Recent work on the FESTCAT database for speech synthesis

Antonio Bonafonte[1], Lourdes Aguilar[2], Ignasi Esquerra[1], Sergio Oller[1], Asunción Moreno[1]

[1]TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2]Departament Filologia Espanyola, Universitat Autònoma de Barcelona, Bellaterra, Spain

Abstract

This paper presents our work around the FESTCAT project, whose main goal was the development of voices for the Festival suite in Catalan. In the first year, we produced the corpus and the speech data needed for build 10 voices using the Clunits (unit selection) and the HTS (Markov models) methods. The resulting voices are freely available on the web page of the project and included in Linkat, a Catalan distribution of Linux. More recently, we have updated the voices using new versions of HTS, other technology (Multisyn) and we have produced a child voice. Furthermore, we have performed a prosodic labeling and analysis of the database using the break index labels proposed in the ToBI system aimed to improve the intonation of the synthetic speech.

Index Terms: speech synthesis, databases, Festival voices, prosody analysis.

1. Introduction

Some years ago, the Catalan Government promoted the production of *Linkat*, a Linux distribution aimed to schools. Speech synthesis is a key component in many accessibility tools, as Gnopernicus or Orca, but Catalan voices were not available at that time in the open-source domain. Therefore, we set up a project to produce speech synthesis corpus and to build voices for the *festival* engine [1]. Not only the *festival* voices, but also the corpus will be released to allow their use in other synthesis engines.

In 2008, the first version of *festival* voices were released. During this year, new version of *festival* voices are being produced. Furthermore, the labeling of the corpus is being improved to analyze the prosody so that better intonation models can be derived.

2. The FestCat Corpus

The primary goal of the FestCat corpus was to produce two synthetic voices (one male and one female voices) for *Festival* [1] with similar quality to the best available English-voices included in Festival. Furthermore, the speech corpus would be public and should allow to produce the best quality when used on state-of-the-art engines.

The design and production process is based on the specifications introduced in the EU TC-STAR project and are described in [2]. Table 1 summarize the *FestCat* corpus.

3. Festival Voices

As stated above, the original goal of the project was to produce 2 high quality voices (one male and one female). However, the speaker selection process produced, as a by product, speech corpora for 10 speakers of reasonable size (1 hour) and phonetic coverage. In the 2008 release we only used the $10\,\text{speakers} \times 1\,\text{hour}$ corpora to produce voices. Two versions of the voices were produced, using two technologies included in Festival:

- *Clunits*: Concatenative speech synthesis using (specific) unit selection
- *HTS*: HMM-based speech synthesis

The *clunit* voices sound more natural that HTS. The vocoder speech model included in HTS and the flat intonation generated resulted in voices clearly synthetic and monotonous. However, while the *HTS* voices were very smooth and stable, the *clunit* voices produced relatively frequent concatenation errors. The reason was either segmentation errors or just spectral, phase or pitch discontinuities. On the other hand, the footprint of *HTS* voices is much smaller. For these reasons, the HTS voices were included in the default LinKat installation while the other voices can be downloaded as additional packages. The different voices can be tried in the *FestCat* web page of the *FestCat* project [3].

During the last year we have produced new voices based on the *FestCat* corpus:

- HTS voices have been trained using the last HTS release. The new version includes HSMM (Hidden Semi-Markov model) to improve duration modeling and, more important, Global Variance (GV). This new feature produces richer prosody and much more natural voices.
- For the big voices (1 male + 1 female) *Clunits* and *HTS* voices have been produced using the whole database. The quality of the new voices are significantly higher than the quality achieved with one-hour corpus.
- Voices using the festival technology *Multisyn* have been build. *Multisyn* is a general (classic) unit selection technology and produces better results than *Clunits*. However, the reasons to prefer *HTS* vs *Clunits* voices in the default Linkat installation are still valid for selecting *HTS* vs. *Multisyn*.
- A child voice has been produced using a new 1-hour corpus.

3.1. Prosody boundaries labeling

As a continuing project, we have been labeling a subset of the *FestCat* database with prosody information. This will allow a study of Catalan prosody both from a linguistic and a technology point of view. The goal with respect to speech synthesis is to assess the generation of synthetic prosody using a symbolic representation.

As a first step, we are labeling information of prosodic boundaries using the break-tier proposed in the Cat_ToBI proposal [4].

