

Using an ASR database to design a pronunciation evaluation system in Basque

Igor Odriozola, Eva Navas, Inma Hernaez, Iñaki Sainz, Ibon Saratxaga, Jon Sánchez, Daniel Erro

Aholab – Dep. of Electronics and Telecommunications. Faculty of Engineering
University of the Basque Country, Urkijo zum z/g 48013 Bilbo
E-mail: {igor, eva, inma, inaki, ibon, ion, derro}@aholab.ehu.es

This paper presents a method to build CAPT systems for under resourced languages, as Basque, using a general purpose ASR speech database. More precisely, the proposed method consists in automatically determine the threshold of GOP (Goodness Of Pronunciation) scores, which have been used as pronunciation scores in phone-level. Two score distributions have been obtained for each phoneme corresponding to its correct and incorrect pronunciations. The distribution of the scores for erroneous pronunciation has been calculated inserting controlled errors in the dictionary, so that each changed phoneme has been randomly replaced by a phoneme from the same group. These groups have been obtained by means of a phonetic clustering performed using regression trees. After obtaining both distributions, the EER (Equal Error Rate) of each distribution pair has been calculated and used as a decision threshold for each phoneme. The results show that this method is useful when there is no database specifically designed for CAPT systems, although it is not as accurate as those specifically designed for this purpose.

Keywords: CAPT system, GOP score, ASR database

1. Introduction

Speech technologies are becoming more and more important in the second language (L2) learning field, especially in systems aimed at teaching pronunciation, the so called CAPT systems (*Computer-Assisted Pronunciation Training*). These techniques have gained interest among researchers, among companies involved in the integration of new technologies into L2 tutoring systems and even among L2 teachers and students. One of the advantages of these systems is that students can learn and practise by their own, at their own pace, thus complementing the learning process and refining what has been learned in the classroom.

The first scoring methods in the early 90s were focussed on the word and phrase level (Hamada et al., 1993; Hiller et al., 1993; Rogers et al., 1994). In the late 90s systems designed to teach selected phonemic errors were developed (Kawai and Hirose, 1997; Kim et al., 1997; Ronen et al., 1997), where models trained on non-native speech were used. In that period, the use of Automatic Speech Recognition (ASR) with HMMs to score each phone in an utterance was described as well (Eskenazi, 1996).

In the last years the development of CAPT systems has focused on integrating multimedia technologies to give a more complete feedback to the user. In this context, the state of the art in the PC-based pronunciation teaching systems includes Pronunciation Power, American Sounds, Phonics Tutor and Eyespeak (*LearningVillage*, 2009). Recently, the *Euronounce* European project has also been developed focused to create a multimodal teaching application (Demenko, G. et al., 2009). All these systems allow the learners to use analysis and recognition of the input speech by means of recording their own utterance. This recording is then displayed as spectrum thus enabling acoustic and visual comparison between the learner's and the reference voice's utterances. In most of the systems, a three colour scheme is used to indicate the correctness of a phone, as in Plaser (Mak et al., 2003).

The method of automatic pronunciation scoring most used as a 'Goodness of Pronunciation' (GOP) measure in most of the applications is the frame-based posterior

probability of a phoneme. For each acoustic segment X_u of phoneme y_u (u denotes the phoneme index):

$$GOP(y_u) = \log \Pr(y_u | X_u) \approx \frac{1}{T_u} \cdot \log \left[\frac{p(X_u | y_u) p(y_u)}{\sum_{k=1}^N p(X_u | y_k) p(y_k)} \right] \approx \frac{1}{T_u} \cdot \log \left[\frac{p(X_u | y_u)}{p(X_u | y_{j_{max}})} \right] \quad (1)$$

where N is the number of phonemes and j_{max} is the phoneme model that gives the highest likelihood for the given segment. In practice, the denominator in equation 1 is replaced by the Viterbi likelihood of the segment given by a phone loop. Some works show that this is a good measure (Witt & Young, 2000; Franco et al., 2000), though some refinements are used to further improve the scores (Kanters et al., 2009). The GOP is used to set a threshold to decide if the phoneme is pronounced correctly, but usually, instead of a global GOP, a different score is used for each phoneme or phoneme group.

