Design of a message verification tool to be implemented in CALL systems

Igor Odriozola[1], Inma Hernáez[1], Eva Navas[1]

[1] Aholab signal processing laboratory, Urkixo zumarkalea z/g, Bilbao, Basque Country
{igor,inma,eva}@aholab.ehu.es

Abstract. This paper presents basic research for the development of a message verification system in Basque to be implemented in CALL (*Computer-Assisted Language Learning*) applications. The system aims to verify a sentence uttered by the user in real time, word by word, in order to display the verified word as soon as it is detected. First a decision threshold for the PS (*Phoneme Score*) is calculated by means of inserting artificial errors in the system. Then, the structure of the ASR internal lattice is described, which includes a phoneme loop between words to absorb the effects of unexpected speech. A last experiment has been carried out to check the behavior of the whole system simulating the insertion of an erroneous extra word by the user. The results of the experiments show that the proposed system is suitable for message verification tasks.

Keywords: CALL systems, utterance verification (UV), message verification, PS scores, L2 acquisition

1 Introduction

The use of spoken language technology for language learning started in the late 1970s, but nowadays the so called CALL (Computer-Assisted Language Learning) systems are witnessing a great development due to the fact that speech technologies have made unprecedented progress and have achieved some kind of stability. CALL applications include a wide range of ICT applications, from the “traditional” drill-and-practice programs of the 1960s and 1970s to the recent applications in virtual learning environments and web-based distance learning. CALL also extends to the use of interactive whiteboards [1], computer-mediated communication (CMC) [2], language learning in virtual worlds, and mobile-assisted language learning (MALL) [3].

Naturalistic, implicit learning is not always enough to achieve high-quality L2 proficiency, according to second language (L2) acquisition research theories. Explicit instruction helps to overcome some of these learning problems [4][5]. This is the reason why nowadays software for CALL systems includes all types of material, mostly audiovisual. Nevertheless, the actual turning point in which CALL systems began to be more useful and sophisticated and came into more extended use was the implementation of speech technologies in them, especially automatic speech recognition (ASR). This fact allowed the interaction with users and the provision of automatic feedback, something absolutely essential in the process of learning a language. The modalities of interaction with the student comprise detection and assessment of pronunciation errors, perception training and use of talking heads, and detection and correction of prosody errors [6]. ASR systems are being used, above all, for pronunciation assessment; however, there are many other applications designed specifically over an ASR to provide intelligent feedback on important aspects of L2 learning such as morphology and syntax [7, 8], by means of using utterance verification techniques along with a predefined list of possible (correct and incorrect) responses for each exercise.

In this paper an ASR-based message verification system for Basque is presented. In this system the user's utterance is checked word by word in real time. The system displays a positively verified word in the same instant that it has just been uttered; otherwise, it waits until the expected correct word arrives. This system shows to be useful for exercises or tasks where the user is intended to choose or create an answer that has a strict word order, such as reordering sentences, answering questions and so on. Although the verification process can be carried out at word level or phoneme level, in this paper word level analysis is presented.

The paper is organized as follows: after the introduction, the theoretical basis in which the message verification system relies on is described. Then, several experiments and their corresponding results are presented. Finally, some conclusions and a reflection about development, improvement and future work are presented.

2 The basis of the system

2.1 The database

Currently there is no suitable database for the development of CALL systems in Basque. Regarding ASR databases, the only one that is publicly available for Basque is a SpeechDat database recorded over the fixed telephone network [14], at 8 kHz and 16 bits. As the recording conditions of this database are very different of the ones that can be expected in CALL systems, this database was not appropriate for our experiments. The database selected for the experiments in this paper, which at the moment is only available for research, is a Speecon-like one, recorded at 16 kHz and 16 bits using one headset and one desktop microphone. In the experiments presented in this paper, only the part recorded by means of the headset microphone has been used. The database contains recordings from 230 speakers, both native and non-native, as well as dialectal and standard Basque data for the formers. The native speakers' subcorpus is composed by 149 speakers, and the non-natives' subcorpus includes 81 speakers who speak Basque as L2 at different levels. All this information is labeled and can be easily extracted from the textual data files.

