FREQUENCY ANALYSIS OF PHONETIC UNITS FOR CONCATENATIVE SYNTHESIS IN CATALAN

Ignasi Esquerra, Albert Febrer and Climent Nadeu

Universitat Politècnica de Catalunya
Campus Nord UPC, 08034 Barcelona, Spain
ignasi@gps.tsc.upc.es

ABSTRACT

Knowledge of phonetic unit frequency is very necessary for developing databases in both concatenative synthesis and continuous speech recognition. In the present work, a large corpus of text was processed and phonetically transcribed to obtain allophone and diphone frequencies for the Catalan language. The corpus was acquired from newspaper articles, in which there were a lot of foreign words that represented a problem in the normalisation process. After automatic transcription, units were counted to get their relative frequency and results were compared to other analysis. Finally, diphones found in the corpus were compared to units of a synthesis database to validate both the normalisation and transcription modules and the synthesis unit database.

1. A PHONETIC FREQUENCY ANALYSIS FOR THE CATALAN LANGUAGE

Many current research areas in spoken language technology make extensive use of corpus data to get knowledge on the characteristics of actual speech or language. In order to train models for automatic speech recognition (ASR), large databases are collected whose words, sentences and other items have been carefully selected to cover all the relevant phonetic units for a given language. Therefore, it is important to know how many units are relevant and how often they occur to produce those phonetically balanced databases

In the field of concatenative text-to-speech (TTS) synthesis, a database of pre-recorded units is required to create the acoustic signal. Databases are usually defined a priori by means of phonological rules. However, very often, this process is independent to the actual transcription tool of the system. When transcription fails, either by a typing error or an extraneous word, a non-available unit can be requested. Large corpora transcription can help in solving this kind of problems by studying what units are delivered and their relative occurrence compared to reference analysis of phonetic frequencies.

The main objective of this work is to get the frequencies of phonetic units in Catalan, to be applied in both speech synthesis and speech recognition system development. For this purpose, a corpus of newspaper articles was processed and phonetically transcribed to get allophone and diphone frequency statistics. As a by-product, this analysis allowed to test the performance of the transcription and normalisation modules.

Although this approach has been in use for many years in the development of TTS and ASR systems for other languages, the need of frequency statistics of phonetics units in Catalan is required to be able to develop up-to-date applications for this language. In our department there are several on-going projects in ASR and TTS using Catalan as a focus language.

Catalan is the native language of large area that comprises the eastern coast of Spain (Catalonia, Valencia and Balearic Islands), Andorra, Rousillion (France) and l'Alguer (Italy). According to the last sociolinguistic surveys, Catalan is spoken by eleven million people, though they are all bilingual because it is not the official language in none of these countries, with the exception of Andorra.

In the following, the analysis methodology and the results are presented. In the first section, the corpus normalisation process is described, paying special attention to some of the problems in dealing with foreign words. Afterwards, frequency results are given for allophone and diphone units and they are compared to two previous works on phonetic statistics in Catalan. Finally, the list of diphones in our TTS synthesis database is compared to the units found in the analysis in order to extract some conclusions about the TTS modules. A brief discussion on the results achieved in this work is presented to conclude the paper.

2. CORPUS PREPARATION

The textual corpus used in this work is a small subset of files from the Catalan part of the PAROLE project that is being carried out at the *Institut d'Estudis Catalans* [1]. This corpus of over 20 million different words was acquired from recent books, newspapers and magazines written in Catalan. Among all available texts, opinion and interview newspaper articles were selected, because it is thought that this kind of text has the greatest lexical variety and thus they are probably the phonetically richest.

Many words were found in the corpus that could not be directly read by the transcription module. Some alphanumeric strings require to be converted into an orthographical form, whereas other symbols need to be transformed somehow. The process of normalising the corpus consisted mainly in getting rid of some header lines, expanding numerical expressions and abbreviations and coding special characters.

Several problems were faced during the corpus normalisation process. Acronyms and abbreviations not present in our TTS normalisation dictionary derived in strange words that were wrongly transcribed and, as a consequence, they resulted in rare phonetic diphones.

