Evaluating different non-native pronunciation scoring metrics with the Japanese speakers of the SAMPLE Corpus

Vandria Álvarez Álvarez, David Escudero Mancebo, César González Ferreras, and Valentín Cardeñoso Payo

Department of Computer Science, Universidad de Valladolid, Spain

Abstract. This work presents an analysis over the set of results derived from the goodness of pronunciation (GOP) algorithm for the evaluation of pronunciation at phoneme level over the SAMPLE corpus of non native speech. This corpus includes several recordings of uttered sentences by distinct speakers that have been rated in terms of quality by a group of linguists. The utterances have been automatically rated with the GOP algorithm. The phoneme dependence is discussed to suggest the normalization of intermediate results that could enhance the metrics performance. As result, new scoring proposals are presented which are based on computing the log-likelihood values obtained from the GOP algorithm and the application of a set of rules. These new scores show to correlate with the human rates better than the original GOP metric.

Index Terms: Computer Assisted Pronunciation Training (CAPT), Goodness of Pronunciation (GOP), Hidden Markov Models (HMM), Automatic Speech Recognition (ASR), L2 Pronunciation [1]

1 Introduction

The Computer-Assisted Pronunciation Training (CAPT) has grown in the last years due to the necessity of L2 learners at improving their pronunciation using an automatic system. By using CAPT, students can benefit from continuous feedback without a teacher by their side all the time, providing a self-service way to practice [13].

The CAPT field growth has been almost parallel to the evolution of technologies, since computers and mobile devices computing capacity and portability has greatly increased. As pointed out in [12] the CAPT commercialization and work started in the earliest 2000's but it was not until 2007 when its relevance

[1] We would like to thank Ministerio de Economía y Competitividad y Fondos FEDER project key: TIN2014-59852-R Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española, and Junta de Castilla y Leon project key: VA145U14 Evaluación Automática de la Pronunciación del Español Como Lengua Extranjera para Hablantes Japoneses

showed up again. With this, the SLATE (Speech and Language Technology for Education) group was created. This group of research is dedicated to the development of education applications by means of automatic speech processing, natural language processing and in some cases spoken dialogue processing. These systems (SLATE/CAPT) make necessary multidisciplinary groups of researchers, from spoken language technologists, language teachers and experts, statisticians, among others [3].

According to several authors the error detection and teaching of pronunciation is a very hard job for CAPT systems, researchers major concern is to derive systems that are capable of identifying errors accurately and reliably to provide correct feedback [12,3,1,13] .

The Automatic Speech Recognition (ASR) often refers to technologies used in the detection and assessment of pronunciation errors, perception training, etc. [3]. The use of ASR started in the 1980s, but it was not useful for all speakers due to the acoustic differences among them and other related aspects. Then, ASR emerged in actual systems at the beginning of 2000s parallel to the advancements in computing technologies.

The evolution of ASR technologies during the last years has left encountered opinions among researchers about whether or not they are suitable for CAPT systems. In [9] is pointed out that probably the problems found in research are not due to mere ASR but also to the lack of familiarity with the ASR-based CAPT.

Another fact, is that ASR needs to be adapted for non-native speakers, ASRs developed specifically with native speech have demonstrated worst performance when tested with non-native [9]. To overcome this issue, usually the ASR engine is trained with both native and non-native speech. Knowing the users language (L1) is a must, since the problems that arise when learning a second language (L2) are different in every case and for that matter the ASR needs to understand these before providing feedback [3].

The recognition task also depends on the types of learning activities. Results won't have the same accuracy if there is a huge set of possible answers than when is limited to a small number. Also evaluating an ASR system scoring phase can be done by comparing the scores provided by the human judges for a given speech sample.

In this paper we present an experiment for scoring the quality of non-native speech with automatic methods. In section 2.1 we present the corpus of Japanese students of L2 Spanish that has been used in this study. Section 2.2 revises the literature of the state of the art measures for scoring L2 speakers. The application of these metrics to our corpus encouraged us to test other alternative metrics that are described in section 2.3. Results presented in section 3 show that the new metrics correlate better than the previous ones with the scores assigned by human evaluators to L2 students. We end with discussion and conclusions.

