Grapheme-to-phoneme transcription rules for Spanish, with application to automatic speech recognition and synthesis

Patrizia Bonaventura
Cluster Reply
Corso Francia 110
Turin, Italy, 10143
p.bonaventura@reply.it

Fabio Giuliani
Cluster Reply
Corso Francia 110
Turin, Italy, 10143
f.giuliani@reply.it

Juan María Garrido
Departament de Filologia Espanyola
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
juanma@liceu.uab.es

Isabel Ortín
Departament de Filologia Espanyola
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
isabel@liceu.uab.es

Abstract

Large phonetic corpora including both standard and variant transcriptions are available for many languages. However, applications requiring the use of dynamic vocabularies make necessary to transcribe words not present in the dictionary. Also, additional alternative pronunciations to standard forms have shown to improve recognition accuracy. Therefore, new techniques to automatically generate variants in pronunciations have been investigated and proven to be very effective. However, rule-based systems still remain useful to generate standard transcriptions not previously available or to build new corpora, oriented chiefly to synthesis applications. The present paper describes a letter-to-phone conversion system for Spanish designed to supply transcriptions to the flexible vocabulary speech recogniser and to the synthesiser, both developed at CSELT (*Centro Studi e Laboratori Telecomunicazioni*), Turin, Italy. Different sets of rules are designed for the two applications. Symbols inventories also differ, although the IPA alphabet is the reference system for both. Rules have been written in ANSI C and implemented on DOS and Windows 95 and can be selectively applied. Two speech corpora have been transcribed by means of these grapheme-to-phoneme conversion rules: a) the SpeechDat Spanish corpus which includes 4444 words extracted from the phonetically balanced sentences of the database b) a corpus designed to train an automatic aligner to segment units for synthesis, composed of 303 sentences (3240 words) and 338 isolated words; rule-based transcriptions of this corpus were manually corrected.

The phonetic forms obtained by the rules matched satisfactorily the reference transcriptions: most mistakes on the first corpus were caused by the presence of secondary stresses in the SpeechDat transcriptions, which were not assigned by the rules, whereas errors on the synthesis corpus appeared mostly on hiatuses and on words of foreign origin.

Further developments oriented to recognition can imply addition of rules to account for Latin American pronunciations (especially Mexican, Argentinian and Paraguayan); for synthesis, on the other hand, rules to represent coarticulatory phenomena at word boundaries can be implemented, in order to transcribe whole sentences.

Introduction

Grapheme-to-phoneme conversion is an important prerequisite for many applications involving speech synthesis and recognition [1]. Large corpora used for these applications (e.g. WSJ, CMU, Oxford Pronunciation Dictionary, ONOMASTICA, SpeechDat) include phonetic transcriptions for both standard pronunciations and for variants, which can represent either differences in dialectal or individual realisation of single words (intra-word variants) [2, 3] or variations in the standard form produced by coarticulation between words (inter-word variants) [4].

These alternative pronuciations have been shown to improve recognition accuracy [5] and they need to be present in large phonetic database: the variants can either be realised manually on the basis of expert phonetic knowledge, or by a rule-based system. However, maintenace of such systems is complex, because insertion of new rules often causes to change the overall performance of the module.
Therefore, new techniques to derive automatically rules for grapheme-to-phoneme conversion from training data have been investigated. Generally rules are obtained through forced recognition, according to the following procedure: 1) aligning of the canonical pronunciation to the alternative ones by means of a dynamic programming algorithm, in order to generate an aligned database 2) use this database to train a statistical model or a binary decision tree to generate variants of words or proper names [1] [3] [6] [5] or to model context-dependent variations at word boundary [4]; neural networks can also be used to generate variants in pronunciation of words [2] or of surnames [7], on the basis of pre-aligned or non-aligned training data [8]. Finally, a mixed approach combining knowledge obtained from training data and *a priori* phonetic expertise has also been experimented to derive possible non-native pronunciations of English and Italian words [9].

All these techniques have proven to be very effective to generate plausible alternatives to canonical ones. However, rule-based approaches can still represent an effective tool to automatically obtain standard transcriptions of large corpora built *ad hoc* for special applications, in particular oriented to synthesis: a letter-to-phone rules component is very suitable to represent allophonic and allomorphic variations [10] [11] [12] which are essential to allow segmentation and diphone extraction from an acoustic database [13].

