A PRONUNCIATION TUTORING SYSTEM FOR BASQUE - FIRST DEVELOPMENT STEPS

Igor Odriozola[1], *Oliver Jokisch*[2], *Inma Hernáez*[1] *und Rüdiger Hoffmann*[2]

[1]*University of the Basque Country, Signal Processing Laboratory (Aholab)*
[2]*TU Dresden, Chair for System Theory and Speech Technology*

[1]{*igor, inma*}*@aholab.ehu.es*
[2]{*oliver.jokisch, ruediger.hoffmann*}*@tu-dresden.de*

Abstract: In this paper, we introduce the first steps of the integration of Basque language in the AzAR system, the pronunciation tutoring software developed for German and Slavonic languages. At the time of designing the curriculum for Basque, we noticed that developing prosodic (suprasegmental) aspects is as necessary as teaching pronunciation (segmental) aspects. Therefore, the design of the system includes both parts. On one hand, the initial steps of the development of a CAPT (computer-assisted pronunciation teaching) system are introduced. It relies on a standard automatic speech recognition (ASR) system based on hidden Markov models (HMMs), which uses GOP scores (Goodness of Pronunciation) as confidence scores. The process of calculation the decision thresholds is explained for languages like Basque that do not have a specifically designed database for verification. Some preliminary results and conclusions are explained. On the other hand, a new module for AzAR is introduced, which aims to automatically assess the prosody of students. The module consists of computing the RMSE between the f0 curves of the student's speech and the reference voice's speech, after aligning and subtracting the mean. First results show that the global distance scores thus obtained are smaller among Basque speakers than comparing speakers of different nationalities with the reference voice. We conclude that, although more experiments and results are needed, it can be useful to receive an automatic feedback from the system. Finally, some conclusions and reflections about future works are introduced.

1 Introduction

The speech data collection and the adaptation of automatic speech recognition (ASR) algorithms are essential tasks for the development of pronunciation tutoring (CAPT) systems. Using the expertise from the software project “Automat for Accent Reduction (AzAR)” [1], we created similar development and test data for the Basque language, and optimized a specific Basque ASR engine [2] for the pronunciation assessment on phonemic level. The article describes methods, procedures, interfaces and some results of the data collection and ASR adaptation as first development steps.

The AzAR system was originally designed for the pronunciation tutoring of learners with native language L1 from the Slavonic language group and for the target language L2 German. Within the cooperation project *Euronounce* [3], the concept was extended to Slavonic target languages Polish, Slovak, Czech and Russian. The *Euronounce* database includes special lessons for phonetic peculiarities but also sentences to evaluate the prosodic aspects, although the prosodic aspect has not an automatic feedback provided by the system. The database contains 130 speakers and about 200 hours of speech. In further steps, the *Euronounce* concept was also tested for Mandarin Chinese learners of German [4] which showed that specific challenges in target language L2 or L3 are dominating the variation from standard pronunciation rather than the influence of the mother tongue L1.

Basque is an isolated language which does not belong to the Indo-European language group, as one would expect from its geographical location [5]. It has two major languages as neighbors, Spanish and French, and the influence of these languages on Basque is noticeable, since Basque is in a situation of diglossia both in its north part (under the French administration) and in its south part (under the Spanish administration), and has an heterogeneous, quite complex, legal status and degree of official status, depending on the region.

As mentioned above, the neighboring languages have a strong influence on Basque language, especially for people who study it as L2. In addition, *Euskaltzaindia*, the Academy of the Basque language, has not yet made a decision on what the intonation and prosody for Standard Basque should be, due to the variety of accents and intonations that it shows from one dialect to another. That makes L2 students of Basque not have a clear reference of which pitch they must use, and this situation leads often to opt for that of their own L1. On the other hand, at the phonetic level, Basque has certain phonemes that do not exist in the inventory of the neighboring languages, and so they must be focused in the development of the evaluation system of pronunciation of phonemes.

