Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus

Mario Carranza

Departamento de Filología Española, Universitat Autònoma de Barcelona, Edifici B, Campus UAB, Cerdanyola del Vallès, 08193 Bellaterra, Barcelona, Spain

Abstract

This paper addresses the issue of manual transcription of non-native speech in an attempt to establish rule-based strategies for labelling intermediate realizations. The problems of transcribing non-canonical realizations of L2 sounds which present shared features of the target (Spanish) and the source language (Japanese) will be considered. We introduce a Japanese accented non-native L2 Spanish corpus, and exemplify the use of decision trees in manual transcriptions as a systematic method for dealing with ambiguous realizations. This approach could help a potential error detection system to detect both canonical and erroneous realizations, contributing to the development of CAPT tools.

Index Terms: non-native speech transcription, ASR, CAPT, L2 Spanish, L1 Japanese, non-native spoken corpus

Introduction

Adapting Automatic Speech Recognition (ASR) systems to non-native speech raises several difficulties. Since canonical native speech is used during the training phase in general-purpose recognizers, the generated acoustic models cannot fit properly with the non-native data and this leads to an increase in Word Error Rate (WER). Many proposals have been put forward to find methods of adapting the ASR systems to non-native data ([1], [2]). The adaptation can be done at the HMM level or by means of knowledge-based rules, that can automatically generate variants which are incorporated in the pronunciation dictionary ([1]). A pronunciation error detection system can be developed by generating new acoustic models for the non-native realizations of L2 phonemes and by the systematization of L1-based typical errors by means of rules ([1], [2]). In order to do so, phonetically transcribed non-native spoken corpora are needed; however, manual transcription of non-native speech is a time-consuming costly task, and current spoken corpora are needed; however, manual transcription of non-native speech is a time-consuming costly task, and current spoken corpora are not available. In order to resolve this problem, we will describe a non-native corpus compiled for this research. The difficulties of labelling and transcription of intermediate realizations and the use of decision trees for disambiguation are presented in section 2, followed by a conclusion in section 3.

1. Database description

A Japanese accented L2 Spanish corpus was compiled, transcribed and manually annotated in view of its future adaptation as a training corpus for developing CAPT tools oriented to Japanese learners of Spanish. The data was obtained from the recordings of oral exams at the Tokyo University of Foreign Studies. Pronunciation errors were systematically encoded using a notation system suited to the automatic processing of errors and to the development of transformation rules.

1.1. Contents

The recorded data consists of 8.6 hours of spontaneous, semi-spontaneous and read speech. Each participant was recorded four times—once every 6 months—throughout the first two academic years at the university. The recordings took place from April 2010 to March 2012. Twenty students (10 male and 10 female) participated in the project. The participants were carefully selected to be representative of the population; therefore, students were chosen considering, among other criteria, their oral proficiency level: 6 students with a high oral proficiency level, 8 students with an intermediate level and 6 students with a low oral proficiency level.

The corpus contains different tasks designed according to the learning stage: semi-spontaneous speech (students were asked to speak about a topic given some weeks in advance) followed by a conversation with the examiner after 6 months and after 12 months of study; spontaneous speech (students were asked to act in a daily situation with no previous preparation), a conversation and a reading task after 18 months; finally, two samples of spontaneous speech and a reading task were required after 24 months.

The recordings were segmented into individual audio files that were transformed to WAV format and labelled using a code which represents speaker number, test period and task type. This will allow to automatically compute error rates according to proficiency level, learning stage and task. The total recording time sorted by task type and learning stage is shown in Table 1.
### 1.2. Transcription levels and error annotation

The recordings were manually segmented and transcribed using Praat ([4]). The TEI conventions ([5]) and the EAGLES guidelines ([6]) for the orthographic transcription and encoding of spontaneous speech were followed. The text was annotated in XML format; some specific labels were added to cover non-native speech phenomena. Since the corpus is aimed at the analysis of errors produced by non-natives speakers, a canonical phonemic level representing the standard pronunciation of the word, and a narrow phonetic transcription level for the actual pronunciation by the speaker were included ([7]); the alignment of both levels is intended to allow the automatic generation of pronunciation variants and the retrieval of statistical information for all words in the corpus. Finally, an error tier to label and encode the pronunciation errors for their further processing plus two additional tiers for non-lexical phenomena were added. In total, the transcription consists of six levels of representation, as shown in Figure 1.

