INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

EpaDB: a database for development of pronunciation assessment systems

Jazmín Vidal[1,2], *Luciana Ferrer*[2], *Leonardo Brambilla*[1]

[1]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
[2]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina

jvidal@dc.uba.ar, lferrer@dc.uba.ar, lbrambil@dc.uba.ar

Abstract

In this paper, we describe the methodology for collecting and annotating a new database designed for conducting research and development on pronunciation assessment. While a significant amount of research has been done in the area of pronunciation assessment, to our knowledge, no database is available for public use for research in the field. Considering this need, we created EpaDB (English Pronunciation by Argentinians Database), which is composed of English phrases read by native Spanish speakers with different levels of English proficiency. The recordings are annotated with ratings of pronunciation quality at phrase-level and detailed phonetic alignments and transcriptions indicating which phones were actually pronounced by the speakers. We present inter-rater agreement, the effect of each phone on overall perceived non-nativeness, and the frequency of specific pronunciation errors.

Index Terms: computer assisted language learning, phone-level pronunciation assessment, resources.

1. Introduction

Using computers to help students learn and practice a new language has long been seen as a promising area for the use of speech processing technology. It could allow spoken language to be used in many ways in language-learning activities, for example by supporting different types of oral practice and enabling feedback on various dimensions of language proficiency, including language use and pronunciation quality. Moreover, freely distributed through public educational programs, it could reach those less privileged sectors of society that otherwise would not have the possibility to study a second language.

A desirable feature for a computer-aided language learning (CALL) system is the ability to provide meaningful feedback on pronunciation quality. Many systems have been proposed in the last decades that produce pronunciation scores for each paragraph, phrase, word or phone pronounced by the student [1, 2, 3]. Some of the approaches in the literature reach agreements with human annotators that are comparable to those across humans when scores are computed over long chunks of speech [1]. Yet, word- and phone-level scoring are still challenging tasks, with performances that are far from that of human annotators. In practice, scoring at phone level is, arguably, the most useful approach, since can provide detailed feedback on where the errors are and what to do to fix them.

The development of pronunciation assessment systems requires the use of databases of non-native speech with pronunciation quality annotated at the same resolution (speaker, phrase, word or phone-level) as the desired predictions. This data is used to estimate the system performance and, for some systems, to train the prediction models. Several non-native speech databases annotated at phrase- or speaker- level are used in the literature: ATR-Gruhn [4], ERJ [5], Tokyo-Kikuko [6] and, Hispanic-English Database [7].

Examples of databases with phone-level annotations are SRIs Latin American Speech database [8] and TBALL (Technology Based Assessment of Language and Literacy) [9], a corpus of Latin American speakers of English. To our knowledge, none of these databases are publicly available for research use. For this reason, we set out to collect a database for pronunciation scoring with detailed phone-level annotations focusing on native Spanish speakers from Argentina reading English phrases.

Our end goal is to use the collected database to develop a phone-level pronunciation assessment system for use by Argentinian children and adults. Furthermore, we will release the database free of charge for research use. Requests for the databases should be submitted to the first author by email.

2. Database Description

We collected a total of 3200 utterances from 50 Spanish speakers from Argentina, 25 women and 25 men. The participants were asked to record 64 short English phrases designed to contain a balanced sample of English phones with an emphasis on those phones that are more likely to be mispronounced by native Spanish speakers from the region. The prompts were designed to be short in length to allow speakers from different levels of English literacy and age to participate. We targeted Argentinian Spanish only. Specifications such as age, gender, region of birth, and living place were also collected.

The set of common pronunciation problems was determined by a Linguist and an English instructor based on a contrastive analysis of English and Spanish phonetic systems and the experiences from previous works [8, 9, 10]. Examples of such hard-to-pronounce English phones for Spanish speakers are the English flap sound [r] pronounced in some contexts as an alveolar trill [rr] or the devoicing of the English labio-dental fricative [v] resulting in an [f]. In Section 3.1 we show which phones were mispronounced more frequently by our subjects.

2.1. Collection Protocol

The speech data was digitally recorded on the personal computers of each participant through an online application developed for the task in order to mimic the envisioned use scenario where users will be practicing their pronunciation at home with their own computers. The application required registration using an email account so that the recordings could be done in more than one session. There was no payment for the subjects, but a raffle was made of an English book among the 50 participants.