2 Experimental procedure

2.1 The SAMPLE corpus

In the framework of the SAMPLE research project, a corpus of spoken Spanish by non-native speakers was developed as a means to support future CAPT studies. The central part of the corpus includes a set of sentences and paragraphs selected from news database of a popular Spanish radio news broadcasting station. The texts cover various information domains related to everyday's life. They were obtained from the Glissando corpus, which was developed in connection to another project related to automatic prosodic labelling. The materials used in this study belong to the subset of prosodically balanced sentences in Glissando, which statistically resemble the prosodic variability found in Spanish [5].

The whole corpus is described in [2]. It contains different materials: read sentences, the Aesop's Fable 'The North Wind' and news paragraphs. In this study, fifteen read sentences from the news paragraphs of the Glissando corpus [5] were selected to be read by a group of non-native Spanish speakers. The list of sentences is described in [2] (see table 1 of that paper). All sentences followed a phonetic coverage criterion. In this study we only focus on the Japanese speakers for the sake of simplicity. These speakers are referred as *f11*, *m03*, *f12*, *f14* and *f13* in the database where *f* means female and *m* means male. The orthographic transcriptions used to identify the uttered phones in every sentence were realized by a group of linguists that collaborated with our research group.

The training of the phonetic models is based on a standard parameterization by using cepstral coefficients over mel frequencies (MFCC) and a 39 dimensions feature vector. More precisely, 12 MFCCs and the normalized power logarithm along with the first and second order derived. The features vectors are obtained with a time slot of 25 ms and time offset of 10 ms. The Albayzin corpus was used to train the acoustic mono-phoneme models since this contains recordings of phonetically balanced phrases [8]. Finally, all the sentences uttered by the 22 speakers of the SAMPLE corpus were computed by the forced alignement algorithm using HTK. The results present the logarithmic scores (likelihoods) of all existent phonemes for every expected phoneme in the utterances of the speakers.

2.2 Confidence measures

The confidence measures help determine the certainty of the recognizer when it identifies if an utterance (a part or all) was pronounced properly. Confidence measures can be computed with low difficulty by ASR engines and does not vary greatly among different sounds [11].

A confidence measure can be considered as a statistic that quantifies the fitting of a model with the corresponding data. For speech recognition acoustic and language models are typically used (together or separately) to extract these confidence measures [10].

Likelihood ratios convert to a useful statistic the outcome of HMM-based ASRs. The HMMs help finding the value of H, that maximizes the joint probability $P(X, H)$, where H is the acoustic model and X is the acoustic observation [10].

$$P(H|X) = \frac{P(X, H)}{P(X)} = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

The Bayes theorem is applied usually to calculate the relation between the joint probability with the posterior probability of the model H given the acoustics X, $P(H|X)$ and also the likelihood given the model H, $P(X|H)$.

There are different proposals that have showed good and well accepted results for auto matic scoring, next some of them extracted from the different previous work are listed, considering their results and continuous application.

The use of likelihood based phoneme level error detection started back in the 1990s [12]. The **log-likelihood** logarithm of the speech data is computed by Viterbi algo- hrithm using the HMMs from native speakers. According to [4,10] this is a good way for measuring similarity or matching between native speech and users speech.

The log likelihood score $\hat{l}$ for each phone segment is [6]:

$$\hat{l}_i = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log p(y_t|q_i) \tag{2}$$

where $p(y_t|q_i)$ is the likelihood of the current frame, y_t is the vector of observations, d is the duration in frames of the phone segment and t_0 is the starting frame index of the phone segment. Dividing over d eliminates the time duration of the phone, with this, the score is normalized. Also according to [4] the likelihood-based score for a whole sentence L is the average of the individual scores $L = \frac{1}{N} \sum_{i=1}^{N} \hat{l}_i$ with N, the number of phones in the sentence.

