Assessment of Non-native Prosody for Spanish as L2 using quantitative scores and perceptual evaluation

V. Cardeñoso-Payo, C. González-Ferreras, D. Escudero-Mancebo

Department of Computer Science. Universidad de Valladolid.
Paseo de Belén 15. 47011 Valladolid. SPAIN.
(valen, cesargf, descuder)@infor.uva.es

Abstract

In this work we present SAMPLE, a new pronunciation database of Spanish as L2, and first results on the automatic assessment of Non-native prosody. Listen and repeat and read tasks are carried out by native and foreign speakers of Spanish. The corpus has been designed to support comparative studies and evaluation of automatic pronunciation error assessment both at phonetic and prosodic level. Four expert evaluators have annotated utterances with perceptual scores related to prosodic aspects of speech, intelligibility, phonetic quality and global proficiency level in Spanish. From each utterance, we computed several prosodic features and ASR scores. A correlation study over subjective and quantitative measures is carried out. An estimation of the prediction of perceptual scores from speech features is shown.

Keywords: computer assisted pronunciation training, perceptual evaluation, non-native corpora

1. Introduction

The demand for foreign language learning has dramatically increased as a consequence of globalization. Constant growth of computing capabilities of smart-phones and other kinds of personal computing assets contributes to a consolidation of computer-assisted pronunciation teaching (CAPT) as a basic tool to help boosting second language acquisition. Pronunciation learning is one of the key aspects of foreign language learning, specially when face to face communication skills are in focus of competence. Teaching correct pronunciation of a foreign language has traditionally been assumed to require the highest level of teacher student interaction. As pointed out in (Witt, 2012), it involves different aspects related to speech recognition, linguistics, psycholinguistics and pedagogy. All of those research fields have to be brought together in the conception, design and evaluation of automatic pronunciation teaching solutions. Research works on automated pronunciation error detection carried out since its offset more than twenty years ago have recognized the problem in its entirety as a difficult one and, thus, have addressed separately the two main components: phoneme level or prosodic level (Eskenazi, 2009). Many different features have been used to measure errors at these two levels and most of the research proposals up to date require a manually annotated database of non-native pronunciations which is very costly and scales poorly.

It is well known that prosodic level pronunciation errors limits proficiency and mutual understanding for non-native speakers (Tepperman and Narayanan, 2008). A large number of metrics have been used along the years in order to measure this pronunciation dimension (Witt, 2012). Nevertheless, subjective evaluation of perceptual aspects is still a must when trying to compare automatic solutions to the prescriptions of experts.

This work has been funded by *Ministerio de Ciencia e Innovación* of the Spanish Government through research project GLISSANDO [FFI2008-04982-C003-02]) and by *Consejería de Educación of the Junta de Castilla y León* through research project SAMPLE [VA322A11-2].

In this work we present a new pronunciation database of Spanish by native and non-native speakers which has been designed to support comparative studies and evaluation of automatic pronunciation error assessment both at phonemic and prosodic level. A correlation study between off-the-shelf likelihood scores provided by freely accessible ASR commercial systems and subjective perceptual scores of linguistic competence provided by expert linguists using several dimensions is also included.

2. Corpus description

In the framework of the SAMPLE research project, we faced the development of a corpus of spoken Spanish by foreign speakers as a means to support future CAPT studies. The central part of the corpus is made of a set of sentences and paragraphs selected from the news database of a popular spanish radio news broadcasting station. The texts cover various information domains related to every day's life and is itself a subset of the GLISSANDO corpus (Garrido et al., 2013), developed in the framework of a project related to our research line in automatic prosodic labelling. We chose the textual material from the subset of prosodically balanced sentences in GLISSANDO corpus, which statistically resembles Spanish language prosodic variety (Escudero et al., 2009).

We recorded 14 Spanish L2 speakers: 9 American English and 5 Japanese. All of them were students of Spanish at a university level. We also recorded 8 native Spanish speakers of different speaking styles, to have a set of reference pronunciations. The set of foreign speakers was selected with the guidance of educational personnel of the Languages Center of our University, among students ranging from A2 to B2 Spanish proficiency levels.