Therefore, the first clear conclusion that has been obtained is that the design of a CAPT system for Basque must take into account two aspects: the segmental or relative to a single phoneme (in this case, the phonetic realization or pronunciation) and suprasegmental or relative to a group of phonemes (in this case, the accent and intonation).

2 The Basque curriculum for AzAR

As explained in the previous section, a part for the segmental evaluation and another part for the suprasegmental evaluation was designed to be implemented in AzAR.

2.1 The segmental part

In regard to the segmental part, the highlights of Basque, both phonetic and phonological, are:

- Phonetic features:

a) The vowel system and the system of diphthongs and hiatuses.

b) Phonemes not found in neighboring languages, for example, the /*ts`*/ (see Sampa Basque reference: [6]).

c) Differentiation between the 6 voiceless sibilants: /*s`*/, /*s*/ y /*S*/ (fricatives) and /*ts`*/, /*ts*/ y /*tS*/ (affricates).

- Phonological features:

a) The palatalization: The palatalization process of consonants /*l*/ and /*n*/ in the context /*i*CV/ (C is /*l*/ or /*n*/, and V is any vowel).

b) The voiceness loss process of the first phoneme of the verb following the negative particle *ez* (*ez dator* > /*es`tatorr*/, –he is not coming–) or disappearance of the sibilant of the negative particle (*ez nator* > /*enatorr*/, –I am not coming–).

This part of the curriculum consists of 60 word pairs (contrasts) and 125 phrases, which have been automatically extracted from the *Contemporary Reference Prose* [7] Basque textual corpus, collected by the Basque Language Institute of the University of the Basque Country (UPV/EHU). This corpus contains 25,1 million words, of which 13,1 million are drawn from books chosen for their quality (287 volumes) and 12 million are from newspaper articles published in both South and North Basque Country.

The reference voice was recorded from a native Basque speaker. The signals were recorded in digital format of 16 bits at 16 kHz, in the recording studio of the *Institut für Akustik und Sprachkommunikation* of the *Technische Universität Dresden*.

2.2 The suprasegmental part

The first steps for assessing the suprasegmental part, specifically the prosody, was implemented in AzAR under the project *Euronounce*, but the system was based only on the auditory perception of the students, since they should compare the intonation of their recorded phrases with the references. Therefore, there was no automatic evaluation, ie no feedback, from the system. For Basque, the evaluation of the suprasegmental part is essential and so a new prosody analysis module is being devised consisting of a plot of curves of fundamental frequency (f0) and an automatic score by comparing the f0 curve of the signal recorded by the student with that of the reference signal, calculating the RMSE as the distance metrics.

In this part, two features will be taken into account:

- The word-level stress:
- The phrase-level intonation.

The word-level stress chosen as reference is the one from the central dialect of Basque, since it is the most stable in a wider area. The stress of this area, although is not unique in the central dialect, follows a stress pattern [+2, -1] for isolated words over three syllables [8]. It is important to remark that Basque is an agglutinative postpositional language and that grammatical case marks are added to the noun as suffixes, after the article. The words formed in this way follow also the same pattern.

This part of the curriculum comprises a total of 20 isolated words and 50 phrases, selected from very different sources, in order to create a heterogeneous set of expressive phrases. They have been read by speakers of different mother tongues in a session where the speakers did not listen to any reference voice, not influencing their way of sentence intonation.

3 Development of the segmental part

3.1 The ASR and the acoustic database

To develop the segmental part, the phoneme verification system for Basque developed by the Aholab research group has been used used [2]. The verification is performed by a standard ASR based on hidden Markov models (HMM), by the forced alignment procedure. Thus, the system produces a GOP (Goodness of Pronunciation) score for each phoneme, which is used as a measure of confidence.

Usually CAPT systems rely on recordings of native vs. non-native speakers to evaluate the signals recorded by the students. Therefore, a database developed specifically for the verification of phonemes recordings must have both types of speakers. Using prior knowledge, one can foresee which some of the conflicting phonemes are for the student of a particular L1, and therefore, the databases are usually completed with the recordings of these phonemes [9]. Basque is an under-resourced language that has not enough acoustic databases for developing speech technology as its neighboring languages [10]. Currently, there are three acoustic databases designed to create speech recognition systems: the SpeechDat_eu database [11], recorded on the fixed telephone network at 8 kHz; the SpeechDat_eu_M, similar for mobile phones, and a Speecon-like database, recorded with various types of microphones at various distances at 16 kHz. The latter two were created with the funding of the Basque Government and have been released only for research.