The rule system described in the present paper was developed on the basis of phonetic knowledge [14] [15] and has two different application domains, which imply different transcription requirements: the recogniser for Spanish uses sub-word units [16] [17] linked to the phonetic representation of isolated words; the units have been trained on the corpus of words extracted from the phonetically balanced sentences included in the SpeechDat database. Therefore, the SpeechDat corpus has been considered as the reference set of words that the conversion rules minimally had to correctly transcribe. Only isolated words were used, with the same phoneme inventory employed in the original SpeechDat transcriptions, including no allophones.
On the other hand, the corpus for synthesis was selected to collect speech material to train the automatic phonetic aligner, in order to extract diphones for a concatenative synthesis system [18] and had to meet different requirements: a) units were to be pronounced both in isolated words and in sentences b) the phoneme inventory had to include the maximum number of allophones so to allow to build a representative acoustic dictionary containing occurences of all units and sequences in every appropriate segmental and prosodic context (stressed and unstressed syllables; initial and final position in the syllable; initial, internal and final position in the sentence; short and long sentences).

Therefore, two partially different sets of rules have been designed for synthesis and recognition, which can be alternatively activated: the latter are a subset of the former. Both systems provide only

one variant in output, i.e. the standard Castilian (Madrid) Spanish pronunciation.

1. Orthographic and phonetic symbols

The orthographic string in input is preprocessed to avoid problems relative to the configuration of the operating system: at this stage, letters with diacritics (corresponding to non-standard ASCII characters) can either be represented by means of extended ASCII (e.g. 'ñ' = ext. ASCII code 241) or they can be converted into a sequence of standard ASCII symbols ('n~' = st. ASCII 110+126). This preprocessing is common to both the recognition and the synthesis systems.

The phonetic symbols used for recognition represent the 30 basic Spanish phones, which are also common to synthesis. For this latter system, however, 11 extra symbols have also been added to represent a set of allophones of the standard Spanish which show an acoustic structure clearly differentiated from the rest of already included allophones. These symbols represent stressed vowels, semivowels (i.e. [i] and [u] allophones in second position of falling diphthongs, distinguished from semi-consonants, or glides, i. e. [i] and [u] allophones in first position of rising diphthongs), the nasal labiodental allophone [ɱ], the palatal voiced stop [ɟ], which accounts for the distribution of the palatal voiced approximant [j], in initial position or after 'l' or 'n'(e.g. 'cónyuge' = [k 'o Gn J u x e]) and the interdental voiced fricative [ð̟], which accounts for the distribution of the the unvoiced interdental fricative [θ̟], in end of syllable before a voiced consonant (e.g. 'llovizna' [L o . B'i Zh . n a]). Finally, two phones typical of most frequent foreign loan words, i.e. the unvoiced dental affricate [ts] (e.g. 'pizza') and the unvoiced palatal fricative [ʃ] (e.g. 'flash') were added. This gives a final set of 43 synthesis symbols.

Front vowels	i, 'i, e, 'e
Central vowels	a, 'a
Back vowels	o, 'o, u, 'u
Semiconsonants	w, j
Semivowels	i~ , u~
Bilabial/labiodental consonants	p, b, β, f, m, ɱ
Dental/alveolar consonants	θ, ð̟, t, d, ð, ts, s, z, n, l, r, r:
Palatal consonants	tʃ, ʎ, ɲ, ɟ, j, ʃ
Velar consonants	k, g, ɣ, x, ŋ

Table 1. Phonetic symbols used for transcriptions (**bold**: synthesis allophones; underlined: phones from loan-words)

2. Rule component

The rule module is composed by a) table look-ups containing pre-stressed roots with hiatuses and words that do not take stress within a sentence b) stress assignment rules c) transcription rules. Vowels are transcribed before consonants. The main complexity in vowel conversion consists in disambiguation of diphthongs and hiatuses: stress position is crucial for correct transcription of these vowel sequences. However, in the rule component, they undergo a different treatment for recognition and synthesis, which is illustrated in the following, before a description of the consonant rules.

2.1. Diphthongs and hiatuses rules

2.1.1.Recognition

Rules for recognition do not transcribe diphthongs and hiatuses according to the stress position. In fact, the SpeechDat transcriptions, that the conversion rules have to reproduce, always stress the first element of a vowel sequence and transcribe all closed vowels as glides (es. 'rehúye' [rr 'e w . jj e], 'reír' [rr 'e j r], 'oír' ['o j r]).