All the recordings were carried out under studio conditions, using a digital recorder at a sampling rate of 48kHz and a professional studio microphone. The recordings for each speaker were conducted in a single session inside a noise free room and following a protocol which included read-

sid	**Sentence to read or listen-and-repeat**
s01	La coalición interpuso esta querella por prevaricación el viernes pasado.
s02	52 denuncias por faltas graves en dos años, 18 de ellas graves por carecer de licencia de funcionamiento. Y el bar sigue abierto.
s03	Para una gala que se celebrará el 8 de febrero del próximo año.
s04	ATT prevé eliminar 12.000 empleos y reducir inversiones de capital.
s05	Notó una foto con flash cuando volvía a su domicilio.
s06	En la cartelera de cine no hay este fin de semana mucha poesía que digamos.
s07	¿Qué sería de una Navidad sin su cesta?
s08	Más de un millón de mujeres trabajan actualmente por cuenta propia.
s09	Y en los mercados los números rojos se extienden hoy por todas las bolsas europeas.
s10	No les han ofrecido hotel, ni tan siquiera a un vaso de agua.
s11	Sin embargo, también hay una buena noticia. existen soluciones.
s12	Sigue con sus trabajos de investigación, en los que ya constan sus conversaciones con la presidenta regional.
s13	Todos ellos, según las últimas informaciones del diario El País, fueron también víctimas de seguimientos.
s14	Esta investigación interna no ha dado aún ningún dato concluyente, y no tiene fecha límite.
s15	Hoy, hay huelga en las escuelas infantiles.

Table 1: Set of sentences used in SAMPLE corpus for the read and listen-and-repeat task.

ing several sets of sentences and paragraphs which are described below. On the average, recording sessions lasted for around 40 minutes and speakers were given the freedom to rest whenever they wanted between consecutive recording runs. Although the contents of SAMPLE corpus include only read speech, for the short sentences task the speakers were asked to read silently each sentence first and then trying to say it as naturally as possible. For every foreign speaker, each recording session included the following steps:

- *First sight read sentences*. Fifteen short sentences were selected from the news paragraphs of the prosodic GLISSANDO corpus, following a phonetic coverage criterion (see table 1). From them, 10 (s01-s10) were selected to be read at first sight by non-native speakers. Ten sentences were read with small pauses between them and the task was repeated three times with resting stops in between. This provides a basis for the experimental study of the influence of simple reading repetition on the pronunciation correctness.

- *Listen and repeat sentences*. A group of 10 (s05-s15) additional sentences was gathered reusing the last 5 of the previous ten sentences and 5 fresh ones from the original set of fifteen sentences. Using a simple tablet application, a reference utterance of each sentence by a native professional speaker was presented to the non-native speaker, who had to carefully listen and repeat it immediately afterwards. Again, this process was repeated three times to provide a means of evaluation of the effectiveness of this guided pronunciation scheme.

- *Short story*. A text with the Spanish translation of the well know Aesop's Fable 'The North Wind' was given to each non-native speaker, who had enough time as to fully understand the meaning and sense of this story. Then, they were required to tell the story as if it were told to a child, trying to provide the best intonation they could. This passage is recommended by the IPA for the purpose of eliciting phonemic contrasts in different languages (Visceglia et al., 2009).

- *News paragraphs*. Finally, fifteen news items of the prosodic GLISSANDO corpus were selected, each one with an associated reading time of around eighty seconds. From the lexical and semantic point of view, they cover different information domains of every day's life and show different levels of pronunciation difficulty, including dates, numbers and common names for places, people and organizations.

Parameter	**Quantity**
Number of speakers	22 = 14 foreign + 8 native
Number of sentences	15 (avg duration: 5.8s)
Number of paragraphs	16 = 15 news + 1 fable
Number of utterances	1179 = 960 sent + 219 par
Recording time sentences	5586s
Recording time paragraph	17646s
Total recording time	6h27m12s

Table 2: SAMPLE corpus statistics.

The summary statistics for the corpus are shown in table 2. For each speaker, the corpus includes fifteen short sentences (5.8s average duration) and up to sixteen longer paragraphs (80.6s average duration).
This design of the corpus contents provides means to support several kinds of studies: the influence of the controlled and progressive repetition of text fragments on the pronunciation quality, both depending on the text and text independent; the differences in pronunciation quality between spontaneous reading and comprehensive storytelling; specific errors associated to L1 of the non-native speaker, etc.
The database was also designed to support the development of intonation and pronunciation quality assessment for nonnative Spanish. It provides feedback for prototypic learning applications, as the one built in the framework of the SAMPLE project (Escudero-Mancebo and Cardeñoso-

Payo, 2013; Vallejo-Alonso, 2013), which was designed to ease the collection of non-native pronunciations of Spanish and to incorporate automatic pronunciation quality assessment in a near future.