The **log-posterior probability** has showed better correlation results with the human judgments [6].

$$P(q_i|y_t) = \frac{p(y_t|q_i)P(q_i)}{\sum_{j=1}^{J} p(y_t|q_j)P(q_j)} \tag{3}$$

where $P(q_i|y_t)$ is the frame based posterior of the i-th phone given the observation vector y_t. $P(q_i)$ represents the prior probability of the phone class q_i. The sum over j operates on a set of context-independent models for all phone classes. The posterior score $\hat{\rho}$ for the i-th segment is the average of the logarithm over the frame-based phone posterior probability over all the frames of the segment, see equation.

$$\hat{\rho}_i = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \log P(q_i|y_t) \tag{4}$$

The complete sentence posterior-based score can be obtained with the average of all individual scores over the N phone segments in the sentence: $\rho = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i$

Compared to the likelihood metric in equation 2 the log-posterior probability score could be less affected since the acoustic matching to the models is in both numerator and denominator [4].

The **GOP** was first introduced byWitt with the purpose of providing an algorithm capable of scoring each phone of an utterance, therefore to accomplish this, the GOP must have previously the following data [14]: (1) The orthographic transcriptions previously annotated by human judges that describes exactly which is the phone sequence uttered. (2) The Hidden Markov models to calculate the likelihood, $p(O^{(q)}|q_j)$, where $O^{(q)}$ is the acoustic segment corresponding to each phone q^j.

Based on the latter the GOP for a phone q_i is computed by:

$$GOP(q_i) = |\log(P(q_i|O))|/NF(O^{q_i}) \tag{5}$$

Based on equation 5, the quality of pronunciation of any phone q_i can be obtained by normalizing the logarithm $P(q_i|O)$, which is the posterior probability that the speaker uttered the phone q_i over the acoustic segment O^{q_i} . The normalization takes place when is divided by the number of frames $NF(O^{q_i})$ in the acoustic segment.

As showed in equation 3 the log-posterior probability can be computed by knowing the likelihood of the acoustic observations given the phone q_i and the likelihood of the acoustic observations given the phone models. By applying this we get:

$$GOP(q_i) = \frac{1}{NF(O^{q_i})} \left| \log \left(\frac{p(O^{q_i}|q_i)P(q_i)}{\sum_{j=1}^{J} p(O^{q_i}|q_j)P(q_j)} \right) \right| \tag{6}$$

For equation 6 J is the total number of phone models for the j possible phones existent in the annotated database. If we assume that all phones are equally likely, meaning $P(q_i) = P(q_j)$ for all j and i, and that the sum of the denominator can be approximated by its maximum, then the GOP is equal to:

$$GOP(q_i) = \frac{1}{NF(O^{q_i})} \left| \log \left(\frac{p(O^{q_i}|q_i)}{\max_{j=1}^{J} p(O^{q_i}|q_j)} \right) \right| \tag{7}$$

The numerator of equation 7 is computed by using the forced alignment block where the sequence of phone models is fixed by the known transcription. On the other hand, the denominator is obtained by the phoneme loop, which realizes an unconstrained loop comparing the acoustic observations of the i-th phone with all the possible phonemes transcription [15].

A variant of the GOP utilizes the forced alignment block as in Witt allowing to set the phoneme boundaries. Once these boundaries are known the computation of the $p(O^{q_i}|q_j)$ (logarithmic scores) for all the j existent phonemes is realized. Thus, to obtain the GOP, is necessary to determine: (1) The annotated(orthographic transcription) utterances that will allow to determine which is the expected phoneme q_i; (2) All the log likelihoods for every phoneme $p(O^{q_i}|q_j)$.