Corpus Size	The corpus size is around 90,000 words (aimed to 10 hours of speech).
Corpus Design	80% of the corpus is designed to achieve high phonetic and prosodic variability. Subcorpus from different domains have been produced (novels, news, teaching books, etc.) applying a greedy algorithm to a big raw corpus. Each utterance is a sentence or a short paragraph. For instance, the mean length of the news subcorpus is 25 words. The rest 20% is designed to improve coverage in doamins relevant in many TTS applications, as numbers, cities (from Catalonia, Spain and the word), commands found in screen readers, etc.
Language and Phoneset	The design goal is the Central Catalan dialect, but also Spanish, Galician, Euskera and English words need to be pronounced. The Catalan phoneset has been extended to include the missing Spanish phonemes (as SAMPA [x] and [T] and some stressed vowels). The corpus include a small Spanish subcorpus (20 min.) and some English and Euskera words.
Recording conditions	Recording studio; 96Khz, 24 bits; 3 synchronous channels (membrane microphone, close-talk microphone and laryngograph)
Labeling	Orthography and phonetic supervision. Automatic phone segmentation using HMM based toolkit.
Speaker selection	10 professional speakers (5 male + 5 female) record 1h corpus. Build 10 TTS voices. Select 1 male and 1 female taking into account articulation and phonetic errors, voice stability on long sessions, pleasantness of the voice, quality of the 1h TTS voices, distortion in front of TD-PSOLA manipulation

Table 1: Summary of the *FestCat* corpus

As in other ToBI systems, the procedure is perceptually-based, although the labeler has visual information of the signal. We have used all the levels in table 2 to better capture the relationship among prosodic constituents. In order to mark the absolute end of the elocution, at the end of the file, the level 5 proposed in [5] has been added. This decision has two advantages: first, in declaratives, it serves as the minimum F0 value of the declination baseline; second, it prevents the processing of the silences in this position, without any linguistic content.

Break	Description
0	Any clear example of cohesion between orthographic forms, such as vowel contacts
1	Any inter-word juncture (provided as default at every word boundary)
2	End of groups with some sense of disjuncture with respect to the following speech chunk
3	End of minor prosodic group
4	End of major prosodic group

Table 2: *Break descriptions*

For each main speaker (1 male + 1 female) approx. 5 hours have been labeled manually by a graduate in linguistics, with no prior training in prosodic labeling. The transcriber was looking at a computer screen with a display of the signal (F0 curve and waveform) together with the phonetic marks corresponding to words, syllables and silences. Nevertheless, she was encouraged to attend preferable to perception. To ensure the consistency of the data, only one transcriber was recruited and one of the authors (L.A.) reviewed the corpus. The annotation has not been considered definitive until the transcriber and the reviewer arrived to a consensus in the labels. Roughly, 10% of the words are followed by a minor break (BI3) and 10% of the words are followed by a major break (BI4 and BI5).

In [6] we present a first study of the correlation between acoustic features and break indexes and a classifier based on these features. The preliminary results show that using only some acoustic measures (presence and duration of pause, value of ending F0, duration of pre-break syllable, etc.) we can predict the presence of break in 90% of cases; and we can differentiate between major and minor break in 80% of the cases.

Further work will include linguistic features (for automatic labeling of databases) and will study the prediction based only on linguistic features (for symbolic prosody prediction, in speech synthesis).

4. Acknowledgments

FestCat has been partially funded by the Generalitat de Catalunya under the LINKAT and TECNOPARLA projects. The work done by L. Aguilar was possible thanks to the visiting position at the Universitat Politècnica de Catalunya during the academic year 2008-09.

Authors want to thank all present and past members of the TALP speech synthesis group that were involved somehow in the *FestCat* project.

5. References

[1] A. W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," 1996–2009. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival.html

[2] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez, "Corpus and voices for catalan speech synthesis," in *Proc. of LREC Conf.*, Marrakech, Morocco, May 2008, pp. 3325–3329.

[3] "FestCat: Catalan corpus and voices for speech synthesis," 2007. [Online]. Available: http://www.talp.upc.edu/festcat

[4] P. Prieto, L. Aguilar, I. Mascaró, F. Torres-Tamarit, and M. Vanrell, "L'etiquetatge prosòdic cat_tobi," *Estudios de Fonética Experimental*, no. XVIII, pp. 287–309, 2008.

[5] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. C. Fong, "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, vol. 90, pp. 2956–70, 1991.

[6] L. Aguilar, A. Bonafonte, F. Campillo, and D. Escudero, "Determining intonational boundaries from the acoustic signal," in *Proc. of INTERSPEECH*, Brighton, U.K., Sep. 2009.