Corpus and Voices for Catalan Speech Synthesis

Antonio Bonafonte, Jordi Adell, Ignasi Esquerra, Silvia Gallego, Asunción Moreno, Javier Pérez

TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
antonio.bonafonte@upc.edu

Abstract

In this paper we describe the design and production of Catalan database for building synthetic voices. Two speakers, with 10 hours per speaker, have recorded 10 hours of speech. The speaker selection and the corpus design aim to provide resources for high quality synthesis. The resources have been used to build voices for the Festival TTS. Both the original recordings and the Festival databases are freely available for research and for commertial use.

1. Introduction

During last decade, speech synthesis has followed the same evolution than other speech technologies, as speech and speaker recognition or spoken translation: moving from rule base systems to data-based systems. In the data-based approach the designed or the system provides knowledge on the problem defining appropriate models and features; but the particular features or parameter values are derived from training data. In speech synthesis this strategy has been applied to all the involved tasks. Obviously to task which are common to other speech or language technologies, as POS tagging, grapheme-to-phoneme mapping and normalization of non-standard words. But also to TTS-specific tasks as prosody generation and waveform generation.

The advantage of the corpus based approach is that once models and algorithms are ready, they can be ported to applied to create new voices in the same or different language with limited effort. Furthermore, the debugging and improving of the models is relatively easy, without critically depending on the expert that developed the models. On the other hand, if the corpus does not exist, the development of such corpus can be very expensive. This makes difficult the portability of the corpus-based approach to languages with limited resources. The lack of available corpus is an important handicap for doing research and using speech technologies in languages with small or medium markets, as Catalan.

We believe that funding the production of public resources and making them accessible with as few restrictions as possible is the best way to contribute to the development of the technology in the language. The TALP Research Center, at UPC, has already released public resources for Natural Language Processing as lexica, taggers, morphological analyzers (Carreras et al., 2004). Recently, we have produced and released data for training acoustic models for speech recognition in different acoustic environments, as fixed and mobile telephony, car, office, etc. (Moreno et al., 2006). In the LC-STAR project, we also have produced public lexica, including POS tags and phonetic transcription for their use in speech recognition and speech synthesis. But currently there was not data for developing the prosody generation and waveform generation modules of speech synthesis systems.

The Education Ministry of the Catalan Government is producing *LinKat*, a Linux distribution with Catalan support aimed to primary and secondary schools in Catalonia. Speech synthesis is a key component in many accessibility tools, as *gnopernicus* or *orca* but Catalan voices were not available. In this paper we describe *FestCat* (FES, 2007), a specific project to produce Catalan voices for *Festival*(**?**), a text-to-speech system included in most of the Linux distributions. In order to promote the research and the use of Catalan speech synthesis, we have created resources which exceed the basic Festival voices. The corpus produced allow to build state-of-the-art synthetic voices and are open for either research or commercial application.

Section 2. describes the design principles and the production of the speech synthesis databases produced in the project. It includes the preparation of the textual corpus, speaker selection, recording settings and labeling of the recordings. Section 3. explains the voices that have been developed for the Festival system. This includes other resources which can be used for Catalan TTS, as word normalization, phonetic transcription, etc. Finally, section 4. summarize the paper and draws some conclusions.

2. Corpus for building Catalan synthetic voices

The corpus produced in the *festcat* project has been designed based on two principles: it has to be useful for producing synthetic voices for Festival, and it has to be useful for producing state-of-the-art voices with prevalent technologies. Furthermore, the corpus should be useful for doing research either to our laboratory or other interested researchers.

Festival allows to use several technologies. *Diphone synthesis*, which was popular in the early and middle 90s, generates the speech waveform by concatenation of basic speech segments, *diphones*, which are modified to have the desired intonation. Production of databases for diphone synthesis is relatively cheap but the quality of the voice is usually not very good. It is still used by some companies for mobiles and other embedded devices, and in fact, it can produce quite acceptable results using sophisticated signal processing algorithms.

Newer and better technologies are corpus based synthesis, as *clunit* voices and *Multisyn* voices. Corpus-based synthesis creates the waveform also by concatenation of basic segments, but the segments are selected from a corpus. If the corpus is large, there are many candidates to create the waveform. As a result, the segments can be selected to reduce discontinuities (an in particular, favoring longer segments) and also to reduce the need of manipulation to get the correct prosody. Public voices are available for Festival based on the ARCTIC databases (Kominek and Black, 2003), which are approximately 1 hour of read speech.

Corpus based synthesis is also used in the best commercial systems. It seems that most of these systems use databases with several hours (3-12).