The audio files have associated their corresponding orthographic transcription file, and a rule-based P2G transcriptor for Basque has been used to obtain their standard Basque phonetic transcription. The HMMs were trained using only the subcorpus of native speakers, leaving the non-native speakers' part for future research about improvements and adjustments for real implementations of CALL systems. Two thirds of the native speakers were used to train the acoustic models and the remaining third for testing. 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.

2.2 The ASR and calculation of PS scores

The message verification system has been built on the ASR-based utterance verification system developed at the Aholab Signal Processing Laboratory [9]. The system relies on a standard ASR based on Hidden Markov Models (HMMs). It processes 16 kHz signals with 16 bits and extracts vectors of 39 MFCC (Mel-frequency cepstral coefficients) —including first and second derivatives— from 25 ms duration frames each 10 ms. The decoding process is carried out by means of the Viterbi algorithm over an HMM lattice. The HMMs are context-dependent (triphones).

The verification process needs a second HMM lattice running in parallel, so that the *Goodness of Pronunciation* (GOP) scores of a phoneme y_u are computed as its posterior probability, over the acoustic segment X_u provided by the Viterbi decoder (u denotes the phoneme index). The parallel lattice consists of a free loop of context-independent HMMs, in order to avoid excessive increase of the processing time. Thus, equation 1 is used to calculate the GOP score of a phoneme:

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

where N is the number of phonemes and j_{max} is the phone model that gives the highest likelihood for the given segment. The denominator in equation 1 is replaced by the Viterbi likelihood of the segment given by the phoneme loop. Although some refinements are used to further improve the scores, many works show that this is a good confidence measure [10, 11].

The overall phoneme score (PS) for a word can be readily defined as a weighted sum of the normalized GOPs of its composing phonemes:

$$PS(word) = \sum_{k=1}^{N} w_k \cdot GOP(phoneme_k) \quad (2)$$

where w_k is the weight of the k-th phoneme among the N phonemes composing the word. Typically, the weights are equal for all the phonemes [12].

An example of the behavior of the PS scores can be seen in Fig. 1. The analyzed utterance contains three words separated by silences. The PS scores have been calculated just at the output state of the final HMM of each word for each frame. The three PS score sequences are represented with different line types, over the spectral representation of the signal, in order to have a better idea of the limits of each word. As can be seen in the figure, each line reaches its maximum (of the –PS curve) when the corresponding uttered word finishes. That leads us to think that the three words would be correctly detected when the –PS score of each of them reaches a maximum over a certain threshold (around zero in the figure). A similar analysis could be done at phoneme level, where the PS score sequence would be calculated at the output state of each HMM for each frame.

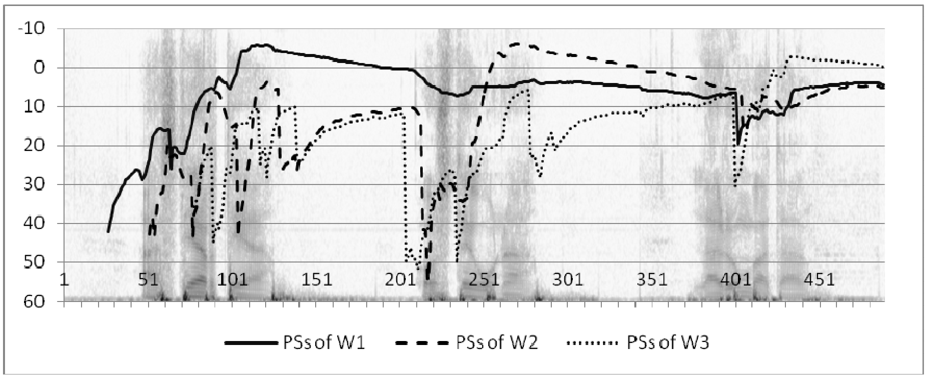

Fig. 1. PS score sequences of three words: W1 (“asteartea”, *Tuesday*), W2 (“osteguna”, *Thursday*), W3 (“larunbata”, *Saturday*), over the spectral representation of the signal.