These GOPs are calculated considering the same standard dictionary pronunciation for both correct utterances and utterances with some kind of errors (from now on: *correct utterance* vs. *erroneous utterance*). The erroneous utterances are usually defined beforehand by linguistic experts, and databases are designed specifically to show the most samples as possible of these phones to be analyzed. Then, two score distributions are obtained for each phone (or group): one created using correct utterances and the other using erroneous inputs, and thus an EER (*Equal Error Rate*) threshold can be fixed.

If there is no such database for a language, as in the case of Basque, we propose a solution in this paper that can be useful to obtain a distribution of the verification scores for erroneous utterances, by means of using a general ASR database. This solution consists of inserting changes (or controlled errors) in the pronunciation dictionary itself. One phone is replaced in each word of the dictionary by another phone of the same group (the groups are automatically obtained by performing a phonetic clustering), and then the distribution of the scores obtained for these phones is used as the distribution of the erroneous utterances.

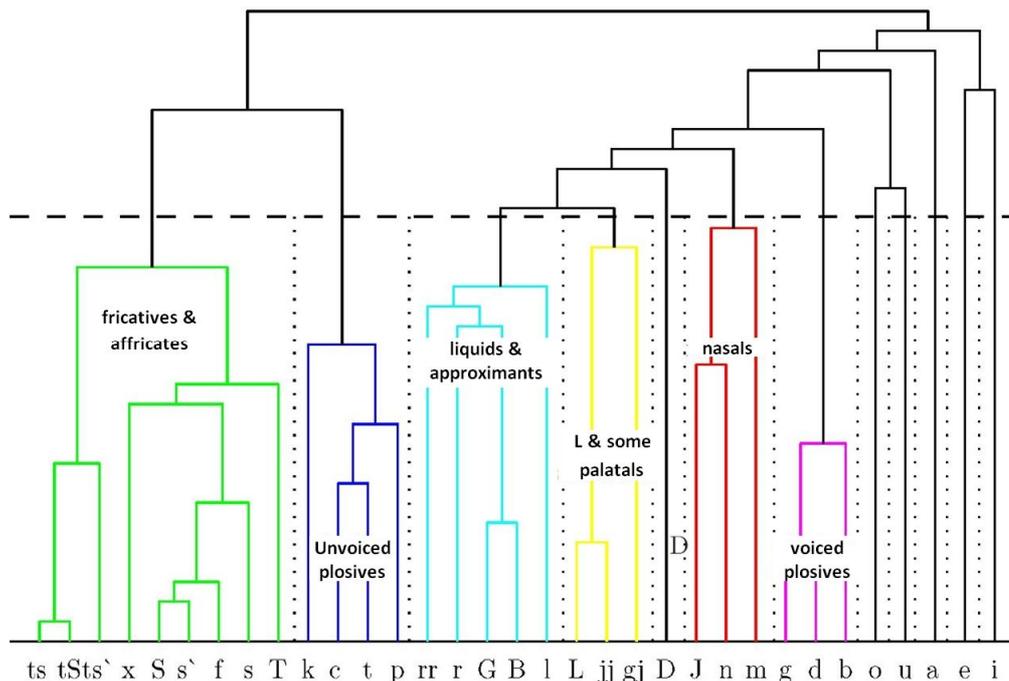


Figure 1: Dendrogram of the phoneme clusters

The aim of this paper is to show how to use an ASR database to develop CAPT systems for under-resourced languages and to evaluate how GOP scores behave in this environment, i.e. if we use a database not specifically designed for pronunciation evaluation. The rest of the paper is organized as follows: section 2 describes the database used in the experiments detailed in section 3 and section 4 presents the conclusions of the work.