A more problematic source of problems was the presence of many foreign words. Because texts were extracted from a newspaper, the number of Spanish names (and from other countries as well) was extremely high in some files. For instance, former Spanish Prime Minister, Felipe González, whose name is almost always pronounced with the Spanish phoneme /T/, appears 294 times in the corpus. Moreover, some family names appear written in a mixed Catalan-Spanish orthography that makes things more difficult (e.g. *"Pedro Balañà"*, where the *"ñ"* letter is exclusive of Spanish and the *"à"* vowel can only have a grave accent in Catalan). With few simple regular-expression rules, it was possible to detect 720 different Spanish words, which were tagged and filtered out to not introduce distortion in the statistics computation. Without those words and some other foreign words that were casually detected, and after getting rid of header lines, the normalised corpus was reduced less than 1% of its original size [Table 1].

	Original	Normalised
Texts	2443	2443
Paragraphs	27349	13235
Lexicon	73074	70050
Words	1258189	1216182

Table 1. Corpus contents.

3. FREQUENCY ANALYSIS OF PHONETIC UNITS

In the Catalan language, there are 36 different sounds, including some common allophonic variations. The central dialect of Catalan, the one that was taken as a reference for transcription, has the peculiarity of having two vocalic systems: one for stressed syllables ([a], [e], [E], [i], [o], [O], [u]) and one for unstressed syllables with only three phones ([i], [u], [@]). In order to simplify the results, no distinction was made in relation to vowel stress or other prosodic features.

Grapheme-to-phoneme rules for the Catalan language are not as easy as for Spanish. In particular, an important source of errors in automatic transcription is the case of *"e"* and *"o"* vowels in stressed syllables with no orthographic accent, which can be pronounced either as open-mid ([E], [O]) or closed-mid ([e], [o]) without any known general rule. For example, *nena* /n'En@/ (little girl) vs. *neva* /n'eB@/ (it snows), or *rosa* /rr'Oz@/ (rose) vs. *rossa* /rr'os@/ (blonde). Another transcription ambiguity is consonant "r" in a word final position, that is usually mute (at least for the dialect taken as a reference for our system), but that is pronounced in many common monosyllabic words. For instance, *cor* /k'Or/ (heart) but *por* /p'o/ (fear), or *mar* /m'ar/ (sea) vs. *mà* /m'a/ (hand). In all those cases, an exception dictionary is looked up to disambiguate.

Two types of phonetic units were considered in this frequency analysis: allophones and diphones. Although diphones are defined in the synthesis sense, they represent as well context units for recognition. Left and right contexts are the same units in terms of frequency occurrence. For example, the phonetic unit [a] followed by [n] (i.e. [a]+[n]) appears the same number of times than [n] preceeded by [a] (i.e. [a]-[n]).

3.1. Allophone Frequencies

Relative frequencies of allophones show, not surprisingly, that the sound [@] is largely the most frequent allophone in Catalan [Table 2]. This is because of the large number of unstressed syllables containing the vowels *"a"* or *"e"*, which in the Catalan central dialect are pronounced invariably with a schwa. The following most common vocalic allophones are [i] and [u] because they constitute, together with the [@], the unstressed vocalic system for the central dialect.

In relation to transcription of vowels *"e"* and *"o"*, the more common close-mid phones are taken by default. This decision implies a higher proportion of the [e] and [o] allophones, in opposition to the open-mid vowels [E] and [O] that appear to be less frequent in the analysis that they really are.