2.3 The decision thresholds

The main problem that arises when working with both GOP and PS scores is how to calculate the decision thresholds. Two distributions are needed for this purpose: on the one hand, the distribution of the values of the PS scores when the expected or correct utterance is being verified and, on the other hand, the distribution of the values of the PS scores when an unexpected or incorrect utterance (for instance a different word or part of a word) is being verified. Then, a cut point between the two distributions can be selected, e.g. the equal error rate (EER), at which both the probability of false acceptance and false rejection are equal. This threshold may be moved taking into account the level of the L2 student; for a beginner, for example, a larger amount of errors can be accepted and, on the contrary, the system should be stricter with an experienced and high-skilled student.

The PS scores of the correct utterances can be calculated using the segmentation provided by the ASR in forced alignment mode. In order to calculate the incorrect utterances’ PSs, one valid technique is to artificially introduce errors in the dictionary. This is particularly useful if we consider that, in general, scarcity of data is a common problem in this kind of research [9, 13]. The literature shows that this technique has been successfully used to calculate phoneme GOP distributions. In this paper word level scores will be calculated and assessed, since it can be useful for tasks where not so strict results are needed, as working with beginners. The way to introduce errors in the dictionary consists in substituting a word by any other word in the sentence grammar, maintaining the segmentation obtained by a previous evaluation of the correct word sequence. Thus, the utterance that the system receives does not fit with the one that it is expecting, and so we can consider that for the system the audio file corresponds to an incorrect utterance.

The histograms of both the PS distribution obtained using this procedure for incorrect utterances and the PS distribution of the correct utterances are shown in Fig. 2. PS computation for incorrect utterances has been carried out three times, in order to obtain sufficient data, since the amount of incorrect words is one per file. The EER is located in the point 0.375, with a value of 2.12 %. This value is a priori encouraging, since the error made classifying a new incoming PS score can be considered very small.

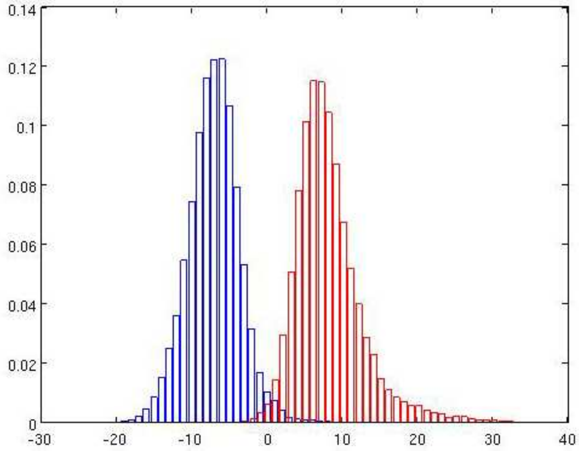

Fig. 2. The normalized histograms of the PS distributions of correct (left) and incorrect (right) utterances.

3 Experiments and results

The measure selected to assess the results is the widely used *SA* coefficient (Scoring Accuracy), which is calculated as in equation 3.

$$SA = \left(\frac{(CA + CR)}{(CA + CR + FA + FR} \right) * 100 \tag{3}$$

where: *CA*: Correctly Accepted; *CR*: Correctly Rejected; *FA*: Falsely Accepted; *FR*: Falsely Rejected.

Two different experiments have been carried out in order to check the consistence of the calculated decision threshold. The first experiment consists in observing whether the utterances that the system is expecting are labeled as correct (*CA*) or as incorrect (*FR*). For this purpose, the part of the database left for testing has been used: 2,218 files in total, with an amount of 7,296 words. The decoding was performed by the ASR in forced alignment mode, where two options were taken into account between words: an optional silence or a coarticulated transition. The PS scores were obtained over the final segmentation of the Viterbi decoder.