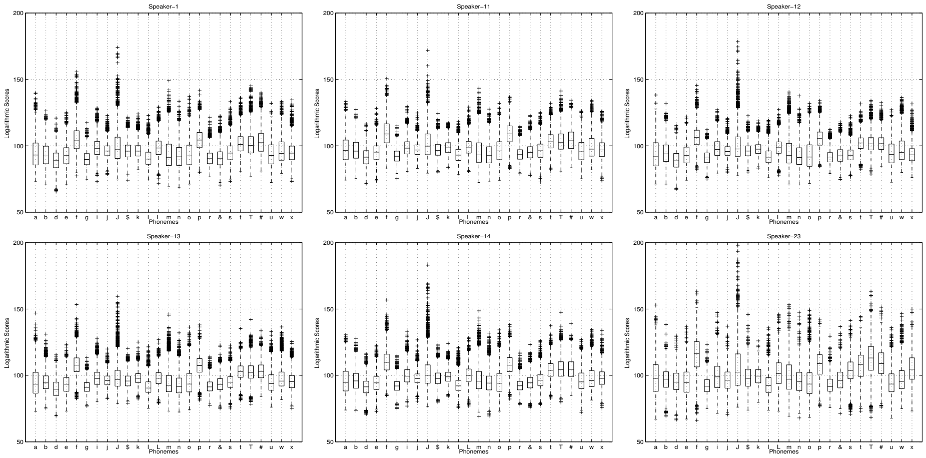

Fig. 1. From left to right and from top to bottom, the figures refer to the speakers m03, f11, f12, f13, f14 and to the tranning corpus Albayzin. Each of the boxplots corresponds with one single phoneme.

Afterwards, the computation of the GOP is realized by subtracting the expected phoneme q_i logarithmic score with the maximum logarithmic score among all phonemes, as in formula 7. This is the implementation that has been used in this work.

The implementation of the GOP metric has been widely used by several groups of researchers with different native and non-native languages. Nevertheless it exists also controversial due to the difficulties on adapting this score to different corpora. In the thesis of Witt [15], some thresholds calculations formulas are proposed to improve results. In order to distinguish between individual and systematic mispronunciations, Witt [13] used a recognition network that comprehended both correct pronunciation and common pronunciation errors ans sub-lattices for each phone. Another GOP variant proposed for the PRASER system [7] normalizes the GOP score by a sigmoid function. Similarly, the proposal from [11] modified the GOP metric focusing on the establishment of thresholds that best suited the data. In this work we also propose to adapt the metric to face up the dependences of the scores on the expected phoneme.

2.3 Alternative scoring proposals

Figure 1 shows that the logarithmic scores depict the same trend over the whole phonemes. Thus, for example, the values for the phoneme f, are in average higher than the values for the surrounding phonemes in the figures (phonemes e and g). This high dependency on the expected phoneme, which seems to be speaker independent, inspired us to normalize these scores before using them to rate the quality of the non-native pronunciation. Based on the latter we decided to create a set of new scorings by computing a new parameter obtained from the logarithmic scores and a set of rules explained next.

Case 1: The computation of a new score ρ for every expected phoneme q_i (that is evaluated in the corpus) is done by using the logarithmic score $\hat{\rho}_i$ associated to it and the average μ_i and standard deviation σ_i previously computed.

$$\underline{\hat{\rho}}_i = \frac{|\hat{\rho}_i - \mu_i|}{\sigma_i} \tag{8}$$

The values μ_i and σ_i are obtained by computing all the logarithmic scores of the expected phoneme when uttered by the same speaker and that also comply with some rules for choosing or not the logarithmic score of q_i. These are explained later for every sub-case.

Thus, there is a value of μ_i and σ_i for every phoneme and a given speaker on every sub-case studied, although as we will discuss later depending on the sub-case there are or not μ_i and σ_i values for all phonemes, and therefore some conditions are applied.

Case 1.a: it computes the μ_i and σ_i values over all the $\hat{\rho}_i$ that correspond to the expected phoneme q_i when its GOP value was equal to zero. As explained before the logarithmic score is only selected if we are analyzing the i-th phoneme.

For most of the speakers there are no values that can be used for every phoneme, because the speaker didn't uttered correctly the expected phoneme. To overcome this issue, we use the Albayzin values, since they are the reference and has μ_i and σ_i values for every phoneme. If only, one utterance is made for a phoneme then σ_i is changed to one to avoid division over zero.