3. Human assessment

Four experts have independently assigned **perceptual evaluation measures** along five different dimensions, using a Likert scale, and a proposed overall proficiency level according to the Common European Framework of Reference for Languages, Teaching and Assessment (CEFR) as applied to Spanish (DELE[1]).

All the labelers already had good competences in the evaluation of Spanish as a second language, developed as part of their training background in the university degree in Spanish Language and Literature. After a selection process, we provided specific training sessions on the evaluation protocol and the expected meaning and scales of the target parameters we proposed to label the utterances in the corpus. Open discussions favored the establishment of a common ground for the criteria to follow for the evaluation along the different dimensions.

The labeling process was monitored in order to detect possible anomalous deviations in the assessment criteria for some of the evaluators. Along the labeling process, we conducted several follow up sessions to try to keep general criteria as homogeneous as possible.

Most of the previous works have used a single dimension to assess pronunciation quality by human experts (Teixeira et al., 2000; Yamashita et al., 2005; Tepperman and Narayanan, 2008; Cincarek et al., 2009; Cheng, 2011). In this work, we follow an approach based on several dimensions, similar to the one recently proposed in (Hönig et al., 2010), because this allows us to evaluate different aspects of the utterances instead of a single overall performance. Perceptual dimensions include:

- *intelligibility* (**int**): the expert provides an integer value to indicate the level of understanding of what has been said (1:very poor, 5:excellent).
- *fluency* (**flu**): the expert provides an integer value to indicate the level of interruptions, hesitations, filled pauses and other phenomena which could affect fluency (1:very poor, 5:excellent).
- *phonetic correctness* (**pho**): the expert provides an integer value in order to evaluate if all the phonemes have been correctly pronounced (1:clearly non-native, 5:native).
- *lexical accent correctness* (**acc**): the expert provides an integer value in order to evaluate if lexical accent (position of the accented syllable within the word) is correctly positioned according to any accepted pronunciation of Spanish (1:clearly non-native, 5:native).
- *rhythm* (**rhy**): the expert provides an integer value in order to evaluate to which extent the prosody clearly resembles the one in a native Spanish speaker or, on the contrary, shows a neat non-native accent (1:clearly non-native, 5:native).
- *Spanish level* (**dele**): the expert indicates which level of proficiency of Spanish appears to have the speaker, according to the DELE scheme (A1, A2, B1, B2, C1 or C2) and using a 1 (A1) to 6 (C2) numeric scale.

The labelers filled their evaluation scores for the perception experiment using a web-based application. A total of 1179 utterances were randomly presented to the evaluator in sequence through a web form. They could listen to the utterance as many times as they wanted and the form was filled with the perceptual scores, the estimated DELE reference level and any additional comments they would like to add for that particular utterance or speaker. The average evaluation time was around 9 times longer than the average utterance duration, which illustrates the high cost of manual annotation. Since the samples were presented at random, the likelihood that the labeler could listen to two of them in the same order they were recorded is negligible, as can be easily computed.

4. Features

A set of prosodic and speech recognizer features were automatically extracted from corpus sentences, as described in this section.

4.1. Speech Recognition Features

ASR scores: In a first step, we have used Google publicly available speech recognition technology[2] to get recognition results for each sentence. This provides simple and affordable global quantitative scores of the **pronunciation quality at phonetic level**. Five recognition hypotheses are requested to the system. A score (**gscore**) related to likelihood to the best candidate hypothesis can be easily obtained from the speech API REST service. The sentence of the best candidate hypothesis provided by Google Speech API v1 for the recognized sentence is then aligned to the reference sentence to compute the Levenshtein distance (**ldist**), normalized with respect to the length of the utterance the speaker had to read. This distance provides a quantitative measure of the matching between what the recognizer understood to be the best sentence candidate and the original sentence which the speaker was assumed to read. Since the API is not designed to facilitate easy tuning of the recognition parameters, **gscore** (and consequently **ldist**) values indicate no recognition at all, when the amount of disfluencies found in the utterance is high.