2. Database

CAPT systems use non-native vs. native recording data to better assess the learners' utterances, so that they all rely on specifically designed databases with both kinds of speech. Besides, most of the systems emphasize some problematic phonemes that are defined beforehand by experts for each of the different L1-L2 pairs. For that purpose, properly designed databases are built for the most usual L1-L2 language pairs (Cylwik et al., 2008).

Basque is an under-resourced language that does not have enough acoustic databases to develop speech technologies that could compete with the major languages. Although there are some speech recognition databases for Basque, the only one which is publicly available (Hernaiz, et al., 2003) was recorded over the fixed telephone network. Currently the priorities do not include the creation of a CAPT oriented database, and pronunciation teaching systems have to be developed with other available data.

The database that had been used in these experiments was designed for the training and development of speech recognition for Basque and at the moment is not publicly available. It is a Speecon-like database, which contains recordings from native and non-native speakers, as well as dialectal and standard Basque data for the formers. It contains data from a total of 230 speakers, collected in different places of the Basque Country, where Basque has different official status, health and phonetic influence of neighbouring languages (mainly French and Spanish).

During the recording of the database the speakers were asked about their knowledge level of the language. The native speakers' subcorpus is composed by 149 speakers; non-native speakers speak Basque as a second language at different levels: the high level non-natives' subcorpus includes 56 speakers and the low level non-natives' subcorpus 25. Due to the dialectal variation and also to the different level of fluency, there are several irregularities in the pronunciation of several phonemes, which are not labelled in the transcription. However, they can be partially deduced from the information provided about the speaker. For example, we can obtain information about the region of the speakers and their Basque level through the labels that indicate their city of birth, city of youth and language level. The audio files have their corresponding orthographic transcription file, and the rule-based P2G transcriber for Basque has been used to obtain the standard Basque phonetic transcriptions.

A mean of 170 files have been used per speaker. 60 of these files contain elicited speech, where the speaker is asked to read different types of texts (dates, numbers, phonetically rich sentences and the like), and the remaining 110 are commands, which are composed mainly by isolated words.

3. Experiments

The strategy that has been chosen in these experiments is the following: first of all, context-dependent (triphones) HMMs have been trained with two thirds of the database, and a monophone loop has been implemented to run in parallel in order to obtain the GOP as defined in equation 1. The first experiment has been to obtain the GOP scores for each phoneme, carrying out a forced alignment in parallel with the phoneme-loop. For the forced alignment step, two possibilities have been taken into account for the subsequent words: coarticulation between them and separation by a silence.

After obtaining a set of scores for each phoneme, transcription errors have been inserted in each dictionary transcription, by selecting one triphone randomly and replacing it by another triphone of the same phoneme group, selected also in a random way. To obtain these groups a GMM model of one mixture has been trained for each phoneme using Basque SpeechDat-EU (Hernaez, et al., 2003) and a phonetic clustering has been performed using regression trees. Figure 1 shows the dendrogram with the obtained phoneme clustering.

When making the phoneme substitutions in the transcriptions, the context-dependency has also been changed, thus keeping the coherency in the transcriptions. Then, a new forced alignment experiment has been carried out to obtain the GOP scores for each inserted phoneme; in other words, the GOP scores corresponding to a phoneme in a simulated erroneous utterance are obtained.

Since the amount of scores in simulated erroneous utterances is much smaller than those for correct utterances the process of random triphone substitution was repeated three times, in order to obtain more data. Then, the histograms of both distributions are calculated for each phoneme, joining all the scores corresponding to the same phoneme. An example of the distribution of the scores corresponding to the phoneme /a/ can be seen in Figure 2, where both separate density functions appear. Note that the phonemes in this paper are represented by their SAMPA code¹.