Unit	# of units	rel. freq.	acc. freq.
@	1037062	18.94	18.94
i	413521	7.55	26.50
s	352822	6.44	32.94
n	336094	6.14	39.08
l	311592	5.69	44.77
t	283283	5.17	49.95
u	275223	5.03	54.98
a	252160	4.61	59.58
k	244778	4.47	64.05
r	202063	3.69	67.74
m	196537	3.59	71.33
z	172096	3.14	74.48
p	166204	3.04	77.51
e	162151	2.96	80.48
D	157826	2.88	83.36
rr	123361	2.25	85.61
o	116172	2.12	87.73
B	112905	2.06	89.80
d	79093	1.44	91.24
E	60705	1.11	92.35
O	60434	1.10	93.46
f	54159	0.99	94.44
w	47058	0.86	95.30
G	38772	0.71	96.01
b	36098	0.66	96.67
Z	33641	0.61	97.29
L	27797	0.51	97.79
N	27624	0.50	98.30
g	20177	0.37	98.67
S	14001	0.26	98.92
j	13946	0.25	99.18
J	13885	0.25	99.43
dz	13105	0.24	99.67
ts	10844	0.20	99.87
dZ	5629	0.10	99.97
tS	1560	0.03	100.00
total	*5474378*	*100.00*	*100.00*

Table 2. Allophone frequencies (relative and accumulative values in %)

It is also interesting to note that only six phones are required to achieve half of the total number of allophones in the corpus, whereas to reach a 90% is possible with the first 18 allophones, i.e. half of the phonetic units considered in the analysis.

These results can be compared to two previous works. The first one was a phonology analysis of Catalan units carried out in 1979 by Joaquim Rafel [2]. Because in his study only phonemes were taken into consideration, several allophones had to be grouped into single units (i.e. phonemes) to make comparisons possible. The other work, carried out in our department, was used in the design of a phonetic corpus for a speech recognition database in Catalan (VOCATEL) [3]. The methodology was very similar to the present work, but the corpus used then was ten times smaller and the normalisation process was simpler.

A relative error distance was computed between the three corpora results, referred as UPC, RAF and VOC respectively [Table 3]. Frequencies in UPC column are slightly different to the previous section results due to the rearrangement of allophones into phonemes to make statistics comparable.

In both cases, the major differences occur for the phoneme pairs [e]/[E], [o]/[O]. Again, the frequency of the open-mid vowels is below the actual frequency in spoken language, even though distances are compensated when their frequencies are added giving a distance around 2% for both vowels. The application of an exception dictionary with words presenting this ambiguity clearly improves the transcription correctness with respect to results obtained for the VOCATEL corpus.

Other important differences are caused by a different voicing assimilation rule between words applied in transcription. Such is the case of the phonemes [s]/[z]. Finally, the fact that the RAF corpus was very small, frequencies for the phonemes at the bottom are not very reliable since they appeared less than 10 times and this may explain those differences.

Unit	UPC	VOC	RAF	UPC-VOC	UPC-RAF
@	18.91	18.76	20.10	0.78	-6.29
i+j	7.79	7.77	6.16	0.36	20.96
s	6.63	6.86	8.57	-3.40	-29.29
n	6.13	6.13	6.38	0.01	-4.20
u+w	5.88	5.83	5.63	0.75	4.18
l	5.68	5.79	6.00	-2.00	-5.66
t	5.36	5.70	5.24	-6.23	2.33
a	4.60	4.61	4.85	-0.32	-5.39
k	4.46	4.36	4.41	2.30	1.10
d+D	4.32	4.06	4.48	5.89	-3.77
r	3.68	3.55	3.77	3.56	-2.24
m	3.58	3.73	3.83	-4.13	-6.75
z	3.14	3.03	0.76	3.34	75.93
p	3.03	2.99	2.76	1.33	9.04
e	2.96	3.67	2.34	-24.04	20.71
b+B	2.72	2.63	2.96	3.23	-9.04
rr	2.25	2.54	2.32	-12.88	-3.36
o	2.12	2.97	1.77	-40.00	16.64
E	1.11	0.35	1.63	67.93	-47.11
O	1.10	0.31	1.43	71.63	-29.97
g+G	1.07	1.00	0.95	7.30	11.48
f	0.99	1.04	1.18	-5.32	-19.20
Z	0.61	0.65	0.44	-5.48	28.04
L	0.51	0.57	0.85	-12.12	-68.38
N	0.50	0.32	0.36	36.75	27.94
S	0.26	0.32	0.35	-25.18	-38.33
J	0.25	0.25	0.28	1.25	-12.36
dz	0.24	0.07	0.02	72.13	91.79
dZ	0.10	0.10	0.08	6.66	23.54
tS	0.03	0.06	0.10	-112.23	-244.86

Table 3. Frequency and error distances between the three phonetic analysis on different corpora (values in %).