Forced-alignment scores: In order to get a more precise and controlled parameterization of phonetic and prosodic units within a utterance, a phonetic segmentation of all the utterances has been carried out using the HTK toolkit[3]. The utterances were segmented by forced alignment using continuous density hidden Markov models. A standard 39-dimensional feature vector was used for feature representation (12 MFCCs and normalized energy, along with the first and second order derivatives). Feature vectors were

[1] http://www.dele.org/

[2] http://research.google.com/pubs/SpeechProcessing.html

[3] http://htk.eng.cam.ac.uk/

Evaluators	int	flu	pho	acc	rhy	dele
A,B,C,D	0.30	0.53	0.17	0.17	0.13	0.34
A,C,D	0.41	0.57	0.31	0.34	0.28	0.56
A,C	0.47	0.61	0.23	0.43	0.44	0.58
A,D	0.33	0.60	0.50	0.28	0.17	0.51
C,D	0.40	0.47	0.16	0.27	0.22	0.58

Table 3: Krippendorff's α for several combinations of evaluators (ordinal metric).

extracted using a 16ms Hamming window and 10ms frame rate. Phone models were 4-state left-to-right mono-phone HMMs with six Gaussian mixtures. Forced-alignment phonetic tier is then used to support the computation of the specialized prosodic features and to obtain two quality measures from the HMM automatic speech recognizer: the accumulated log probability per utterance (**AP**) and the average log probability per frame (**PPF**).

4.2. Prosodic Features

Although we have developed an algorithm for multiclass automatic prosodic labeling based on SpTOBI (see (González-Ferreras et al., 2012), in this work we have only computed specialized prosodic feature sets (according to the nomenclature in (Hönig et al., 2010)). These correspond to well known features sets for scoring methods already presented in (Kim et al., 1997) and (Neumeyer et al., 2000).

Speech rate measures: For each utterance, we compute a rate of speech (**ros**) as the number of phones per second.

Global interval proportions: We computed the proportion of vocalic intervals (**v**) (sum of the lengths of vocalic intervals divided by the total duration of the sentence, excluding pauses), as proposed by (Ramus et al., 1999). The standard deviation of the duration of vocalic intervals (**deltav**) and of consonantal intervals (**deltac**) are computed at utterance level. At speaker level, the average and standard deviation of this three features bring six Global Proportions of Intervals (GPI) features per speaker. Following (Dellwo and Wagner, 2003), we also computed the standard deviation of consonantal (**varcov**) and vocalic (**varcoc**) interval durations divided by mean consonantal or vocalic duration within the utterance.

Variability indexes: We identify vocalic and consonantal segments and computed two forms of the Pairwise Variability Index proposed in (Grabe and Low, 2002):

$$rPVI = 100 \times \frac{\sum_{i=1}^{N-1} |d_i - d_{i+1}|}{N-1} \quad (1)$$

$$nPVI = 100 \times \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|d_i - d_{i+1}|}{(d_i + d_{i+1})/2} \quad (2)$$

With these, four utterance-level features are extracted: **rPVIv**, **nPVIv**, **rPVIc**, **nPVIc**, which could later be computed as 8 speaker-level PVI features, after computing average and standard deviation across each speaker utterances.

	int	flu	pho	acc	rhy	dele
int	1.00					
flu	0.57	1.00				
pho	0.66	0.47	1.00			
acc	0.70	0.57	0.67	1.00		
rhy	0.58	0.58	0.59	0.72	1.00	
dele	0.82	0.72	0.80	0.82	0.78	1.00
gscore	0.33	0.25	0.33	0.41	0.39	0.41
ldist	-0.36	-0.15	**-0.40**	-0.40	-0.46	-0.45
ros	0.33	**0.70**	0.22	0.34	0.35	0.43
v	0.25	0.32	0.08	0.12	0.13	0.25
deltav	-0.02	-0.21	-0.14	-0.14	-0.13	-0.08
deltac	-0.33	**-0.60**	-0.24	-0.31	-0.34	-0.45
varcov	-0.05	-0.24	-0.19	-0.17	-0.17	-0.14
varcoc	-0.29	-0.52	-0.26	-0.32	-0.34	-0.43
rPVIv	-0.06	-0.17	-0.15	-0.19	-0.12	-0.12
rPVIc	-0.34	**-0.60**	-0.23	-0.30	-0.35	-0.45
nPVIv	-0.03	-0.05	-0.10	-0.16	-0.12	-0.08
nPVIc	-0.27	-0.46	-0.17	-0.28	-0.31	-0.37
PPF	**-0.45**	-0.51	-0.37	**-0.43**	**-0.63**	**-0.61**
AP	0.25	0.52	0.26	0.29	0.19	0.30