Recently, HMM-based synthesis has become popular and has also been integrated in Festival. In HMM synthesis, the database is used to estimate mathematical models (HMM) which are used for generating speech parameters which are finally transformed into speech. The quality is still not as good as corpus based synthesis, but it has several advantages. First of all, it is a very flexible technique where speaker/style/language adaptation can be imposed. Furthermore, the memory requirements are low, similar to diphone synthesis and much more lower than corpus based. Finally, as this is a very active research area, the quality is improving year after year. The size of the database is not as critical as in the corpus-based synthesis. Anyway, as bigger the database is, more accurate are the HMM models and better quality.

Based on the previous discussion, we decided to produce corpora for building two voices (one male voice and one female voice). The size of each corpus will be around 10 hours, i.e., approximately 90,000 words. Such database allows to create corpus-based synthesis with different footprint and qualities, including state-of-the-art voices. It also allows to build HMM voices. The specification of the databases are based on the ones defined defined in the EU TC-STAR project (Bonafonte et al., 2006) for creating high quality voices, but several modifications have been done is the selection of the domain of the voice, in the speaker selection and to cope with bilingualism (Catalan/Spanish) and foreign words.

2.1. Corpus design

The design of the corpus follows the principles established in TC-STAR (Bonafonte et al., 2006).

We have divided the corpus in two parts: the first part contains 80% of the corpus (see table 1). The goal is to achieve good phonetic and prosodic coverage. Some domains have been defined to favor large variability and also taking into account the first interest of the project (education and accessibility) or other projects from our lab (parliament). Dialog and monologues (from literary plays) have been included to have questions, dialog verb tenses and colloquial expressions. Furthermore, a specific corpus of questions has been added to improve the prosody variation of the corpus.

For each subdomain, we have collected several millions of words. The raw text has been divided in short paragraphs which have been labeled with context dependent phonemes. Furthermore, each unit has been labeled with information about the stress and the position in the sentence(beginning, middle, ending), as these features have a strong correlation with prosody. Based on the analysis of the whole text we have fixed that, for each domain, we want to include all the diphones, including the stress feature, that in the analysis appear more than 10 times per million. And all the diphones, including stress and position features that in the analysis appear more than 100 times per million. For each subdomain, we have selected the paragraphs from all the data available using our public tool *CorpusCrt* (Sesma and Moreno, 2000). This tool uses a greedy algorithm to select sentences, according to some specification criteria, from a large collection of sentences.

In most of the domains, the corpora have been selected using a greedy algorithm applied over a large collection of data.

This is applied to all the corpus, except the one which we have named *phonetically rich sentences*. These two-hour corpus is designed to achieve more variability in the triphones and to achieve the design criterion defined in (Bonafonte et al., 2006): the triphones included in all the corpus have to include 95% of the intraword triphones that appear in our lc-star lexicon with 50,000 common words. For the missing triphones, we have designed sentences using the words in the lexicon. The *questions* has also been designed by hand using some sentences found in the corpus.

The second part contains 20% of the corpus (table 2), and includes some domains which are important for some applications. Some of the domains are very generic, as numbers, cities and villages, and spellings, while others are more specific, as sentences which represent typical web sites (aimed to Internet accessibility) or the IVR commands (addressed to call centers and other telephone services). The expressivity part includes a few non-verbal expressions (laugh, breath, etc.) and also filler pauses and repetitions, to support our research on that topic. 10 calibration sentences have been repeated several times so that the speaker adjust his voice to the normal tone and speed. Furthermore, we have keep these sentences as multiple references to evaluate some components of the TTS system, for instance, intonation modeling. The sentences here have been designed manually trying to get representative lexicon and structure.

For most of the subdomains, the corpora have been produced using a greedy algorithm applied over a large collection of data. The procedure to ensure phonetic and prosodic coverage will be described in the paper. However, in some domains, the design has been done by hand. For instance, the phonetically-rich sentences has been designed by hand to include specific words. The words have been selected comparing the statistics of the other corpora with the LC-STAR Catalan dictionary. In order to achieve the highest coverage of triphones, rare words of the LC-STAR dictionary have been selected. In general, the percentage of questions in the corpus is small. One specific corpus has been designed to include questions, including Wh- questions, very short questions, etc.

All the text has been reviewed before the recordings. The goals of this revision are:

- Correct typos and misspellings.