This target can be attained by deterministic rules, that account for three realisations of [u]: a) deletion b) full vowel [u] and c) semivowel [w]. In particular, instance (a) applies when letter 'u' (henceforth letters are included between apices) appears within sequences 'gu', 'qu' before front vowels (e.g. 'burguesía' [b u r . G e . s 'i . a]);

transcription (b) occurs when 'u' either precedes a rising diphthong or it follows a consonant different from 'g', 'q' and it is the stressed first element of a hiatus (e.g. 'cuyo' [k u . jj o], 'muy' [m 'u j]).
In all other positions, both as a first element of a rising diphthong ('abuela' [a . B w 'e . l a]) or as a second element of a falling diphthong ('acaudalados' [a . k a w . D a . l 'a . D o s]), 'u' is transcribed as the glide [w].

On the other hand, [i] can be transcribed either as a voiced palatal approximant [ʝ] when it occurs after 'h' before a vowel (e.g. 'hiedra' [jj 'e . D r a]) or like the glide [j] when it is the second element of a falling diphthong ('afeitar' [a . f e j . t 'a r], 'prohibido' [p r o j . B 'i . D o]), or in first position of a rising diphthong ('sociedad' [s o . T j e . D 'a D]). Otherwise, when stressed, it is realised as [i] ('pingüino' [p i N . g w 'i . n o]).

Most of these transcriptions are incorrect from a linguistic point of view, but they are functional to the recognizer they are designed for, which does not distinguish semi-vowels from full vowels, and unstressed vowels from stressed ones.

2.1.2.Synthesis

However, correct rendition of hiatuses and diphthongs is crucial for synthesis, in order to select appropriate correspondent units. A different, more complex treatment of these groups is therefore required, which involves: a) initial retrieval of pre-stressed roots containing hiatuses from a table look-up b) stress assignment to other vowel sequences c) transcription according to stress position.

Only primary stress is assigned by the following procedure, which searches in the string whether:

a) the word ends either by a simple vowel (i.e. preceded by a consonant e.g. 'moc'illa', 'ov'illo') or by a rising diphthong (i.e. by a vowel preceded by 'i','u' or 'y', e.g. 'l'impio', ''agua', 'desm'ayo'); then stress is assigned to the vowel preceding the last vowel or diphthong, if it is 'a,e,o' (es. 'Can'aria'). Also 'i','u' and 'y' can be stressed in that position, if they are preceded by 'qu', 'cu', 'gu', or by a consonant or if they are initial (e.g. 'Chiqu'illlo', 'engu'anta', 'b'urgo', 'argent'ina', ''urna').
b) the word ends by a vowel, preceded by a vowel different from 'i,u,y'; then the second-last is stressed (es. 'Paragu'ay', 'can'oa', 'Bilb'ao', `cefal`ea').
c) the words ends by 'n' or 's' preceded by a simple vowel; then stress falls on the second-last vowel if it is 'a,e,o' (e.g. ''orden', 'c'asas'); if the second-last is 'i', 'u' or 'y' and one of these vowels is either initial, or preceded by a consonant or preceded by 'qu', 'cu', 'gu', then even 'i,u,y' can be stressed (es. ''urnas', 'b'urgos', 'chiqu'illos').
d) the word ends by a consonant different from 'n,s' preceded by a single vowel; then that vowel is stressed (e.g. 'pap'el', 'muj'er').

Stressed vowels in the sequence are then transcribed as full vowels and unstressed ones either as semi-vowels when in second position of falling diphthongs ('afeitar' [a . f e i~ . t 'a r]), or as semi-consonants if in first position of a rising diphthong ('propia' [p r 'o . p j a]).

2.2. Consonant rules

Also, consonants undergo a different treatment for synthesis and recognition:

'b, d, g' are transcribed as voiced stops [b d g] if initials (e.g. 'bueno') or preceded by a homorganic nasal (e.g. 'hombre', 'conde', 'mingo'), or by [l] for the dental stop (e.g. 'toldo').
Otherwise, if they are internal, preceded by a consonant different from a nasal, they are realised as the corresponding voiced bilabial or velar fricative [β,ɣ] or dental approximant [ð] (e.g. 'amaba', 'arruga', 'crudo').
For synthesis, voiced stops are devoiced when they precede an unvoiced phone.

'p, t, k, c' are transcribed in the following way:
'p' is deleted before 's' (e.g. 'psicólogo'), otherwise it is realised with the corresponding bilabial stop [p] (e.g. 'papel').

