1 were employed.

5. Discussion
According to Fig. 2, if the Spanish pronunciation variant is not added to competitive lexicons, NON-SPA, the best subjective-objective score correlation is equal to 0.55 when MNCW = 25. Also in Fig. 2 subjective-objective score correlation as high as 0.67 and 0.77 are achieved with SPA-EP and SPA-SP, respectively, when MNCW = 10. Observe that the limitation of the ASR technology is also evident in the effect of MNCW on the subjective-objective score correlation. MNCW, and thresholds $D_{\text{min}}$ and $D_{\text{max}}$ attempt to include a trade off between the required discrimination between correct and wrong pronunciations, given a target word, and the limitations of the ASR technology. The competitive vocabulary is designed to contrast the target word with others in the Viterbi process. It improves the competitiveness in the search but it cannot be composed of words that are too similar phonetically.

![Figure 2: Averaged subjective-objective score correlations vs. MNCW in TDB-1 with POS classifier.](image)

The required data to estimate context-dependent models in the HMM training procedure is estimated. In this case, the threshold of minimum number of repetitions in the training database allowed for triphones was established in 5. Results with this limitation training are shown in Table 1. As can be seen, the classifier based on POS provides the best performance in TDB-1: subjective-objective score correlation equal to 0.82. Preliminary experiments with non-controlled pronunciation mistakes and non-seen vocabulary (TDB-2) is also presented in Table 1. As can be seen, the classifiers based on REC and POS can lead to a subjective-objective score correlation equal to 0.75. When compared with TDB-1, TDB-2 gives a subjective-objective score correlation 9% lower, which in turn strongly validates the generalization ability of the method. Notice that in TDB-2 the recording environment was less controlled than in TDB-1. It is worth mentioning that this experiment shows that the method presented here can handle unpredictable mispronunciation and non-seen lexicon.

Table 1: Averaged subjective-objective score correlations in TDB-1 and TDB-2 with MNCW = 10 and SPA-SP.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Average subjective-objective score correlation with TDB-1</th>
<th>Average subjective-objective score correlation with TDB-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogWDCM</td>
<td>0.78</td>
<td>0.68</td>
</tr>
<tr>
<td>REC</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td>LogRatio</td>
<td>0.74</td>
<td>0.40</td>
</tr>
<tr>
<td>POS</td>
<td>0.82</td>
<td>0.75</td>
</tr>
</tbody>
</table>

6. Conclusions
In this paper an automatic generation of ASR competitive text-dependent lexicon for CAPT without human assistance and without a priori information of mother language dependent errors is evaluated. Also, the results reported here were achieved with low-cost desktop microphones. The presented method can lead to subjective-objective score correlation equal to 0.82 and 0.75 with two levels of pronunciation quality depending on the task. The incorporation of techniques to improve the robustness to additive and convolutional noise are proposed as further research.

7. Acknowledgements
This work was funded by Conicyt-Chile under grants Fondecyt No. 1070382, Fondecyt No. D051-10243, and AT 24080121.

8. References