At first glance, it was observed that there were some potentially problematic phonemes with overlapping distributions. This could be due to the differences in the pronunciation of some phonemes between native and non-native speakers, especially in those phonemes that do not exist in the L1 of the speaker, as the /ts/ or the /s/. Since the HMMs had been trained ignoring this fact, new HMMs were built again, using only the native speakers that were born and live in the east area of the country, since in the west area the /s/ phoneme is not pronounced today —it merged with the /s/—. So, 76 eastern Basque speakers have been used to train the new HMMs and 13 eastern Basque speakers have been reserved for testing

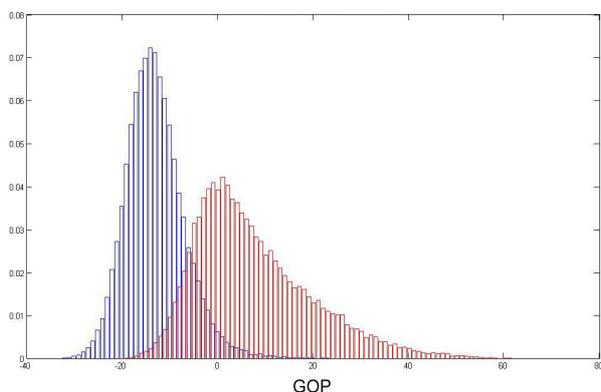


Figure 2: Normalized histograms of the GOP scores of the phoneme /a/: the blue bars represent the distribution of scores in a correct utterance; the red bars represent the distribution of scores in a simulated erroneous utterance

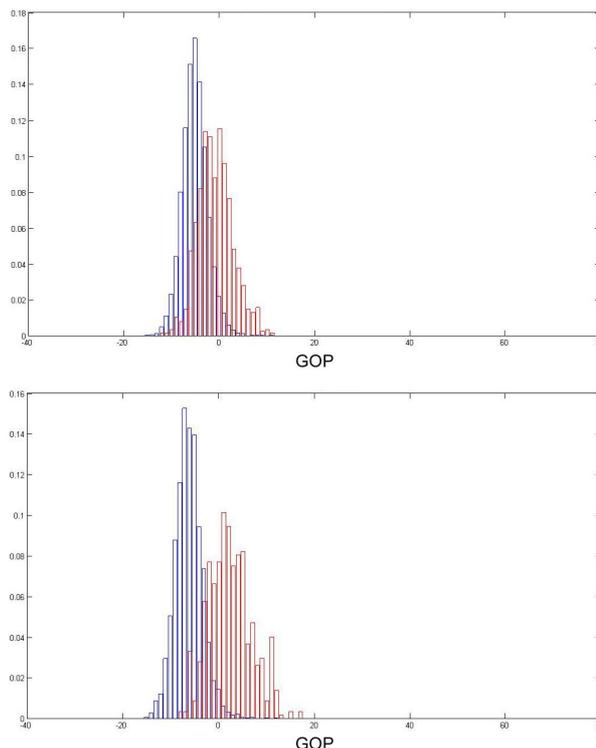


Figure 3: Normalized histograms of the GOP scores of the phoneme /ts/ obtained with HMMs trained using all kind of speakers (top) and using only native speakers' data (bottom)

purposes. With these new HMMs all the procedure was repeated, and new score distributions were obtained. A significant separation between the distributions of the problematic phonemes was observed. An example is shown in Figure 3.

The next step was to set the thresholds that we will use to classify each phoneme as correctly or wrongly pronounced. These thresholds are obtained by calculating for each phoneme the EER derived from the GOP score distribution functions.

Three different tests were performed in order to evaluate the designed strategy. In all the tests, the GOP scores of all the phonemes have been calculated using a forced alignment procedure and have been compared with the thresholds calculated beforehand.

In a first experiment (Test 1) the tests were performed using the recordings of the 13 native speakers reserved beforehand. The transcriptions of the speech corresponding to these speakers are not checked, but we assume that there is no error in these transcriptions and that all the speakers pronounced all the phonemes properly. So, we assume that the phonemes that the system evaluates as correct are really correct (CA, *correctly accepted*) and the phonemes that the system evaluates as incorrect are considered false rejects (FR, *falsely rejected*).