The results obtained are shown in the first result row of Table 1. We can see that SA is 97.18 % or, considering the error instead of the accuracy, 2.82 % error rate is achieved, which coincides approximately with the value of EER obtained for the decision threshold. It is worth mentioning that 75.24 % of the FR words contain 3 phonemes or less. So, it is evident that, as one could expect, the longer the word is, the more robust the result is. This may be due to the fact that if one of the phonemes is not correctly pronounced it affects more to the overall scoring in a short word than in a longer one. As an example, we noticed that in the results there are many short words containing the character *j*, which in standard Basque must be pronounced as /*jj*/ but in many dialects is pronounced as /*x*/[1].

Table 1. Results of the experiments 1 and 2

	CA	CR	FA	FR	SA
Experiment 1	7,090	---	---	206	97.18 %
Experiment 2	---	1.174	0	---	100.00 %

For the second experiment, only isolated words have been taken into account. In this case, artificially inserted errors consist in randomly substituting the input textual word with another word in the dictionary. The files containing isolated words in the test part of the database are 1,174, and the PS score provided over the ASR segmentation shows that none of them is classified as correct. So, a scoring accuracy of 100 % is obtained, as can be seen in the second result row of Table 1.

4 System design

In a realistic environment the student will make mistakes. That means that the verification system must be able to manage these extra voice segments in order to absorb the effects on the Viterbi decoder. Since the system will receive more voice frames than it expects, this must be modeled somehow. So, in the design of the final system an optional phoneme loop has been added to the decoding lattice of the ASR at the beginning, at the end and between words, in the way shown in Fig. 3. If this phoneme loop was not added, the segmentation resulting from the Viterbi algorithm would not be predictable, and the verification or scoring could not be calculated over this segmentation.

The system will be first assessing the PS scores for the first word W1, frame by frame (each 10 ms), until one of them goes over the threshold calculated previously as indicated in section 2.3. The PSs will be calculated over the segmentation that the

[1] http://aholab.ehu.es/sampa_basque.htm

incoming Viterbi token has in the last HMM state of the word. Then, the system will wait for a maximum (in the –PS curve), checking that the scores of the next *N* frames are smaller than this value. When a maximum is detected in that way, the word will be displayed for the user, and the same process will start for the next word W2.

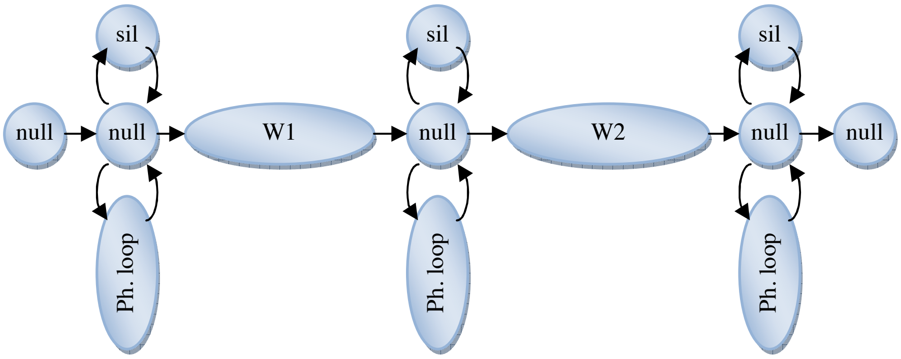

Fig. 3. The decoding lattice scheme for a sentence of two words, with optional silences and phone loops between words

In order to assess the designed system in a more realistic environment, a new experiment has been carried out, where expected and unexpected speech are combined. The test has been devised as follows: for each sentence or word sequence to be evaluated, one word has been deleted in the transcription. The aim of this experiment is to simulate a situation where a user utters some words correctly, in the same order as expected, then an incorrect word, and finally the remaining words correctly again. Notice that the word that is being evaluated during the incorrect segment is the next word that the system expects. So, we can obtain the score of that word when an incorrect word is uttered, and afterwards the score of the same word when the correctly uttered word arrives.

886 sentences have been used in this experiment, containing 5,080 uttered words in total. Each sentence ranges from two to seventeen words. In each written input sentence one word has been deleted, so that in the gaps created by the deletion the next word is verified. Thus, the word following to a deletion is verified twice: at first while the deleted word is being pronounced, and then when its corresponding utterance arrives. Thus, the PSs of 4,194 words —which are considered as correct— will be evaluated, and, in addition, 886 of them will be also evaluated for the unexpected speech segment, 5,080 scores in total.