When subjects first registered in the system they were presented with an informed consent form and were asked for personal information including age, gender, region of birth, and place of living. They were then presented with instructions, including a request to do the recording in a relatively quiet envi-

http://dx.doi.org/10.21437/Interspeech.2019-1839

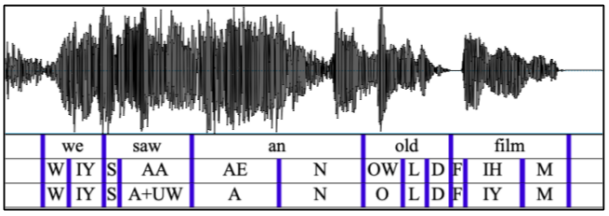

Figure 1: *The PRAAT interface used for annotation showing one example phrase for which the transcription includes one insertion (in "saw"), one deletion (in "film"), and two substitutions (in "an" and "old"). The deletion cannot be seen because it has duration 0.0, but the tier contains a label of "0" which aligns with the target phone [L]. Three tiers are not shown here to save space: the spectrogram, the one used to record the phrase-level nativeness score, and an extra tier for general comments about the waveform (e.g., to indicate issues like background noises). The remaining three are, from top to bottom the words, the canonical transcription and the annotation.*

ronment. After that, a series of four screens were shown including 16 phrases each, organized by levels of difficulty. The user had to read the phrases, recording them with a push-to-talk-like interface. Recordings were stored using an anonymous identification number for the speaker.

2.2. Annotations

As the first step in the annotation process, a Kaldi-based open source forced aligner was used to generate time-aligned versions of the audio files from orthographic transcriptions. The Montreal Forced Aligner [11] offers a pre-trained acoustic model trained on the LibriSpeech corpus [12] and an English pronunciation dictionary based on the ARPAbet phone set [13]. This results in a canonical phone-level annotation, which is used to initialize the annotation tier that will then be modified by the annotators.

The task of transcribing speech at phone level with the desired level of detail is a complex one, which is prone to subjectivity and errors. For this reason, we recruited two annotators: a linguist and an English instructor both with a strong background in phonetics and native-like English expertise. In the near future we plan to add a native English speaker as third annotator. The annotators were asked to narrow down the canonical phonemic transcriptions provided by the forced aligner to a phonetic level. The task involved representing the phones with additional details about the contextual variations in pronunciation that occur in normal speech using the ARPAbet symbols and an extra set of extensions needed to represent the non-native variants encountered in our data. The annotation was done using PRAAT [14]. One of the annotators labelled all utterances, while the second one only labelled 4 phrases per speaker for this first version of the database. She will annotate all phrases in the near future. Figure 1 shows the interface used for annotation with one example phrase with several pronunciation errors.

Annotators were asked to flag utterances that were truncated, did not match the phrase they were asked to read, or had other problems not due to pronunciation (e.g., severe noise masking the speech). The remaining utterances were labelled with utterance-level ratings corresponding to perception of nativeness on an ordinal scale of 1 to 5 and with phone-level alignments and transcriptions, as described in more detail in the next sections.

2.2.1. Phone Level Annotation

Our primary goal is to create a database to investigate methods for phone-level pronunciation scoring. In particular, we want to explore L1-dependent methods, which are trained on non-native data with pronunciation annotations such as the works in [15, 16, 17]. To this end, we annotated all recordings with detailed phonetic transcriptions, which will allow us to train and evaluate phone-level pronunciation scoring systems. In this process, we encountered two main challenges: defining the appropriate level of detail in the annotations and finding a common set of symbols to transcribe both native-sounding and non-native-sounding phones. We closely followed the work in [8] to address these challenges.

First, we decided that a narrow phonetic transcription was needed to describe the standard errors of Spanish English learners correctly. Apart from phone deletions, insertions and substitutions, non-native speakers also change the place of articulation of several phones, a phenomenon that could significantly affect the perception of nativeness. As a consequence, the data was transcribed taking into account those allophones necessary to pinpoint the emerging problems. This task was eased by the fact that all non-natives shared the same L1 and thus a standard set of pronunciation problems was expected.