A second test (Test 2) was performed using the recordings of the speakers with “low level” skills in L2, 25 in total. Even though we cannot predict a priori if a phoneme was correctly or incorrectly pronounced, the tests were carried out without having into account any kind of linguistic knowledge and labelling, in order to see the performance of the system at first sight.

¹ http://aholab.ehu.es/sampa_basque.html

The results have been evaluated using the SA measure (*Scoring Accuracy*):

$$SA = ((CA+CR)/(CA+CR+FA+FR))*100; \quad (2)$$

where: CA: *Correctly Accepted*; CR: *Correctly Rejected*; FA: *Falsely Accepted*; FR: *Falsely Rejected*. In these experiments (Test1 and Test2) CR is 0.

		/a/	/u/	/ts̺/	/s̺/
Test 1	Realiz.	5.524	1.937	750	1.317
	SA	86,22 %	89,67 %	83,73 %	74,26 %
Test 2	Realiz.	9.923	3.481	1.438	2.469
	SA	84,06 %	87,33 %	41,59 %	49,49 %

Table 1: Number of realizations and SA for the phonemes /a/, /u/, /ts̺/ and /s̺/ in tests 1 and 2.

As it is shown in Table 1, the SAs of the phonemes /a/ and /u/ have a small variation from one test to the other, but the SAs of the /ts̺/ and /s̺/ (phonemes that do not exist in Spanish) fall significantly, especially in the case of the /ts̺/. This means that the realizations of some phonemes by Basque learners are very close the native speakers' realizations, while other phonemes are pronounced in quite a different way.

Finally, we focused the next experiment on the phonemes that show a worst SA in Test 2. A linguist expert labelled the realizations of these phonemes of 14 speakers from the subcorpus of the low level learners as correct or not correct. In the case of the /ts̺/ there are 813 realizations, where 375 have been labelled as correct and 438 as not correct. For the phoneme /s̺/, 1.348 realizations have been labelled; 720 were labelled as correct and 628 as not correct. Considering these labels, Test 3 was performed, and the results of Table 2 were obtained.

Test 3	ts̺`	s̺`
CA	32,84 %	33,97 %
CR	43,67 %	29,15 %
FA	10,21 %	17,43 %
FR	13,28 %	19,43 %
SA	76,51 %	63,13 %

Table 2: Results of comparing the automatically generated labels with the labels assigned by a linguist expert, for phonemes /ts̺/ and /s̺/.

We can see that the SAs of Test 3 (Table 2) are now closer to the ones obtained in Test 1 (Table 1) compared with the results of Test 2. This confirms the existence of incorrectly pronounced phonemes in the database. The results of the phoneme /ts̺/ show that the distribution separation that we have obtained training HMMs with only native speakers' data (shown in Figure 3) has been useful to calculate decision thresholds. The less accurate results for the phoneme /s̺/ can be interpreted taking into account that nowadays this sound does not exist in a big area of the Basque Country, so that even some native speakers of the eastern side may be influenced by this

fact.

From a general point of view, the results show that our system is able to detect both correctly pronounced phonemes in an utterance and incorrectly pronounced ones, focusing especially on phonemes that do not exist in the L1 of the learners. We have found that there are great differences between the skills of the speakers of the "low level" subcorpus, maybe because the classification of the database comprises only two skill levels: low and high. The system, as expected, discerns much better the incorrect pronunciations of the lowest level speakers. Comparing the results with those obtained in (Kanters et al., 2009), we have achieved smaller SAs for the phonemes that do not exist in the L1 of the speaker, but better results for phonemes that the speaker already knows from his/her native language. So, we can consider that the mean SA is higher than those corresponding to the conflictive phonemes. Further analysis is needed to check the exact overall value of SA, but, as a conclusion, the results validate the followed strategy to obtain the thresholds for the decisions.