The SA obtained in this experiment is, as can be seen in Table 2, 96.63 %. Regarding just the unexpected or incorrect words, a scoring accuracy of 84.88 % has been obtained (752 out of 886). Nevertheless, the total amount of incorrect words (886) is smaller than that of the correct ones (4,191) and that is why their influence is not so evident on the global rate.

Table 2. Results of the experiment simulating errors

Experiment results	
CA	4,157
CR	752
FA	134
FR	37
SA	**96,63 %**

5 Conclusions and future work

In this paper the design of a message verification system is introduced suitable to be implemented in CALL systems, in tasks that need word-by-word verification in real time. As soon as a word is detected, it is displayed to the user. In such a system, rejecting unexpected speech is as necessary as detecting correct words (words that the system is expecting). In this paper we describe a way to calculate the decision threshold, inserting artificial controlled errors. The results of the experiments show that the system has better scores when accepting expected words than when rejecting unexpected speech, although the decision threshold has been calculated using the EER. So, further analysis must be done, in order to detect why this asymmetry happens and adjust the decision threshold.

An experimental approximation to a realistic environment has been applied to better evaluate the system, but it lacks of real users' evaluations. Thus, some more tests must be devised to complete the analysis and to assess the system in a real environment. It would be interesting to test the system with students in different points of their language acquisition process, in order to adjust the thresholds to different situations. Another interesting experiment to do is to perform phoneme-level verification and compare the results of both systems.

The system presented in this paper has been designed for Basque, since Basque acoustic models have been used. However, the strategy can be easily followed to develop a message verification system for any other language, by means of creating the corresponding acoustic models for that language.

6 Acknowledgements

This work has been partially supported by the Basque Government (Berbatek, 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).

References

1. Schmid, E.C.: Interactive whiteboard technology in the language classroom: exploring new pedagogical opportunities, Saarbrücken, Germany: VDM Verlag Dr. Müller (2009)
2. Lamy, M.N., Hampel R.: Online communication in language learning and teaching, Houndmills: Palgrave Macmillan (2007)
3. Shield, L., Kukulska-Hulme, A.: Special edition of ReCALL (20, 3) on Mobile Assisted Language Learning (2008)
4. Norris, J.M., Ortega, L.: Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis, Language Learning, vol. 50, pp. 417-528 (2000).
5. Ellis, N.C., Bogart, P.S.H.: Speech and Language Technology in Education: the perspective from SLA research and practice, In Proceedings ISCA ITRW SLaTE, Farmington PA (2007)
6. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10): 832–844 (2009)
7. DISCO project website, http://lands.let.ru.nl/ strik/research/DISCO/.
8. Van Doremalen, J., Strik, H., Cucchiarini, C.: Utterance Verification in Language Learning Applications. Proceedings of the SLaTE-2009 workshop, Warwickshire, England (2009)
9. Odriozola, I., Navas, E., Hernáez, I., Sainz, I., Saratxaga, I., Sánchez, J., Erro, D.: Using an ASR database to design a pronunciation evaluation system in Basque. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4122-4126 (2012)
10. Witt, S., & Young, S. J.: Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30(2-3), 95-108 (2000)
11. Franco, H.: Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30(2-3), 121-130 (2000)
12. Mak B., Siu M., Ng M., Tam Y., Chany Y., Chan K., Leung K., Ho S., Chong F., Wong J., Lo J.: PLASER: Pronunciation Learning via Automatic Speech Recognition. In Proc. of HLT-NAACL, May, 2003, Edmonton, Canada (2003)
13. Kanters, S., Cucchiarini, C. and Strik, H.: The Goodness of Pronunciation algorithm: a detailed performance study, In Proceedings of SLaTE 2009, Birmingham (2009)
14. Hernaez, I., Luengo, I., Navas, E., Zubizarreta, M., Gaminde, I., Sanchez, J.: The Basque speech_dat (II) database: a description and first test recognition results, In Eurospeech-2003, 1549-1552 (2003)