The Speecon-like database contains recordings of native and non-native speakers and both dialectal and standard Basque speech. It contains audio signals of 230 speakers, recorded in different parts of the Basque Country. In each session, each speaker was asked about his/her language skills level, to choose between the options "native", "high level" or "low level". The native speakers' subcorpus consists of 149 speakers, the high level speakers' subcorpus of 56 speakers, and low level speakers' of 25 speakers.

3.2 The decision thresholds

The HMMs used to develop the pronunciation evaluation system have been trained using only native speakers. 76 speakers' sessions were selected, and the remaining ones were reserved for testing. These models were created for context-dependent phones (triphones) using vectors of 39 MFCC (Mel Frequency Cepstral Coefficients) parameters.

One of the most important tasks in the design of a pronunciation evaluation system is the calculation of the decision thresholds. To do this, the distribution pairs of the GOP scores of each phoneme were obtained in two different situations: when the phonemes are correctly pronounced and when they are incorrectly pronounced. The decision thresholds can be obtained by calculating the equal error rate (EER) of both distributions. To obtain the distributions of incorrectly pronounced phonemes, a simulation of incorrect uttered phonemes was performed, introducing controlled changes in the dictionary, i. e. replacing one phoneme in a particular position with another one of the same phonetic group (vowels, plosives, nasals, liquids and sibilants), at random. An example of one distribution pair can be seen in Figure 1.

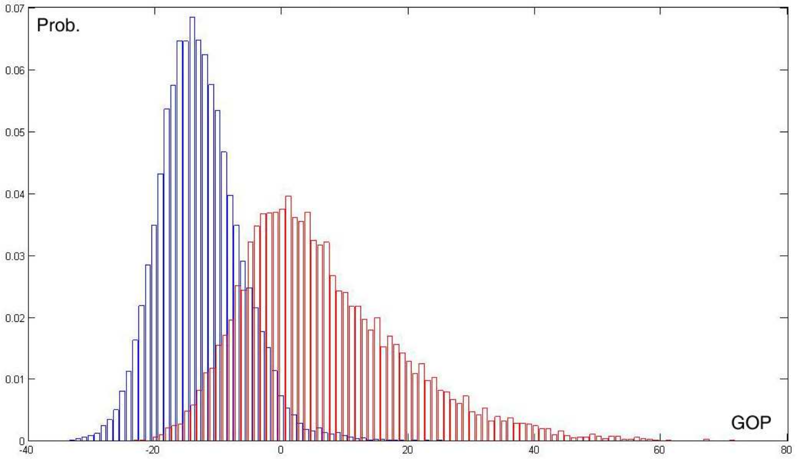

Figure 1 – Normalized GOP distr. for phoneme /*a*/; correctly (left) and incorrectly (right) uttered.

Figure 1 shows that a typical distribution pair overlaps in some extent. So, it is important to check which the EER value for each phoneme is, in order to see the ability of discerning phonemes of the system. The EERs thus obtained are shown in Table 1. While some phonemes have a low EER value, others have worst values that should be improved. In the case of voiced plosives, it may be due to the two different realizations that they have: as plosives (in the beginning of the words) or as approximants (between vowels). So, different acoustic models must be trained for these phonemes. In the case of the sibilants, which are important since they are included in the curriculum, two problems arise: on one hand, the /*s`*/ is pronounced as /*s*/ in some areas of the Basque Country, and, on the other hand, the /*ts*/ is pronounced as /*tS*/ in some other areas. So the corresponding HMMs have not been properly trained. Some results of the application of these ERR are shown in [2].