**Figure 1: Levels of transcription and annotation.**

#### 1.2.1. Orthographic transcription

In the orthographic tier each word was transcribed in its standard orthographical form. No punctuation marks were used due to the difficulties of establishing sentence boundaries in spontaneous speech. The vocal content, such as filled pauses, was also orthographically transcribed in a standardized form and labelled in the vocal tier (see 1.2.4). XML labels were used to mark phenomena such as foreign words, non-existing words, repetitions, truncations, unclear and unintelligible utterances.

#### 1.2.2. Canonical phonemic transcription

The canonical phonemic transcription level shows the phonemic representation of words as pronounced in isolation. Thus, coarticulatory phenomena taking place within words and at word boundaries are not considered. Northern Castilian Spanish was adopted as the standard form for the transcription; although some of the Japanese students received input from speakers of other Spanish dialects, Castilian Spanish was the main variety used in their courses. An adaptation of SAMPA to Spanish ([8]) was used for the transcription; the symbols corresponding to allophonic variants were not considered in this level. Only linguistic content was phonemically transcribed; the noise-corrupted cases were marked as "unclear" in the orthographic transcription and left with no phonemic transcription.

#### 1.2.3. Narrow phonetic transcription

The actual pronunciation of words by the speakers is presented in the narrow phonetic transcription tier. In order to better reflect the Japanese pronunciation of L2 Spanish, 11 new symbols and 7 diacritics from X-SAMPA ([9]) were incorporated to the initial inventory. A preliminary version of the narrow phonetic transcription was automatically generated from the canonical phonemic tier. The resulting new tier was manually checked, realigned and corrected; the additional X-SAMPA symbols and diacritics were used to represent the Japanese pronunciation of Spanish sounds.

#### 1.2.4. Other levels of representation

Vocalized but non-lexical content (semi-lexical elements according to TEI [5]), such as hesitations (including filled pauses), laughter and interjections were marked in the vocal tier. Since the acoustic realization of these phenomena is usually very similar to that of linguistic sounds—such as long vowels for the hesitations—the aim of doing this was to explicitly identify the segments which should not be considered for the development of the acoustic models in the ASR training. Other phenomena such as coughs, breathing, ambient noise, as well as overlapping speech of the examiner (non-lexical phenomena in the TEI guidelines) are marked in the incident tier.

#### 1.2.5. Error annotation

Whenever the canonical phonemic tier and the narrow phonetic tier presented a dissimilarity (not derived from the Spanish coarticulation rules), this was considered an error to be included in the error tier. All phones in the transcriptions have a corresponding two digit code used for error annotation. The encoding procedure is similar to that proposed in [10] and consists on a string of six numeric characters separated by (space) and (.) symbols. The first two digits correspond to the code of the affected phone, separated by a # character; the following four represent the previous and following sounds (the phonologic context of the error) separated by a _ character; finally, a letter (a, b or c) was added at the beginning of the string to represent the type of error. The letter “a” corresponds to errors of substitution of one phone by another; “b” represents errors of insertion; and “c” stands for the deletion of one phone that should have appeared in the Spanish canonical form. Since this annotation system was originally created to cover all type of errors, it is not possible to encode the resulting phone in substitution cases; however, as the error label is aligned with the narrow phonetic and canonical phonemic transcription tiers in the corpus, this information can be automatically recovered. Figure 2 shows an example of the error encoding. In this case, a substitution of the target phone [e] (code: 02) by the phone [i] has occurred between the phones [β̞] (26) and [s] (23).

The error annotation system combined with the speaker and task type codification in the file name were used to automatically quantify error types considering also the affected phone and the phonological context. Frequency of occurrence and likelihood ratios for each error will be also available after the processing of the corpus, and can eventually become a useful source of statistical information in the generation of data-driven pronunciation rules.
2. Intermediate realizations

One of the difficulties in the manual transcription of the corpus is due to ambiguous realizations of L2 sounds that present features of both the target and the source language. In these cases transcribers should establish clear criteria for deciding whether the realization should be classified as correct or incorrect based on the acoustical information, since perceptual judgments might be biased and the level of disagreement might be high if more than one transcriber is involved ([6]).