3.2. Diphone Frequencies

Diphone frequencies were calculated in the same way and there were found 1046 different units. However, the analysis of diphone frequencies was somewhat more difficult because the set of possible units is undetermined. Some C+C combinations are impossible to be realised because of articulatory constraints, but a lot of them depend on the criteria followed to deal with consonant assimilation between words.

If units are ordered in a descending frequency order and plotted in a semilog scale, it can be observed that, at least for the first hundred values, their frequency can be quite well interpolated with a line [Figure 1]. It was expected that, being the corpus quite large, the less frequent diphones should appear at least several times. But, for the higher rank diphones the curve decreases smoothly instead of ending with an abrupt step to zero for the last diphone. The explanation for this behaviour is that all of those diphones result from a transcription error, mainly because of foreign words.

Frequency results for diphones, as well as the number of units, are presented grouped by phonetic classes since it is impossible to show them all [Table 4].

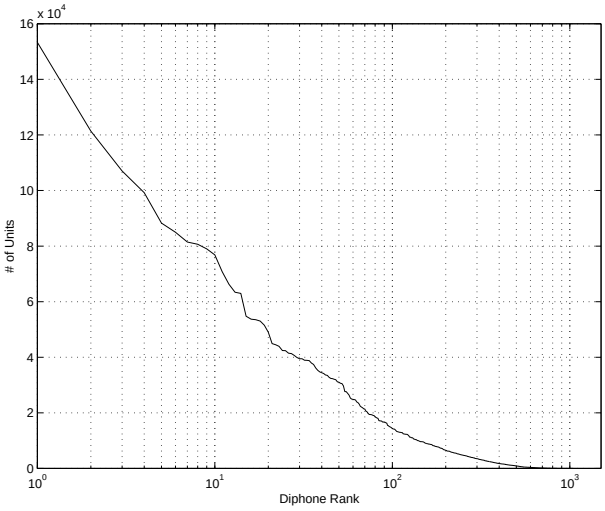

Figure 1. Diphone absolute frequency by rank

	Vowel	Plosive	Fricative	Affricate	Approx.	Nasal	Liquid	Silence	*total*
Vowel	4.76 (64)	6.71 (47)	7.15 (40)	0.51 (32)	5.03 (40)	8.77 (31)	8.33 (32)	1.10 (8)	*42.39* (294)
Plosive	11.00 (48)	0.91 (34)	0.45 (21)	0.00 (6)	0.19 (14)	0.17 (11)	1.80 (20)	0.24 (5)	*14.79* (159)
Fricative	6.64 (40)	2.14 (20)	0.37 (18)	0.00 (5)	0.80 (18)	0.34 (9)	0.29 (14)	0.56 (4)	*11.18* (128)
Affricate	0.35 (32)	0.05 (9)	0.01 (11)	(0)	0.05 (10)	0.01 (7)	0.00 (9)	0.05 (2)	*0.55* (80)
Approx.	5.29 (40)	0.17 (24)	0.15 (17)	0.00 (7)	0.11 (12)	0.06 (11)	0.74 (14)	0.05 (3)	*6.60* (128)
Nasal	5.42 (32)	2.58 (24)	1.56 (19)	0.01 (8)	0.00 (5)	0.17 (10)	0.24 (11)	0.22 (4)	*10.23* (113)
Liquid	7.96 (32)	1.39 (20)	1.21 (17)	0.00 (7)	0.40 (14)	0.52 (11)	0.18 (11)	0.14 (4)	*11.83* (116)
Silence	0.96 (8)	0.80 (6)	0.24 (5)	0.00 (1)	0.00 (2)	0.16 (3)	0.21 (3)	(0)	*2.39* (28)
total	*42.41* (296)	*14.79* (184)	*11.18* (148)	*0.53* (66)	*6.60* (115)	*10.23* (93)	*11.83* (114)	*2.39* (30)	*100.00* (1046)

Table 4. Diphone frequencies by phonetic class (row+column) (frequencies in %; in parenthesis number of different units)

4. SYNTHESIS UNIT DATABASE

One of the objectives of this frequency analysis was the validation of a phonetic unit corpus for speech synthesis in Catalan. In a previous work, a table of possible diphones for our bilingual TTS system was defined taking into account the phonetic rules of Catalan [4]. Currently, the system's unit database consists of 800 diphones, plus some polyphones to cope with difficult coarticulation transitions.