4. Conclusion

A new method to use a general purpose ASR speech database to build CAPT systems for under-resourced languages has been described in this paper. A method to automatically determine the threshold of goodness of pronunciation (GOP) has been proposed. This method uses a strategy of introducing controlled errors in the transcriptions of the dictionary words, in order to obtain the distribution of the GOP scores of incorrectly pronounced phonemes. Different experiments have been described using different sets of speakers to test the system, and the results show that this method can be very useful to evaluate the utterances of learners with lowest levels.

The final goal of this work is to include it in a CAPT system for Basque that is currently under development. Future works will analyze how these GOP score distributions (and, therefore, the thresholds) behave in real conditions channels. Under these circumstances, the possibility of adaptation of the HMM models will be an important point to develop. Future works will analyze, as well, the needs to deal with the pronunciation errors of the L2 learners of higher level.

5. Acknowledgements

This work has been partially supported by the Basque Government (Berbategi, IE09-262), the University of the Basque Country (UE-08/11) and the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02).

6. References

- Cylwik, N, Demenko, G, Jokisch, O., Jäckel, R., Rusko, M., Hoffmann, R., Ronzhin, A., et al. (2008). The use of CALL in acquiring foreign language pronunciation and prosody—general specifications for Euronounce Project. In Proc. SASR, pp. 123-130.
- Cylwik, N., Wagner, A., Demenko, G. (2009). The Euronounce corpus of non-native Polish for

- ASR-based pronunciation tutoring system. In *SlaTE*.
- Demenko, Grazyna, Cylwik, Natalia, & Wagner, A. (2009). Applying speech and language technology to foreign language education. In *2009 International Multiconference on Computer Science and Information Technology*, 2, pp. 457-463.
- Eskenazi, M. (1996). Detection of foreign speakers' pronunciation errors for second language training – preliminary results. In *ICSLP'96*. Philadelphia, PA, USA.
- Franco, H. (2000). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(2-3), pp. 121-130.
- Hamada, H., Miki, S., Nakatsu, R. (1993). Automatic evaluation of English pronunciation based on speech recognition techniques. *IEICE Trans. Inform. Syst.* E76-D (3), pp. 352-359.
- Hernaiz, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., Sanchez, J. (2003). The Basque speech_dat (II) database: a description and first test recognition results, In *Eurospeech-2003*, pp. 1549-1552.
- Hiller, S., Rooney, E. Laver, J., Jack, M. (1993). SPELL: An automated system for computer-aided pronunciation teaching. *Speech Communication* 13, 463-473.
- Kanters, S., Cucchiari, C., Strik, H. (2009). The goodness of pronunciation algorithm: a detailed performance study, In *SlaTE-2009*, pp. 49-52.
- Kawai, G., Hirose, K. (1997). A CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruent. In *Proc. EUROSPEECH'97*. Rhodes, Greece.
- Kim, Y., Franco, H., Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. In *Proc. EUROSPEECH'97*. Rhodes, Greece.
- Learning Village. Educational Software Review, Retrieved on 15th July 2008 from <http://www.learningvillage.com/html/guide.html>
- Mak, B., Siu, M., Ng, M., Tam, Y.-cheung, Chan, Y.-chung, Chan, K.-wah, Leung, K.-yee, Ho, S., Chong, F.-ho, Wong, J., Lo, J. (2003). PLASER: Pronunciation Learning via Automatic Speech Recognition. In *HLT-NAACL Workshop on Building Educational Applications using Natural Language Processing*.
- Rogers, C., Dalby, J., DeVane, G. (1994). Intelligibility training for foreign-accented speech: A preliminary study. *J. Acoust. Soc. Amer.* 96 (4), pt. 2.
- Ronen, O., Neumeyer, L., Franco, H. (1997). Automatic detection of mispronunciation for language instruction. In *Proc. EUROSPEECH'97*. Rhodes, Greece.
- Witt, S., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3), pp. 95-10