Case 1.b: it is based on choosing the logarithmic scores that correspond to the expected phoneme q_i when this was uttered by the same speaker, there is no dependency on the GOP value.

Case 1.c: it is based on only choosing the logarithmic scores that correspond to the expected phoneme q_i when its GOP value is different from zero to compute μ_i and σ_i , in opposite to case 1.a.

Although, for this case there is at least more than one case in which a phoneme was uttered incorrectly then μ_i and σ_i can be obtained. In the worst scenario if only one utterance is made for a phoneme then σ_i is changed to one to avoid division over zero. Or if no μ_i and σ_i were obtained for a phoneme then the Albayzin values are used.

Case 2 The case 1 made use of the logarithmic score of the phoneme q_i. The case 2 on the other hand chooses the maximum logarithmic score among all the logarithmic scores calculated for all the phonemes in that specific utterance. With these values it is possible to calculate the average μ_i' and standard deviation σ_i' of all the maximum logarithmic scores for every expected phoneme.

$$\underline{\underline{\hat{\rho}}}_i = \frac{|\hat{\rho}_i - \mu_i'|}{\sigma_i'} \tag{9}$$

This Case 2 is sub-divided into two, whose difference is basically using or removing the absolute value from the numerator in equation 9.

SPK	PHO	DELE	GOP	CASE 1.a	CASE 1.b	CASE 1.c	CASE 2.a	CASE 2.b
f11	2.92	2.86	3.81	1.85	0.78	0.75	1.11	-0.70
f12	3.01	3.28	3.00	2.04	0.78	0.76	1.01	-0.58
m03	3.08	3.09	5.14	2.69	0.77	0.78	1.31	-1.15
f14	3.12	3.18	3.17	2.60	0.78	0.77	1.06	-0.65
f13	3.77	3.92	2.96	1.76	0.79	0.82	0.98	-0.58
Albaycin	-	-	0.87	1.00	0.75	1.61	0.84	-0.14

Table 1. Comparison of all the different scores analyzed.

Case 2.a use the formula 9 as it is and **Case 2.b** does not use absolute values in the numerator.

Once all these new scores have been obtained for every expected phoneme in the corpus, a global new pronunciation score is obtained for every speaker: $\underline{\underline{\rho}} = \frac{1}{N}\sum_{n=1}^{N} \underline{\underline{\rho}}_n$; where N is the total of all phonemes uttered by the given speaker.

3 Results

Table 1 shows the quality measures of the Japanese speakers of the Sample corpus. All the figures represent the mean values of the metrics computed by sentence. The row entitled Albayzin presents the metrics computed with the sentences of the training corpus so that it is a baseline that indicates the degree of quality of the non-native pronunciation.

The columns PHO and DELE are subjective metrics given by a team of human evaluators [2]. PHO indicate the phonetic quality and DELE is the overall quality. The index PHO has been used to sort the speakers in terms of his proficiency: the better the pronunciation of the speaker, the lower he or she is in the table.

The GOP results do not correlate with the subjective metrics so that the speaker m03 has GOP value (5.14) that is greater than the one obtained by apparently worse speakers like f11 and f12. The metric CASE 1.b and specially the metric CASE 1.c seem to correct this fact, with values that highly correlate with the subjective metrics: best and worst speaker (f13 and f11 respectively) are match with the ranking obtained by using the columns PHO and DELE.

Table 2 confirms this correlation when it is computed with all the non-native speakers of the corpus (American and Japanese ones). Case 2 metrics reproduce the behavior of the GOP metric getting to very high correlations of 0.88 and 0.99. Case 1.c seems to reproduce the best the subjective evaluation with a correlation of 0.81 with the PHO metric and 0.74 with the DELE metric.