Having decided on the allophones that we needed to use, we went on to define the set of necessary symbols. We started with the ARPAbet symbols (used by the Montreal Forced Aligner and popularized by the CMU pronunciation dictionary [18]) and extended them with a set of extra ARPAbet-like symbols to represent expected and unexpected non-native variants. A catch-all diacritic, '*', was included to represent a sound that was perceived as a nonnative rendition of a phone but for which no specific allophone was available. Insertions were handled by adding the new phone to the expected one with a '+' (plus sign) and deletions were marked as '0' (zero) (see Figure 1 for an example of both cases). In this way, annotators were in charge of deciding which specific phone had been deleted or converted into two sounds. This avoided the problem of doing automatic alignments between the canonical transcription and the annotated transcription to determine the mapping between the target phones and the produced ones. This is a non-trivial problem that requires a set of heuristic rules to be solved, which we wanted to avoid.

Table 2 shows the English ARPAbet and the set of ARPAbet-like extensions for vowels. Table 1 shows the corresponding version for consonants.

Annotators were also asked to determine the alignment of each phone to the waveform. This was done using the forced-alignment as a basis and correcting the phone boundaries only when necessary. Deletions were turned into 0-length phones (with a label of '0', as mentioned above). Insertions, indicated by concatenating the inserted phone with the phone it was inserted to (as judged best by the annotator) were left with an undefined boundary between the two phones separated by the '+' sign. This could prove to be a problem, depending on the specific method used to implement the pronunciation scoring system. We might eventually add this information in the database if necessary.

2.2.2. Utterance Level Annotation

In addition to aligning and annotating utterances at phone-level, annotators were asked to rate each phrase according to their perceived level of nativeness. In an attempt to unify the criteria used by both annotators to assign these scores, they were pre-

Table 1: *ARPAbet symbols for consonants with extensions. Extensions are: [Ph/Th/Sh/Kh], for aspirated versions of [P/T/S/K]; [TE/DE] for Spanish dental [T/D]; [BH/GH] for Spanish affricate [β/γ]; [RR], [X] and [LL] for Spanish trill and [x] and English dark [l]. Where symbols appear in pairs, the one to the right is voiced. "-" indicates articulations judged impossible.*

	Bi-Labial	**Labio-Dental**	**Dental**	**Alveolar**	**Post-Alveolar**	**Palatal**	**Velar**	**Glottal**
Stop	P/Ph B	-	TE DE	T/Th D	-	-	K/Kh G	-
Affricate	BH	-	-	-	CH JH	-	GH	-
Nasal	M	-	N	-	-	-	NG	-
Fricative	-	F V	TH DH	S/Sh Z	SH ZH	-	X	HH
Approximant	-	-	R	-		Y	-	-
Flap	-	-	DX	-	-	-	-	-
Trill	-	-	RR	-	-	-	-	-
Lateral	-	-	L	-	-	-	LL	-

Table 2: *ARPAbet symbols for vowels with extensions. Extensions are: [A/E/O] for Spanish [a/e/o]. Spanish [i/u] were judged equivalent to English [IY/UW]. Plus diphtongs [AY/EY/OY/AW] not included in the table.*

	Front	**Central**	**Back**
	IY	-	UW
High	IH	-	UH
	E	-	O
Mid	EH	AX	AO
Low	AE	AH A	AA

sented with a small set of phrases to rate and asked to discuss their decision for these cases until they converged on a criterion.

2.3. Reference Transcriptions

For each utterance, a phonetic transcription was created as a reference. Transcriptions were made using the ARPAbet set of symbols. Common alternatives in Standard American English pronunciation were considered for each phrase. Any of these alternatives would be considered as correctly pronounced. For example, "I don't like rainy days" may be transcribed as:

AY D/DX OW N T/0 L AY K R EY N IY D/DX EY Z

where the "/" is used to indicate that either allophone (or a deletion in the case of the [T] in "don't") should be considered correct in that context. Elisions in co-articulation are considered as correct, but more colloquial contractions are not included since they are not expected to occur in read speech.

Reference transcriptions for different variants of English were not created since all subjects annotated for this version of the database were clearly targeting the Standard American English pronunciation. Eventually, we will add references for the other most common variant learned in Argentina, Standard British English. In practice, each student would be targeting one of the two variants, and the corresponding reference transcriptions would be selected for each case.

3. Statistics and Analysis

The phone level transcriptions in this database will be used to train automatic systems for pronunciation assessment. Our goal is to know, given a prompt, if a speaker was able to produce the expected phones correctly, and if not, which variations occurred. This level of detail will allow us to give detailed feedback to the students about their mistakes, imitating human teachers.