Table 4: Pearson correlation (r) among perceptual scores (upper part) and features and perceptual scores (lower part), always at utterance-level.

	int	flu	pho	acc	rhy	dele
int	1.00					
flu	0.69	1.00				
pho	0.92	0.45	1.00			
acc	0.92	0.68	0.92	1.00		
rhy	0.84	0.62	0.81	0.93	1.00	
dele	0.97	0.66	0.93	0.97	0.92	1.00

Table 5: Pearson correlation (r) among perceptual scores at speaker-level (average accross labellers, sentences and repetitions).

5. Experiments and results

Given that the evaluation task is highly subjective, we first conducted an inter-rater consistency check. With this, we try to detect which evaluators, if any, provided scores consistent enough with the rest. To this aim, we have calculated Krippendorff's α IRR indicator(Krippendorff, 2004) using $ordinal$ metric for all the combinations of sets of labelers, from 2 to 4 members.

The best consistency results are selected in table 3. In general, α values are below the minimum required threshold to ensure consistency according to this indicator, except for **dele** scores when evaluator B is not considered. If we select evaluators (A,C,D) we reach a compromise between level of consistency and number of evaluators to get average scores for the rest of the analysis.

We evaluated pairwise Pearson correlation between evaluated perceptual dimensions, both at utterance level (table 4) and speaker level (table 5). At utterance level, we selected just the four speakers for whom all utterances were evaluated by all labelers. At speaker level, the r values for different scoring criteria are highly correlated among themselves, and give similar results to the ones reported in (Hönig et al., 2011), where a higher number of evaluators was recruited.

In the bottom part of table 4, we show correlation among

	gscore	ldist	ros	v	deltav	deltac	varcov	varcoc	rPVIv	rPVIc	nPVIv	nPVIc	PPF	AP
gscore	1.00													
ldist	**-0.70**	1.00												
ros	0.00	0.15	1.00											
v	0.16	-0.11	0.15	1.00										
deltav	0.07	-0.04	-0.38	0.60	1.00									
deltac	-0.14	0.10	-0.60	-0.61	-0.11	1.00								
varcov	0.04	0.05	-0.37	0.43	**0.94**	-0.02	1.00							
varcoc	-0.18	0.18	-0.40	-0.47	-0.06	0.90	0.03	1.00						
rPVIv	0.09	-0.12	-0.42	**0.69**	0.80	-0.17	**0.67**	-0.16	1.00					
rPVIc	-0.14	0.08	**-0.63**	-0.66	-0.15	**0.95**	-0.07	**0.77**	-0.20	1.00				
nPVIv	0.10	-0.14	-0.31	0.61	0.54	-0.23	0.41	-0.24	**0.88**	-0.23	1.00			
nPVIc	-0.17	0.12	-0.44	-0.53	-0.17	0.63	-0.11	0.54	-0.20	**0.79**	-0.17	1.00		
PPF	-0.40	**0.46**	-0.25	-0.37	-0.13	0.41	-0.04	0.39	-0.19	0.46	-0.21	0.41	1.00	
AP	0.12	0.07	0.53	0.19	-0.21	-0.35	-0.25	-0.30	-0.20	-0.29	-0.12	-0.17	0.05	1.00

Table 6: Pearson correlation (r) among computed features at utterance-level.

labeler perceptual scores and the computed speech features described in section 4.2. The r coefficient for features-features correlation is shown in table 6.

It can be seen that **dele** proficiency level is highly correlated to the rest of perceptual dimensions. The poorest correlation corresponds with the pair (**dele**, **flu**), no matter if we work at utterance or speaker level. This could clearly indicate that human evaluators did not always found the same relationship between high quality Spanish production and high speech fluency: sometime, speech production could be as fluent as in Spanish but intelligibility, accent or rhythm could be bad. For the rest of dimensions, $r > 0.80$, which is a good result. Since four of the five perceptual dimensions seem easier to evaluate while **dele** is quite complex to grasp, high correlation could provide an indirect means to assess DELE level.