During the transcription of the corpus three types of intermediate categories have been detected, depending on the phonetic or phonologic status of the related sounds.

2.1. Realizations between two L2 phonemes

In Spanish there is a phonological contrast between the alveolar lateral approximant /l/ and the alveolar rhotic tap /ɾ/. In Japanese, both phones are possible realizations of the alveolar tap phoneme, represented also by the symbol /ɾ/ ([11]). Although several studies have shown that there is a context-dependent preference for one realization or the other ([11], [12]), the realization of this phoneme seems to exhibit a relatively high degree of individual variation ([11], [12]). Furthermore, some cases of an intermediate category [ɺ] (lateral flap) have been detected in the corpus (Figure 3). Intermediate realizations present formant continuity, typical of the lateral realization [l] and a closure release, more characteristic of the tap realization [ɾ].

In order to distinguish between [l] or [ɾ] realizations in intervocalic position, duration must be taken into account, as segmental duration seems to be longer in lateral than in rhotic segments; this tendency is also found in native Spanish ([13]). Therefore, a speaker-based duration threshold can be established as a way of disambiguating intermediate realizations.

In consonant cluster position some realizations that might be described as showing mixed features of both sounds have also been detected. Figure 4 shows a case of vowel epenthesis before [l] phone insertion before the [l] phone.

2.2. Realizations between two L2 allophones

The Spanish phonemes /b/, /d/ and /g/ present two allophonic variants: stops ([b], [d], [g]) after a pause or a nasal consonant ([d] can also appear after [l]), and approximants ([β], [ð], [ɣ]) in all other contexts ([14]). Approximant allophones for these phonemes do not exist in Japanese, so Japanese students of L2 Spanish tend to replace the approximant realizations by their stop counterparts. Nevertheless, intermediate realizations can show the formant continuity characteristic of approximant realizations as well as the closure release typical of stops (Figure 5). Since no clear difference in segment duration between the two allophones has been detected in the non native realizations, alternate ways of distinguishing between stop or approximant realizations should be considered.

In consonant cluster position some realizations that might be described as showing mixed features of both sounds have also been detected. Figure 4 shows a case of vowel epenthesis before [l] phone insertion before the [l] phone.
2.3. Realizations between one L2 phoneme and one L1 phoneme

Intermediate categories can also be the result of the realization of new L2 phonemes which have no correspondence in the L1 inventory, but are similar to one or more L1 categories. For example, the Spanish unvoiced fricative velar phoneme /x/ is usually assimilated by Japanese speakers to the Japanese fricative phoneme /h/, which presents four allophones depending on the following vowel ([12]): voiceless/voiced glottal fricatives [h] / [ɦ], voiceless bilabial fricative [ɸ], and voiceless palatal fricative [ç]. Furthermore, some mixed realizations between these sounds have also been found in the corpus. Figure 7 represents an example of an intermediate realization between a velar fricative [x] sound and a palatal fricative [ç] sound. The decision tree shown in Figure 8 for the disambiguation of these sounds takes into account the following vowel as well as spectral features to disambiguate between various possible realizations of the targeted [x] sound.

3. Conclusion

The adaptation of ASR technology to non native speakers requires a considerable amount of phonetically transcribed non native data; moreover, if the technology is to be applied in computer assisted pedagogical applications, it should be able to successfully detect mispronunciations of targeted sounds. In order to achieve so, a detailed phonetic transcription of the data is required, but manual transcription faces some difficulties when dealing with non native speech.

This paper provides a description of a Japanese accented Spanish L2 spoken corpus, phonetically annotated considering its future application to the development of a Spanish CAPT system aimed at Japanese speakers. During the manual transcription of the corpus intermediate realizations of L2 phonemes were documented, showing the need for disambiguation strategies. We propose the adoption of a decision tree for each case of intermediate realization, based on acoustic criteria. This approach is aimed at helping transcribers in choosing the adequate phone label in narrow phonetic transcription of L2 speech. Furthermore, given that decision trees can be mathematically expressed as algorithms, once manually tested they can be implemented for automatic corpus processing. The advantage of decision trees relies in the fact that they can be systematically applied; since they are not based on perceptual criteria, their adoption could contribute to overcome some of the difficulties found in the transcription of non native speech.

4. References