For each target phone in the reference transcription, there were four options: the speaker may have produced the native phone, produced a nonnative version of that phone, substituted the correct phone for a native or nonnative version of a different one, or deleted the phone altogether. We determined which of these possibilities happened for each expected phone by comparing the transcriptions of each utterance to the best matching reference transcription for the phrase.

The resulting analysis is presented in the following sections. The presented statistics correspond to the first release of the EpaDB database. A document containing these tables will be included along with the release of the database and updated with each new version.

3.1. Frequency of non-native pronunciations

We started by computing the percent of times that each phone was mispronounced in our database. For this purpose we used the labels from the annotator that labelled all phrases from every speaker. We computed the percent of times a target phone was mispronounced by comparing its transcription with the best-matching reference. If the transcriptions for that phone differed in any way, the phone was considered mispronounced. Going downwards in Table 3, we can see that phones that were more frequently mispronounced are not part of the Spanish phonetic inventory, with the only exceptions of [K] and [T]. These phones are hard to pronounce even though they exist in both phonetic inventories. Spanish speakers voice them in English final position since they never occur there for them. The phones that are harder to pronounce can be classified into vowels such as [AX/AH/AE/IH] and consonants [LL/Z/Ph/V/K]. English distinguishes vowels by duration, a feature that goes mostly unnoticed by Spanish speakers causing a proliferation of variations in production. Consonants, on the other hand, imply changes in points and manners of articulation alien to nonnatives and sometimes hard to imitate.

3.2. Annotator Agreement

In order to measure the agreement across the two annotators, we computed the kappa coefficient [19] for each target phone using the four phrases per subject (200 phrases total) that were transcribed by both annotators. The last column in Table 3 shows the kappa values. Note that 11 of the 46 target phones did not appear in the set of 4 phrases labelled by both annotators and, hence, lack a kappa value. Once the second annotator finishes transcribing all phrases, we will update these values in the document released with the database. Furthermore, some kappa values were computed with very few incorrectly pronounced cases, as labelled by either annotator, making the estimation unreliable. These cases (when both annotator found less than

Table 3: *Statistics for each target phone: total number of occurrences, percent of non-native pronunciations, correlation between the scores computed from the phone-level errors and the nativeness score at phrase level, and the kappa coefficient. For an explanation of the asterisks, the "-" and the missing values in the kappa column see Section 3.2. Rows with grey background correspond to phones that also exist in the Spanish inventory.*

Phones	Total	Non-Nat%	Score corr	Kappa	
AO	232	0.0	-	0.00	*
DX	81	0.0	-	0.00	*
F	475	0.6	0.08	-	
B	348	0.9	0.21	-	
M	1050	1.2	0.33	-	
N	1324	1.5	0.24	-0.02	*
L	850	1.6	-0.33	0.00	*
SH	375	1.9	0.46	-	
W	475	2.1	0.47	-	
DH	967	2.9	0.49	0.25	*
S	1024	3.0	0.59	-	
CH	234	3.4	0.31	-	
IY	948	5.0	0.38	0.39	*
G	374	5.3	0.43	-	
AY	899	5.8	0.12	0.00	*
UW	768	7.1	0.53	0.38	
Y	423	8.3	0.71	1.00	*
P	449	9.8	0.04	0.21	*
OY	174	9.9	0.51	-	
AW	243	12.7	0.53	0.66	*
EY	382	13.0	0.63	0.00	*
UH	224	13.4	0.51	-	
R	1098	16.7	0.52	0.46	*
EH	691	17.4	0.78	-	
HH	399	18.0	0.70	0.31	
K	450	18.9	0.38	0.20	-
TH	181	19.0	0.36	0.34	*
D	659	23.6	0.79	0.02	*
ER	820	33.5	0.38	0.31	
AH	424	34.6	0.86	0.21	
T	1092	36.5	0.80	0.24	-
JH	150	36.7	0.60	0.34	
NG	250	38.8	0.70	-	
OW	339	38.9	0.74	0.37	
Th	550	45.1	0.58	0.30	
IH	1362	45.7	0.89	0.27	
AX	1903	55.6	0.93	0.19	
V	374	59.6	0.92	0.18	
ZH	150	63.3	0.76	-	
AA	790	66.2	0.89	0.57	
AE	477	71.5	0.88	0.24	
Kh	324	72.3	0.38	-	
Ph	174	79.3	0.55	-	
Z	723	82.9	0.79	0.45	
LL	399	84.9	0.40	0.07	

5 errors) are marked with an asterisk in the table. Finally, some cases had exactly 0 incorrect cases for both annotators resulting in an undefined kappa value and are marked with a "-".