Phon.	EER	Phon.	EER	Phon.	EER	Phon.	EER	Phon.	EER
a	14,03	p	18,30	r	17,11	J	13,70	s`	34,82
e	18,64	b	30,09	rr	14,66	jj	32,05	s	16,94
i	7,99	t	20,54	l	17,66	f	12,85	S	20,47
o	15,53	d	24,88	L	19,14	T	24,62	ts`	10,61
u	10,92	k	22,34	m	15,66	x	5,10	ts	36,74
c	26,18	g	28,93	n	38,45	gj	-	tS	27,85

Table 1 – ERR values obtained for Basque phonemes.

3.3 Segmentation quality

The audio signals were introduced into the recognition system, and by means of the forced alignment procedure a phoneme level segmentation was obtained. To assess the quality of the automatic segmentation, the audio files were segmented manually as well. The evaluation was done frame by frame. A total of 584.612 frames (all files in the segmental part) were analyzed, and an overall accuracy of 88.08% was obtained.

At first glance, it was observed that the greatest differences occur at the word ends, where the recognition system needs a greater number of frames to leave the last HMM state when this does not correspond to the silence model. This may be because the echoes and reverberations which, although minimal in the studio recording, exist and have some presence in the signal that the human being classifies as silence.

To solve this problem, more robust silence models should be created. However, we can consider that the results of the segmentation are of good quality.

4 Development of the suprasegmental part

To calculate the distance between two intonation patterns several metrics can be found in the literature, like the mean-absolute-frequency-deviation [12], the pitch target points [13] and the use of temporary shifts in f0 patterns [14], but none of them seems to perform better than the classic RMSE (root-mean-square error) or the correlation coefficient of Person R^2. In this paper, RMSE will be used to measure the distance between the pitch (f0, fundamental frequency) curves of different early learners of Basque and the pitch curve of the previously recorded reference voice.

4.1 The distance between f0 curves

The speech signals recorded from the speakers were first recognized by the forced alignment procedure, in order to obtain a segmentation of the utterances. The recognition process was performed using acoustic models trained on a Basque speech database, so manual correction was needed to cope with differently pronounced phonemes and the effects they cause in the segmentation. Then, the f0 curves of the voiced segments were computed and aligned with the corresponding voiced segments of the reference voice using dynamic time warping (DTW) technique (see Figure 2). The distance between each curve pair was calculated using RMSE, after having subtracted the f0 mean of all the voiced segments of each audio file. With all the individual distances obtained in each segment, a mean distance or global distance was obtained.