Domain	Approximate number of words
Novels	18,000 words
Dialogs and monologues	4,750 words
News	10,000 words
Software user's guides	6,500 words
Transcriptions of the parliament	4,500 words
Teaching books	9,000 words
Phonetically-rich sentences	13,200 words
Additional questions	2,300 words
Spanish corpus	3,000 words

Table 1: Corpus domains for phonetic and prosodic variability

Domain	Approximate number of words
Numbers, dates	2200
Catalan cities and villages, world cities	1900
Web sites	4300
Screen reading commands	1000 words
Spellings	300 letters
IVR commands	4700 words
Company names	350 words
Web and email addresses	50 addresses
Expressivity	1800 words
Calibration sentences	1300 words

Table 2: Corpus domains for specific domains

- Simplify the structure of complex prompts. In some cases, the prompts we are recording are long (paragraphs with several sentences). For instance, the mean number of words of the *news* corpus is 25. Some of these paragraphs have a complex structure, with parenthesis, or changes in the word order to get a literary effect. In order to simplify the reading of the corpus we have reformulated some sentences so that in most cases the speaker do not need to plan how to read. We think this helped a lot in delivering fluent speech.

- Detect and solve ambiguities produced by *non-standard words*. All the acronyms, numbers, etc. have been labeled in the prompts. The transliteration is not showed to the speaker (it would have made it harder to read), but the recording supervisors instruct him how to read in case of doubt. This data can also be used in the future for developing normalization rules for the TTS.

- Detect and label words with non standard pronunciations (eg. foreign words, as Spanish and English words, or technology words, as name of software tools).

2.2. Speaker selection

One of the unsolved problems in concatenative speech synthesis is the selection of the speaker. Using the same technology, and similar speech data, the synthetic voices can have very different quality depending on the speaker. In order to minimize the risk of getting poor quality voices, it is important to spend resources in the speaker selection.

The target voices are two speakers, one male and one female. Then, for the speaker selection, 5 males + 5 females professional speakers have been selected. These 10 speakers have recorded 1h of speech each. For the speaker selection we have used the novel corpus. However, not all the speakers read the same corpus: we have designed 5 different novel corpus in order to increase the variability. We think that this data can be very useful to estimate better prosody models. In particular, phrasing algorithms can be trained using speaker independent data. All the corpus share 20 common sentences for evaluation purposes.
The selection of the speakers (1 male + 1 female) was done by the authors taking into account several aspects:

- Phonetic aspects of the original voice as proper Catalan pronunciation, clear articulation, etc.

- Pleasantness of the original voice.

- Quality of a simple synthetic voice. For each speaker we have generated automatically synthetic voices for our own TTS system. We believe that the quality of this 1-hour voice is a good indication of the quality of 10-hour voice. If the voice presents some problem in the segmentation, pitch detection, inconsistencies, etc. this already affect the quality of the 1-hour voice.

- Quality after TD-PSOLA modification of the original recordings.

- Capability of the speakers for doing fast recording (no many reading errors) and to have same voice quality after long recording sessions. Most of the speakers worked in radio broadcasting and where very suited

for the task. But still, there where some differences between speakers.

2.3. Recordings

The recordings have followed the specifications defined in (Bonafonte et al., 2006). They have been done in a recording studio at UPC. Three channels have been recorded: membrane microphone, laryngograph and close-talk microphone. The A/D board was set to 96kHz as the sampling frequency and 24 bits resolution.

The recordings where scheduled in sessions of 2-3 hours, with pauses every hour. Two people attended the recording sessions. The first one controlled the signal levels and the start and stop of the recordings. The second make notes in the prompts if the speaker made a change in the words or a specific pronunciation. If the pronunciation was absolutely natural and correct (but slightly different to the prompt) we did not repeat the recording but changed the text files.

2.4. Labeling

All the speech data is labeled manually prosodically and phonetically. For the prosody, a broad labeling is used (major and minor breaks, accented words, emphasis). The phonetic transcription consist on a manual revision of the canonical transcription. The labeler do the phonetic and prosody annotation at the same time, listening and viewing the signal.

The phonetic segmentation and the pitch labeling is not supervised. It has been automatically created using our own tools. Phonetic segmentation uses HMM-based forced alignments. In the first step, the HMM toolkit finds the pauses and the pronunciation variant. In the second step, the phonetic segmentation is derived. Furthermore, for each unit, we provide a confidence measure. We usually prune the databases based on that measure to avoid segmentation errors (Adell et al., 2006).

2.5. Multilingualism

As has been stated in the introduction, in the reading of many Catalan texts appear many Spanish words and also frequently English words.

The Spanish words are uttered using the Spanish phonetics. Therefore, some phonemes (the SAMPA /T/ and/x/) need to be included in the database. Also, the unstressed vowels /e/ and /o/ are very rare in Catalan but very frequent in Spanish. Therefore, we have designed a Catalan+Spanish phone set which is used to define the coverage criteria in the corpus design.

Also, as in many other languages, foreign words (in particular, English words) appear in Catalan texts. The pronunciation of these words depend on each particular speaker. However, in many cases they are mapped in the Catalan+Spanish phone set. Therefore, we have not extended the phone set for foreign words. We have analyzed also the other languages spoken in Spain: Vasc and Galician. While Galician can be represented with the previous phone set, the Vasc language has several fricatives. We have added this phonemes to the phone set and ensured a minimal coverage at the level of phoneme.