The phones for which a somewhat more reliable kappa can be computed all have kappa values between 0.07 and 0.57. Indicating a different ranges of agreement.

We will continue to analyze the data for the phones with low agreement to further understand this phenomena.

3.3. Effect of each phone on overall perceived non-nativeness

Finally, we measured the influence that mispronouncing a certain phone had in the overall perceived nativeness of a speaker, estimated as the correlation between two scores computed for every speaker. Being the first score the average of the phrase-level nativeness score provided by the first annotator and the second, the percent of times that the speaker correctly pronounced the target phone. The correlation between those two scores is given in the fourth column of Table 3.

The correlation for the phones that had low non-native percents could not be reliably estimated since the score based on the percent of correctly pronounced cases was almost fixed at 100%, with only small deviations, leading to noisy estimates of the correlation that may even be negative. For the frequently mispronounced phones, where this correlation could be reliably estimated, it was generally high, with few exceptions such as [Kh], [LL] and [ER]. The first two are exchanged by their corresponding allophones [K], [L] producing changes in in sound but not in meaning, which seems to be less penalized. In the latter, the same effect is caused by the non-native substitution [E+DX].

4. Conclusions

We introduced EpaDB, a database of 3200 English short utterances produced by 50 Spanish speakers from Argentina annotated at detailed phonetic level. The database is intended for the development of pronunciation assessment systems and will be freely available for research purposes. We showed inter-rater agreement, measured the influence of specific phones in the overall perception of nativeness of a speaker and calculated which phones are most likely to be mispronounced by our target population.

We found agreement to be between fair and moderate for most phones for which enough correctly and incorrectly pronounced samples were available to compute the kappa coefficient. Furthermore, we saw that the phones more frequently mispronounced were those absent from the Spanish inventory and that those that most affected the perception of nativeness were [AX], [V] and [AE]. Finally, we saw that English learners from Argentina find it more difficult to pronounce those phones that are not part of the Spanish phonetic inventory.

In the near future, we plan to add more speakers, adults and children, complete the ratings by the second annotator and add ratings from a third annotator. All updates will be made public as new versions of the database.

5. Acknowledgments

This work was partially supported by Google under the Google Research Awards for Latin America, 2018.

6. References

[1] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub, "Automatic evaluation and training in english pronunciation," in *First International Conference on Spoken Language Processing*, 1990.

[2] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1471–1474.

[3] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[4] R. Gruhn, T. Cincarek, and S. Nakamura, "A multi-accent non-native english database," in *ASJ*, 2004, pp. 195–196.

[5] S. R. Consortium *et al.*, "Ume-erj english speech database read by japanese students," 2002.

[6] K. Nishina, Y. Yoshimura, I. Saita, Y. Takai, K. Maekawa, N. Minematsu, S. Nakagawa, S. Makino, and M. Dantsuji, "Development of japanese speech database read by non-native speakers for constructing call system," in *Proc. ICA*, 2004, pp. 561–564.

[7] L. M. Tomokiyo and S. Burger, "Eliciting natural speech from non-native users: collecting speech data for lvcsr," in *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*. Association for Computational Linguistics, 1999, pp. 5–11.

[8] H. Bratt, L. Neumeyer, E. Shriberg, and H. Franco, "Collection and detailed transcription of a speech database for development of language learning technologies," in *Fifth International Conference on Spoken Language Processing*, 1998.

[9] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[10] P. Ladefoged and K. Johnson, *A course in phonetics*. Nelson Education, 2014.

[11] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[13] A. Klautau, "Arpabet and the timit alphabet," 2001.

[14] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer, version 3.4," *Institute of Phonetic Sciences of the University of Amsterdam, Report*, vol. 132, p. 182, 1996.

[15] W. Hu, Y. Qian, and F. K. Soong, "A new neural network based logistic regression classifier for improving mispronunciation detection of l2 language learners," in *The 9th International Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 245–249.

[16] ——, "A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call)." in *Interspeech*, 2013, pp. 1886–1890.

[17] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[18] R. Weide, "The cmu pronunciation dictionary, release 0.6," 1998.

[19] S. Siegel and N. Castellan Jr, "Jr. nonparametric statistics for the behavioral sciences," *Nonparametri˜ Statistics for the Behavioral Sciences, McGraw-Hill, New York*, pp. 190–222, 1988.