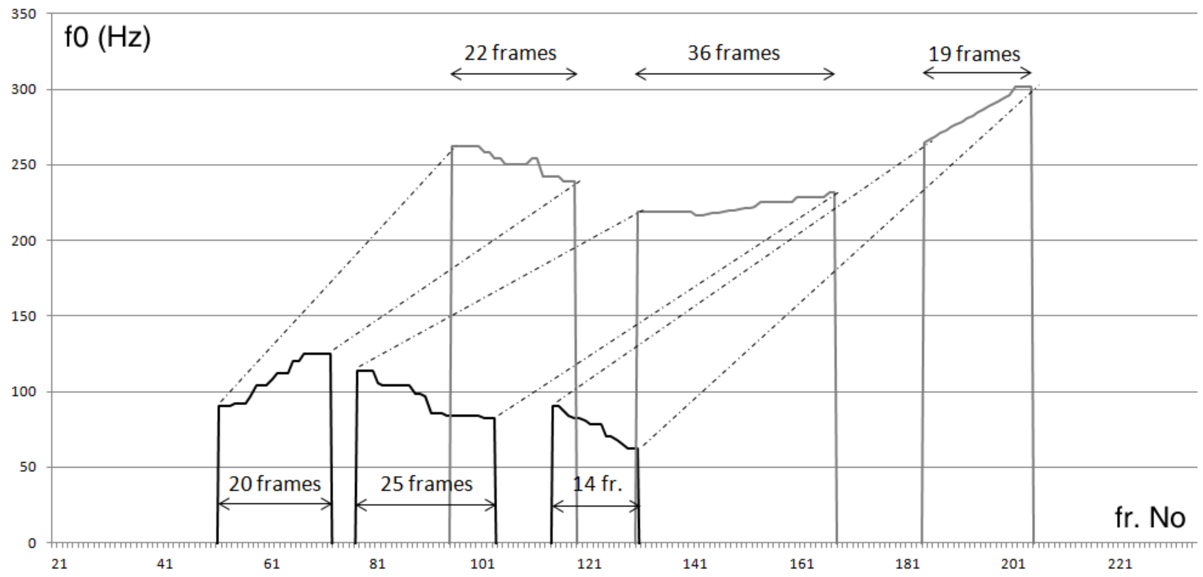

Figure 2 – The f0 curves of the voiced segments of the reference voice (in black) and the Slovakian voice (in gray) uttering the word *independentzia* (/*independents`ia*/).

4.2 First results

The experiment was carried out with eight speakers of different nationalities and different mother tongue. Two speakers were Basque, one male and one female, in order to have a reference of the behavior of the global distance. The remaining six speakers were from different countries, five males and one female. The speakers were asked to read the same set of sentences without listening to any reference voice previously, in order to create no influence on them.

Since the pitch perception of the human ear is proportional to the logarithm of frequency rather than to frequency itself, two experiments have been performed to cope with the effects of working with voices that have different fundamental frequency means (notice that typically the pitch of female voices is at around 200 Hz and the pitch of male voices at around 100 Hz). In the first experiment, f0 curves are used; in the second, log(f0) curves. The results are shown in Table 2.

Speaker number	**Mother tongue**	**Gender**	**Age**	**RMSE f0 curve [Hz]**	**RMSE log(f0) curve**
01	Japanese	m	27	15,4	0,134
02	Macedonian	m	38	17,9	0,140
03	Amharic (Ethiopia)	m	32	16,4	0,134
04	German	m	42	20,1	0,165
05	Urdu (India)	m	31	15,5	0,129
06	Slovakian	f	26	20,3	0,135
07	Basque	f	35	19,3	0,111
08	Basque	m	34	14,0	0,113

Table 2 – RMSE of f0 and log(f0) curves of the speech of 8 speakers compared to the reference voice.

The results show that the global distance differs if logarithms are used or not. Without the use of logarithms, the distance between the reference voice (male) and the female voices is noticeably bigger. Using logarithms however, the shortest distance obtained is precisely that corresponding to the Basque female voice. The Basque male voice achieves the shortest distance without using logarithms, and the second shortest distance using logarithms. Furthermore, the differences among the global distances are considerable: the difference between the Basque speakers is 1,8 % relative to the voice that has obtained the shortest distance, while a difference 16,2 % is obtained for the voice with the shortest distance of the group of non-Basque speakers. These results are encouraging, since they show the behavior one would expect beforehand.

5 Conclusions and future work

In this paper, the development of a curriculum for Basque language to be integrated in the AzAR system has been introduced. It consists of two linguistic aspects: on one hand, the pronunciation evaluation module (segmental part) and the prosody evaluation module (suprasegmental part). The later extends the capabilities of the current implementation of AzAR, since currently the system does not provide an automatic scoring of the evaluation of the prosody of the speaker.

The development of the segmental part relies on the phoneme verification system of the Aholab research group, which is designed using acoustic models (HMMs) trained over a general ASR database instead of being trained over a specifically designed database. This fact creates the difficulty of having not so clean models, and that is what the calculated decision thresholds show. As a conclusion, some allophones of the same phoneme (for example, the case of the voiced plosives that are uttered as approximants between vowels) must have their own HMM instead of using only one. Furthermore, sibilants probe to need more carefully selected signals, since their use differs in some areas of the Basque Country and so the acoustic models are not trained properly. Nevertheless, the consequences of using such models can be reduced in some extent using different levels in the feedback provided by the system. Previous results show that the scoring accuracy, using only two correctness levels (correct vs. incorrect), is over 80 % for properly trained phonemes and around 70 % for problematic phonemes. Using a system of three or even five levels (as in AzAR system) in order to cope with the overlapping part of both distributions, better results will be probably obtained.