't' is realised as the voiced dental approximant [ð] before 'b', 'm' and final (e.g. 'fútbol', 'cenit', 'istmo'), otherwise as 't' (e.g. 'técnica').
'c' is realised as the unvoiced interdental fricative [θ] before a front vowel (e.g. 'exception', 'ceso') or as a velar voiced fricative [ɣ] before 'd, n' (e.g. 'anécdotas', 'técnica').
For synthesis, [p t k] are converted in the correspondent voiced approximant allophones, before a voiced consonant (e.g. 'atmósfera' [a Dh m `o s f e r a]).

Nasals assimilate place of articulation of the following consonant and are transcribed with the correspondent allophones (e.g. 'amplio' [a m p l j o], 'chanfla' [TS `a M f l a], 'berrinche' [b e r: `i Gn TS e], ángulo [`a N g u l o]).

'r' is realised as a geminate [r:] when initial, before 'r', 'n', 'l', 's' (e.g. 'burrito', 'redondo', 'honra', 'alrededores'), otherwise it is transcribed as the alveolar flap [ɾ].

'z' is realised as the (inter)dental voiced approximant [ð] before a voiced consonant (e.g. 'juzgar', 'hallazgo', 'gozne'), otherwise as the unvoiced (inter)dental fricative [θ] (e.g. 'azteca', 'razón', 'zapata').

'v' is transcribed as [b] when inital or preceded by [n] (e.g. 'verdad', 'conviene'), otherwise as the bilabial voiced approximant (e.g. 'ovillo').

'x' is transcribed as [ks] when initial (in Catalan words, present in SpeechDat, e.g. 'Xavier') or as [ɣs] when followed by a vowel or 'h' (e.g. 'examen', 'exhortación'); for synthesis 'x' in these context is realised as [k s].

'y' is always transcribed as the palatal voiced approximant [ʝ] in every condition for recognition, whereas two allophones are distinguished for synthesis: if initial, or internal after [l], [n], 'y' is realised as a palatal voiced stop [ɟ] (e.g. 'yelmo', 'inyectar'), otherwise as the palatal approximant [ʝ] (e.g. 'cayado').

3. Transcription results

The recognition rules have been tested over the SpeechDat Spanish corpus, used to train the recogniser, and on the synthesis corpus, including both isolated words and sentences; however, single words composing sentences have been separately treated and then reassembled, because restructuring rules at word boundary to account for inter-word coarticulation were not yet implemented. Transcription results are reported in Table 2.

	SpeechDat	*Synthesis*
Transcription errors	1.5%	0.6%
Total words	4444	3578

Table 2. Transcription errors

Lack of matching between the transcriptions produced by the rules and those provided by the SpeechDat corpus were due to the following reasons: a) the 23 single alphabetic letters were not transcribed by the rules b) stress on Catalan city names was not correctly placed in the SpeechDat corpus, and the rules provided the correct version (e.g. 'Masnou' SD: [m'asnow] vs. rules: [masn'ow]) b) vowel sequences were always stressed on the first element and [i] and [u] in diphthongs and hiatuses were always reported as glides [j] and [w] respectively. These two SpeechDat conventions obscure the difference between the diphthongs and hiatuses (see above par. 2.1). The rules assign stress on diphthongs and hiatuses and consequently produce correct transcription of the closed vowels ('rehúye' [rr e `u jj e], 'reír' [rr e `i r], 'oír' [o `i r]) c) city names that included a stressed initial capital letter in the SpeechDat corpus (e. g. Ávila) could not be read in input by the transcriber, so their transcription does not match the reference one d) secondary stress is not assigned by the rules, so all transcriptions of SpeechDat which report it, do not match with those provided by the rules, which consider only main stress (e.g. ''unicam`ente' vs. ''unicamente') e) some phenomena occurring at morpheme boundary could not be transcribed (e.g.

absence of voicing of [s] in final position of a prefix before a voiced consonant; e.g. 'disyuntivas' [d i s + jj u n . t i . B a s]); in fact, no morphological parser is included.

Errors over the synthesis corpus chiefly appeared on hiatuses (e.g. 'circuito' [Th i r k w `i t o] vs. [Th i r k `u i t o]) and on words of foreign origin ('boxeador' [Bh o Gh s e a Dh `o r] vs. [Bh o k s e a Dh `o r]).

4. Discussion and further developments

Rules matched effectively the reference phonetic transcriptions in the SpeechDat corpus and showed to constitute a useful tool to generate new standard pronunciations, even when extended to new corpora, like the one adopted for the synthesis application.

Further developments of the present system can consist in adding rules for pronunciation variants of standard Castilian Spanish, like South-American varieties (especially Mexican, Argentinian and Paraguayan), to be implemented within the recogniser as optional pronunciations to the standard forms, according to a procedure used to generate dialectal variants of standard Italian [19].