In order to ensure coverage, we have introduced Spanish words in the *news* corpus. Proper names have been detected and substituted by Spanish proper names to ensure the phonetic coverage. Furthermore, a specific Spanish corpus has been recorded (see table **??**). Some English names have been added in the *news* corpus. Furthermore, several English words appear in the software manuals and in the screen reading corpus. For the Vasc sounds, a Vasc native speaker has selected and recorded proper nouns for each new phoneme. The words have been included in the *news* corpus. The speaker listened to the examples of each sound before reading the set of sentences with the sound. However, the speakers found very difficult to utter those sounds and we need to evaluate the result.

A interesting effect of the bilingualism is the reading of codes, emails or web pages. In Spain (and Catalonia) we try not to spell, because the Spanish language is phonetic and there is an almost direct equivalence between graphemes and phonemes. As the pronunciation ambiguity in Catalan is higher, we have noticed that when precise information needs to be read, many Catalan speakers switch to the Spanish phone set and pronunciation rules. In this way, the ambiguity is avoided.

3. Festival Voices

The final goal of the project was to provide 2 speakers (1 male and 1 female), based on 10h recordings. However, the 10 selection speakers contain enough data (1 hour), similar to the ARCTIC databases (Kominek and Black, 2003) and reasonable TTS voices can be produced.

We have used two technologies provided in the Festival framework: *clunits*, based on the unit selection technology, and *HTS*, based on HMM synthesis.

The HTS voices are very small, less than one megabyte, compared with more than 100 megabytes for each clunit voice. The HTS voices sound are very intelligible. Although it is evident that the voices are synthetic, they can provide the required functionality for the LinKat distribution. The *clunit* voices sound more natural but from time to time, yo notice errors in concatenation or prosody.

All the Festival voices are available at the web page of the project (FES, 2007)

The festival front-end has been extended to cope with Catalan. This includes the normalization rules, lexicon and letter-to-sound rules, and the POS-tagger.

We have produced a lexicon with around 60,000 words. The word forms are based on the lexicon provided in the FreeLing project (Carreras et al., 2004), adding new words found in the text collected to prepare the corpus. These 60,000 words have been transcribed automatically using our in-house transcription tool. Letter-to-sound rules have being derived using Festival modules. The phonetic transcription accuracy is very high for Catalan words found in the lexicon and quite acceptable for unseen words.

4. Conclusions

In the paper have described the resources produced for creating synthetic voices in Catalan. The resources are freely available for research and commercial use. The produced resources can be summarized as:

4.0.1. Summary of the produced resources

- High quality recordings: 10 speakers, 1h. To our knowledge, the only initiative similar to this one is the ARCTIC project (Kominek and Black, 2003) but we provide more speakers and all of them are professional speakers.
- High quality recordings: 2 speakers, 10h. To our knowledge, this is the first time that such amount of data for synthesis is freely available for research and commercial use..
- 10 hts voices and 10 clunits voices ready for using in the Festival TTS
- Catalan front-end for Festival, including word normalization, lexicon, letter-to-phoneme classifier, and POS tagger.

5. Acknowledgment

We thank the Catalan Government for their support in this project. We also thank Inma Hernaez and Eduardo Rodríguez Banga, for their support in defining the Galician and Vasc phone sets and providing the needed data.

6. References

Jordi Adell, Pablo D. Agüero, and Antonio Bonafonte. 2006. Database pruning for unsupervised building of text-to-speech voices. In *Proc. of ICASSP*, volume 1, pages 889–892, Toulouse, France, May.

Antonio Bonafonte, Harald Höge, Imre Kiss, Asunción Moreno, Ute Ziegenhain, Henk van den Heuvel, Horst-Udo Hain, Xia S. Wang, and Marie-Neige Garcia. 2006. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *Proc. of LREC Conf.*, Genoa, Italy, May.

X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. FreeLing: An open-source suite of language analyzers. In *Proc. of LREC Conf.*, Lisbon, Portugal.

2007. FestCat: Catalan corpus and voices for speech synthesis.

John Kominek and Alan W Black. 2003. CMU ARCTIC databases for speech synthesis. Technical report, Carnegie Mellon University.

Asunción Moreno, Albert Febrer, and Lluis Marquez. 2006. Language resources for the development of speech technologies in catalan. In *Proc. of LREC Conf.*, Genoa, Italy, May.

Alberto Sesma and Asunción Moreno. 2000. Corpuscrt 1.0: Diseño de corpus orales equilibrados. Technical report, UPC, December.