An interesting result for the self-consistency of this study is that the scores provided by the speech recognizers and the Levenshtein distance between the best hypothesis and the prompt sentence are highly correlated (corr=0.70 with **gscore** and corr=0.46 with **PPF**). This lack of matching manifests differently for **gscore** and for **PPF** simply because Levenshtein distance is obtained from the sentence hypothesis associated to the first feature. This could be expected in general, and the correlation is not higher because in many cases the recognized sentence does not correspond to the original one. It happens than non-native speakers produced sometimes a different phonetic variation or a very different timing in the phonetic elements of the sentence, associated to hesitations and pronunciation errors. We hope these cases to give clues about the parts of the sentence which should guide new pronunciation exercises under a feedback directed learning environment.

As for the relationship between quantitative scores provided by the recognizer or the sentence distance and the set of perceptual dimensions, Pearson correlation remain usually low except for some exceptional cases, although the p-values are always clearly below p=0.0005. Although a deeper study on feature merging is mandatory before getting any definitive conclusion, some comments on table 4 can be made. Overall, r values among speech features and subjective criteria are lower than the ones among subjective criteria themselves. Nevertheless, the values are acceptable and point out some kind of relationship among speech features and perceptual criteria which should be further investigated.

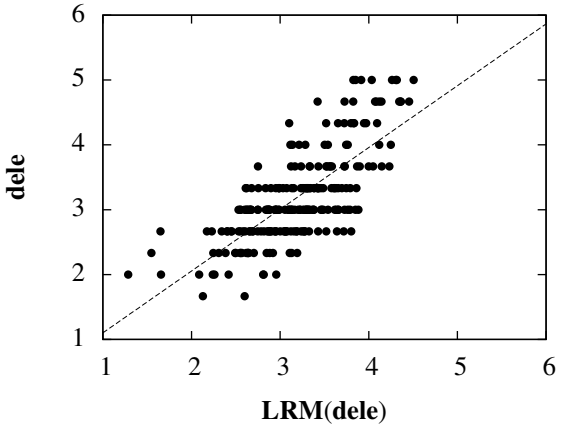

Figure 1: **dele** versus its Multiple Linear Regression prediction LRM(**dele**).

Finally, we evaluated the prediction power of labelers scores from automatic computed speech features. For this, we obtained a 10-fold linear regression model (LRM) using M5 feature selection algorithm, provided by Weka. We experimented with other models too (regression trees, MD5 on quantized features, or SVM) but in all cases results were similar for the moment. Since LRM results are cleaner to understand and represent, we circumscribe to them in this paper. The LRM best fit is given in the following equation:

$$\textbf{LRM}(\textbf{dele}) = -0.7375 * \textbf{ldist} + 0.0469 * \textbf{ros} + 0.009 * \textbf{v} + 2.0701 * \textbf{deltav} - 0.0022 * \textbf{varcoc} - 0.0726 * \textbf{rPVIv} - 0.1081 * \textbf{PPF} - 5.4796$$

In figure 1 we plot the value for **dele** predicted by best 10-fold LRM model (given by previous expression) and the real average value assigned by experts, at utterance level. We used only the data for the four speakers which have been fully evaluated and incorporated the scores of the three evaluators with highest consistency. The dotted line represents the best linear fit $\textbf{dele} = (0.953 \pm 0.055) * \textbf{LRM}(\textbf{dele}) + (0.15 \pm 0.18)$, with $r = 0.985$. Mean absolute error of LRM fit was 0.7473. This means that most of the time the **dele** average score assigned by experts and the value predicted by the model differed in less than the step between consecutive DELE levels (± 1). In practice, this means that the model and the experts predict basically the same DELE level in most cases.

6. Conclusions and Future Work

In this work, we have presented the SAMPLE corpus, an annotated speech database for L2 Spanish. Although the corpus is still at an early stage of development, it is designed to offer interesting research opportunities for CAPT research on Spanish.

We combined standard Automatic Speech Recognition Technology and perceptual evaluation in order to evaluate the degree of correlation between a quantitative score of the recognizer and the qualitative assessment provided by experts. Five subjective dimensions related to intelligibility, accent, fluency, phonetic accuracy and rhythm have been used. Labelers also provided a proposed DELE level of proficiency in Spanish, using just pronunciation skills over a relatively short lexicon and with a limited amount of speech.

The results presented in this work show good correlation levels among perceptual dimensions and acceptable correlation levels among perceptual dimensions and speech features. Further research on the way to increase inter-rater reliability and a deeper analysis of speech feature selection are expected to provide essentially better figures in the near future.