Comparing that table to the list of diphones in the corpus, the main discrepancies were detected for the units made of two consonants. Indeed, although their relative frequency is small compared to that for C+V and V+C units (17.2% in front of 36.7% and 36.5%, respectively), the C+C diphones are the most likely to happen when the transcription module runs into an error because of a difficult consonant cluster.

Looking carefully to origin of the non-common units, it was realised that either a normalisation or a transcription error was the cause: foreign words (e.g. *"Darmstadt"*), acronyms and abbreviations (e.g. *"vda."*, *"ADN"*), or unavoidable typing errors. Furthermore, half of those units are among the 20% less frequent units in the corpus, which explains in part the behaviour of curve in figure 1. Apart from enlarging the exception dictionaries, a table of unit substitutions was created to deal with those units that are not present in the synthesis database because do not correspond to valid phonetic transitions.

Likewise, some rare diphone units are synthesised in our system in a special way. Instead of storing them as usual, they are created from two half-phones belonging to the phone+silence and silence+phone diphones, making the silence segment very small. The knowledge of the relative frequencies of such diphones helped us to decide whether it is worth to treat them as normal units or they can be saved using half-phones.

5. CONCLUSIONS

The analysis of diphone units obtained from the transcription of a large textual corpus like this one has permitted to test the performance of both the transcription and the normalisation modules. The most rare units normally appear due to words that could not be treated in a satisfactory way by the text processing front-end to the TTS system. In particular, a large number of Spanish words (names, but also long quotations) were found in the corpus.

As it was seen in the comparative analysis with other similar works, allophone frequencies depend to some extend on the decisions made about coarticulation rules. This fact was particularly clear comparing the frequencies of voiced/unvoiced pairs of phonemes. Diphones found in the analysis permitted to identify several units that, even though they are not phonetically possible in Catalan, may appear as a result of text transcription.

As a summary, the analysis of this corpus has led to new statistics of frequency distribution of phonetic units in Catalan, which have been applied to the validation of a unit database for concatenative synthesis, as well as for testing transcription and normalisation. Results are also being used in the definition of a phonetic corpus for a speech recognition database. Further work has to be done in the corpus normalisation process because a lot of interesting information can be still extracted from it, improving the robustness of the UPC TTS system.

ACKNOWLEDGEMENTS

Authors want to thank the *Institut d'Estudis Catalans* for providing the text corpus used for this analysis and the Phonetics Laboratory researchers from the *Universitat Autònoma de Barcelona* for their help in the transcription and unit selection process for the UPC TTS system. This work has been supported by CIRIT, Generalitat de Catalunya, through the Centre de Referència en Enginyeria Lingüística (CREL).

REFERENCES

[1] J.Soler i Bou, *"Written Linguistic Resources in Catalan: the DCC project"*, Workshop on Language Resources for European Minority Languages, in LREC'98, Granada, May 1998

[2] J.Rafel i Fontanals, *"Dades sobre la freqüència de les unitats fonològiques en català"*, Estudis Universitaris Catalans XXIII, vol.2, p 473-496, 1979

[3] I.Esquerra, C.Nadeu, L.Villarrubia and P.León, *"Design of a Phonetic Corpus for Speech Recognition in Catalan"*, Workshop on Language Resources for European Minority Languages, in LREC'98, Granada, May 1998

[4] A.Bonafonte, I.Esquerra, A.Febrer and F.Vallverdú, *"The UPC Text-to-Speech System for Spanish and Catalan"*, in these proceedings