Regarding the suprasegmental part, encouraging results have been obtained. The automatically computed RMSE distances between log(f0) curves are smaller for speakers of the same language regardless of his/her gender. These scores can be really helpful for the student, especially if the f0 curves are also plotted so that the student can compare visually the curve that he/she has produced with the reference one, in addition to the current auditory reference.

The system needs to be tested in real conditions. For this purpose, the integration of the Basque curriculum in the AzAR must be completed and then proceed to adjust and improve the system depending on the results obtained by real students. The experiments of the suprasegmental part throw good results evaluating speakers with different mother tongues, but it would be interesting to see how they behave for Spanish or French speakers. Besides, they have been carried out offline, so a way to integrate the scripts for a real time behavior must be devised.

References

[1] Jokisch, O.; Koloska, U.; Hirschfeld, D.; Hoffmann, R.: Pronunciation learning and foreign accent reduction by an audiovisual feedback system. In Proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII), Beijing, pp. 419 – 425, October 2005. Springer LNCS-3784, Berlin.

[2] 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 Proc. 8th Intern. Conf. on Language Resources and Evaluation (LREC), Istanbul, pp. 4122 – 4126, May 2012.

[3] Jokisch, O.; Jäckel, R.; Rusko, M.; Demenko, G.; Cylwik, N.; Ronzhin, A.; Hirschfeld, D.; Koloska, U.; Hanisch, L.; Hoffmann, R.: The EURONOUNCE project - an intelligent language tutoring system with multimodal feedback functions: Roadmap and specifications. In A. Lacroix, editor, Konferenz Elektronische Sprachsignalverarbeitung (ESSV) 2008 (Studientexte zur Sprachkommunikation Bd. 50), pp. 116-123, Frankfurt/M., Germany.

[4] Ding, H.; Mixdorff, H.; Jokisch, O.: Pronunciation of German syllable codas of Mandarin Chinese speakers. In Proc. Konferenz Elektronische Sprachsignalverarbeitung, ESSV (Studientexte zur Sprachkommunikation Bd. 58), pp. 281 – 287, September 2010, Berlin.

[5] Hualde, J. I.: Basque Phonology, Taylor & Francis, 1991.

[6] Aholab Signal Processing Laboratory, University of the Basque Country (UPV/EHU): SAMPA computer readable phonetic alphabet of Basque, retrieved on 1st July 2012 from http://aholab.ehu.es/sampa_basque.htm

[7] Basque Language Institute, University of the Basque Country (UPV/EHU): *Contemporary Reference Prose* corpus, retrieved on 1st July 2012 from http://www.ei.ehu.es/p289-content/en/contenidos/informacion/euskara_inst_erdaretan/en_erdaret/epg.html

[8] Hualde, J. I.: Basque accentuation. In H. van der Hulst (arg.), Word prosodic systems in the languages of Europe, Mouton de Gruyter, Berlin, pp 947 – 993, 1999.

[9] Demenko, G.; Wagner, A.; Cylwik., N: The Use of Speech Technology in Foreign Language Pronunciation Training. Archives of Acoustics, 35(3), pp 309 – 329, 2010.

[10] The META-NET Language White Paper Series: Overview and Key Results, retrieved on 1st July 2012 from http://www.meta-net.eu/events/meta-forum-2012/talks/bolette-pedersen.pdf

[11] Hernáez, 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 Proc. of Eurospeech-2003, pp 1549 – 1552. Geneva (Switzerland), 2003.

[12] Bellegarda, J. R.; Silverman, K.; Anderson, V.: Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation, IEEE Transaction on Speech and Audio Processing, vol. 9, no. 1, pp. 52–66, January 2001.

[13] Campione, E.; Véronis J.: A statistical study of pitch target points in five languages, in Proceedings of ICSLP 98, 1998.

[14] Clark, R. A. J.; Dusterhoff, K. E.: Objetive methods for evaluating synthetic intonation, in Proceedings of Eurospeech 99, September 1999.