7. Acknowledgements

We want to thank prof. Pilar Celma for help in conducting the perceptual experiments and in the selection of expert evaluators. We also thank Pablo Camodeca, Clara Manrique, Alba Pérez and Irene González for their participation as evaluators.

8. References

Jian Cheng. 2011. Automatic assessment of prosody in high-stakes english tests. In *INTERSPEECH 2011*, pages 1589–1592. ISCA.

Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Noth, and Satoshi Nakamura. 2009. Automatic pronunciation scoring of words and sentences independent from the non-natives first language. *Computer Speech & Language*, 23(1):65 – 88.

Volker Dellwo and Petra Wagner. 2003. Relations between language rhythm and speech rate. In *ICPhS*, pages 471–474.

D. Escudero, L. Aguilar, A. Bonafonte, and J.M. Garrido. 2009. On the definition of a prosodically balanced copus: combining greedy algorithms with expert guided manipulation. *Revista de la Sociedad Española de Procesamiento del Lenguaje Natural*, 43:93–102, September.

David Escudero-Mancebo and Valentín Cardeñoso-Payo. 2013. Desarrollo de una aplicación móvil de ayuda a la mejora de la pronunciación del español como lengua extranjera basado en reconocimiento de voz. In *III Congreso Internacional el Español Global*.

M. Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

Juan María Garrido, David Escudero, Lourdes Aguilar, Valentín Cardeñoso, Emma Rodero, Carme de-la Mota, César González, Carlos Vivaracho, Sílvia Rustullet, Olatz Larrea, Yesika Laplaza, Francisco Vizcaíno, Eva Estebas, Mercedes Cabrera, and Antonio Bonafonte. 2013. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language Resources and Evaluation*, pages 1–27.

César González-Ferreras, David Escudero-Mancebo, Carlos Vivaracho-Pascual, and Valentín Cardeñoso-Payo. 2012. Improving automatic classification of prosodic events by pairwise coupling. *IEEE Transactions on Audio, Speech and Language Processing*, 20(7):2045–2058.

E. Grabe and E.L. Low. 2002. Durational Variability in Speech and the Rhythm Class Hypothesis. In *Laboratory Phonology VII*, pages 515–546.

Florian Hönig, Anton Batliner, Karl Weilhammer, and Elmar Noth. 2010. Automatic assessment of non-native prosody for english as l2. In *Speech Prosody*.

Florian Hönig, Anton Batliner, and Elmar Nöth. 2011. How Many Labellers Revisited - Naives, Experts, and Real Experts. In Helmer Strik, Catia Cucchiarini, Rodolfo Delmonte, and Rocco Tripodi, editors, *Proceedings of the ISCA Special Interest Group on Speech and Language Technology in Education*.

Y. Kim, H. Franco, and L. Neumeyer. 1997. Automatic Pronunciation Scoring of Specific phone Segments for Language Instruction. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 649–652.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2–3):83–93.

Franck Ramus, Marina Nespor, and Jacques Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292.

Carlos Teixeira, Horacio Franco, Elizabeth Shriberg, Kristin Precoda, and M. Kemal Sönmez. 2000. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In *INTERSPEECH 2000*, pages 187–190. ISCA.

Joseph Tepperman and Shrikanth S. Narayanan. 2008. Better nonnative intonation scores through prosodic theory. In *INTERSPEECH 2008*, pages 1813–1816.

Diego Vallejo-Alonso. 2013. Evaluación de la entonación en la enseñanza del español como segunda lengua. Technical Report TFG-G287, Escuela Técnica Superior de Ingeniería Informática, Universidad de Valladolid, Valladolid, Spain, September.

T. Visceglia, Chiu yu Tseng, Mariko Kondo, Helen Meng, and Yoshinori Sagisaka. 2009. Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project). In *Oriental COCOSDA International Conference on Speech Database and Assessments*.

Silke M. Witt. 2012. Automatic Pronunciation Error Detection: A Review of the current State–of–the–art. In Olov Engwall, editor, *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, pages 1–8, Stockholm, Sweden, June. KTH Department of Speech, Music and Hearing.

Yoichi Yamashita, Keisuke Kato, and Kazunori Nozawa. 2005. Automatic scoring for prosodic proficiency of english sentences spoken by japanese based on utterance comparison. *IEICE Transactions*, 88-D(3):496–501.