For synthesis, on the other hand, possible extensions would include restructuring rules at word boundary, which enable to automatically transcribe words within sentences. These rules should account for changes like place assimilation of final nasals and laterals to the following word-initial consonant (e.g. 'con ganas' [k o N g `a n a s]; 'el cho`fer' [e Gl TS `o f e r]); deletion of a final unstressed vowel before identical unstressed ones (e.g. 'la atmo`sfera' [l a Dh m `o s f e r a]) and realisation of closed vowels as glides before a word-initial vowel (e.g. 'ni un' [nj`un]). Also syllabification rules can be added, in order to allow a more adequate treatment of vowel sequences.

Acknowledgements

Our thanks go to CSELT Labs synthesis and recognition research groups and to Vahid Junuzovic, for their collaboration and support.

References

[1] Andersen O., Kuhn R. et al. (1996) "Comparison of Two Tree-Structured Approaches for Grapheme-to-Phoneme Conversion", ICSLP '96, V. 3, pp. 1700-1703, Oct. 1996.

[2] Fukada T. and Sagisaka Y. (1997) "Automatic Generation of a Pronunciation Dictionary Based on a Pronunciation Network", Eurospeech '97, V. 5, pp. 2471-2474, Sept. 1997.

[3] Torre D., Villarubia L. et al. (1997) "Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers", ICASSP-97, V. II, pp. 1463-1466, April 1997.

[4] Ravishankar M. and Eskenazi M. (1997) "Automatic Generation of Context-Independent Pronunciations", Eurospeech '97, V. 5, pp. 2467-2470, Sept. 1997.

[5] Cremelie N. and Martens J.-P. (1997) "Automatic Rule-Based Generation of Word Pronunciation Networks", Eurospeech '97, V. 5, pp. 2459-2462, Sept. 1997.

[6] Andersen O. and Dalsgaard P. (1995) "Multi-lingual testing of a self-learning approach to phonemic transcription of orthography", Eurospeech '95, pp. 1117-1120, Sept. 1995.

[7] Deshmukh N., Ngan J., Hamaker J. and J. Picone (1997) "An Advanced System to Generate Pronunciations of Proper Nouns", Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-97), V. II, pp. 1467-1470, April 1997.

[8] Adamson M. and Damper R. (1996) "A Recurrent Network that Learns to Pronounce English Text", Int. Conf. on Spoken Language Processing 1996 (ICSLP '96), V. 3, pp. 1704-1707, Oct. 1996.

[9] Bonaventura P., Micca G. and Gallocchio F. (1998) "Speech recognition methods for non-native pronunciation variations", ESCA Workshop, Rolduc, 4-6 May 1998, pp. 17-23.

[10] Bonaventura P. and Di Carlo, A. (1985) "Regole di trascrizione da grafema a fonema per applicazioni alla sintesi dell'italiano standard", Rivista Italiana di Acustica, vol. 3, pp. 85-105.

[11] Cavalcante Albano E. and Antonio Moreira A. (1996) "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in

Portuguese", Int. Conf. on Spoken Language Processing 1996 (ICSLP '96), V. 3, pp. 1708-17011, Oct. 1996.

[12] Ríos A. (1993) "La información lingüística en la transcripción fonética automática del español", *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13, pp. 381-387.

[13] Salza P. (1990) "Phonetic transcription rules for text-to-speech synthesis of Italian", Phonetica, vol. 47 pp.66-83.

[14] Canepari L. (1979) "Introduzione alla fonetica ", Einaudi, Torino.

[15] Quilis A. (1993) "Tratado de Fonología y Fonética españolas", Madrid, Gredos.

[16] Fissore L., Ravera F., Laface, P. (1995) "Acoustic-phonetic modelling for flexible vocabulary speech recognition", EuroSpeech 95, Madrid, pp. 799-802.

[17] Bonaventura P., Gallocchio F. and Micca G. (1997) "Improvement of a vocabulary- and speaker-independent speech recogniser for English and Spanish", CSELT Working Papers n. DTR 97.0788.

[18] Angelini B. et al. (1997) "Automatic diphone extraction for an Italian text-to-speech synthesis system", Eurospeech '97, V. II, pp. 581-584, Sept. 1997.

[19] Bonaventura P. and Leprieur H. (1994) "Grapheme-to-phoneme transcription rules for synthesis of regional Italian", Rivista Italiana di Acustica, vol. 3.