4 Discussion and conclusions

The results presented in this paper are aligned with others cited in section 2.2 that have also observed the difficulties of the GOP index for correlating with

	PHO	**DELE**	**GOP**	**Case 1.a**	**Case 1.b**	**Case 1.c**	**Case 2.a**	**Case 2.b**
HJ PHO	1							
HJ DELE	0.87	1						
GOP	-0.35	-0.31	1					
Case 1.a	-0.26	-0.45	0.22	1				
Case 1.b	0.50	0.49	-0.34	-0.13	1			
Case 1.c	**0.81**	**0.74**	-0.03	-0.23	0.49	1		
Case 2.a	-0.41	-0.46	**0.88**	0.42	-0.38	-0.09	1	
Case 2.b	0.30	0.38	**-0.90**	-0.36	0.37	-0.01	-0.96	1

Table 2. Correlation coefficients among pronunciation scores at speaker level for non-native speakers of the SAMPLE corpus.

subjective evaluations. In this case we have contrasted this fact with Japanese speakers of L2 Spanish language. The modification introduced by Case 1.3 for taking into account the dependencies of the likelihood ratio with the predicted phoneme, has permitted to improve results so that the correlation between this metric and the subjective one increase up to 0.83.

In this paper we have shown that the computation of a metric permits to assign a score to non-native speakers that correlates with the marks assigned by human evaluators. Nevertheless, scoring non-native speakers is only one of the tasks that concerns CAPT techniques. A straight forward extension of our results is the detection of errors which be identified by prominent values of the scores over phonemes. It is current work of our research team the implementation of a diagnosis module that permits to specify the speaker the type of error and consequently advise him or her. A last phase of feedback comprehends the design issues related on how to present results for the speakers to improve their pronunciation.

References

1. van Doremalen, J., Cucchiarini, C., Strik, H.: Automatic pronunciation error detection in non-native speech: The case of vowel errors in dutch. The Journal of the Acoustical Society of America 134(2), 1336–1347 (2013)
2. Escudero-Mancebo, D., González-Ferreras, C., Cardeñoso Payo, V.: Assessment of non-native spoken spanish using quantitative scores and perceptual evaluation. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 3967–3972. European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014), http://www.infor.uva.es/ descuder/investig/pdfs/LREC2014Assessment.pdf
3. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10), 832–844 (2009)
4. Franco, H., Neumeyer, L., Kim, Y., Ronen, O.: Automatic pronunciation scoring for language instruction. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. vol. 2, pp. 1471–1474. IEEE (1997)

5. Garrido, J.M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., de-la Mota, C., González, C., Vivaracho, C., Rustullet, S., Larrea, O., Laplaza, Y., Vizcaíno, F., Estebas, E., Cabrera, M., Bonafonte, A.: Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. Language Resources and Evaluation 47(4), 945–971 (2013)
6. Kim, Y., Franco, H., Neumeyer, L.: Automatic pronunciation scoring of specific phone segments for language instruction. In: Eurospeech (1997)
7. Mak, B., Siu, M., Ng, M., Tam, Y.C., Chan, Y.C., Chan, K.W., Leung, K.Y., Ho, S., Chong, F.H., Wong, J., et al.: Plaser: pronunciation learning via automatic speech recognition. In: Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2. pp. 23–29. Association for Computational Linguistics (2003)
8. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Marino, J.B., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: Third European Conference on Speech Communication and Technology (1993)
9. Neri, A., Cucchiarini, C., Strik, W.: Automatic speech recognition for second language learning: how and why it actually works. In: Proc. ICPhS. pp. 1157–1160 (2003)
10. Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M.: Automatic scoring of pronunciation quality. Speech communication 30(2), 83–93 (2000)
11. Strik, H., Truong, K., De Wet, F., Cucchiarini, C.: Comparing different approaches for automatic pronunciation error detection. Speech Communication 51(10), 845–852 (2009)
12. Witt, S.M.: Automatic error detection in pronunciation training: Where we are and where we need to go. Proc. IS ADEPT 6 (2012)
13. Witt, S.M., Young, S.J.: Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication 30(2), 95–108 (2000)
14. Witt, S.M., Young, S.J., et al.: Language learning based on non-native speech recognition. In: Eurospeech (1997)
15. Witt, S.M.: Use of speech recognition in computer-assisted language learning. University of Cambridge (1999)