Spanish Synthesis Corpora

Martí Umbert, Asunción Moreno, Pablo Agüero, Antonio Bonafonte

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain

{mumbert|asuncion|pdaguero|antonio}@talp.upc.edu

Abstract

This paper deals with the design of a synthesis database for a high quality corpus-based Speech Synthesis system in Spanish. The database has been designed for speech synthesis, speech conversion and expressive speech. The design follows the specifications of TC-STAR project and has been applied to collect equivalent English and Mandarin synthesis databases. The sentences of the corpus have been selected mainly from transcribed speech and novels. The selection criterion is a phonetic and prosodic coverage. The corpus was completed with sentences specifically designed to cover frequent phrases and words. Two baseline speakers and four bilingual speakers were recorded. Recordings consist of 10 hours of speech for each baseline speaker and one hour of speech for each voice conversion bilingual speaker. The database is labelled and segmented. Pitch marks and phonetic segmentation was done automatically and up to 50% manually supervised. The database will be available at ELRA.

1. Introduction

During the last years a big effort has been devoted to build Language Resources (LR) for Speech Recognition. In Europe many of these resources have been designed by research groups and their specifications became standards (SpeechDat, Speecon, Orientel) that were applied in other projects. The EU project LC-STAR[1] specified contents, production and formats for lexica for TTS in many languages including Arab, Turkish, Mandarin, and Russian among others. However, there is not a standard for defining a database for speech synthesis and many differences between TTS systems become from the kind of data and methodology to produce the data.

In the scope of the EU project TC-STAR[2] "Technology and corpora for Speech to Speech Translation" (FP6-506738), a big effort is done to research and create high quality Speech Synthesis systems. The Synthesis part of the project has three application areas

- building the most advanced state-of-the-art TTS systems.
- performing research on intra-lingual and cross-language voice conversion,
- performing research on expressive speech.

In the framework of the TC-STAR project, an important task was to define the specifications for the TTS baseline databases as well as the specifications for voice conversion and expressive speech databases. A complete description including corpora design, procedures for speaker selection, recording methodology, speech labeling and validation criteria can be found in (Bonafonte et al. 2005) and (Bonafonte et al. 2006). Specifications cover three languages: Mandarin, English and Spanish. Currently are applied to other languages under the scope of the ECESS[3] initiative.

The corpora design is divided in various sub-corpora:

1. Baseline corpora: Intended for the baseline system. Contains 10 hours of read recorded speech. Corpora were built from transcriptions of parliamentary speeches, novels and frequent phrases selected from some specific domains.
2. Voice conversion corpora: Recorded by bilingual speakers, contains one hour of read speech in each language (English and Spanish). Corpus was designed by translating a set of sentences taken from the parliament.
3. Mimic sentences: (same suprasegmental structure): Intended for intra-lingual voice conversion.
4. Expressive speech: Designed in the project for a speech-to-speech translation framework in a parliament translation application.

The complete recordings were automatic labelled (transliteration, phonetic, prosodic, phoneme segmentation and epoch labelling). For the baseline voices, phonetic and prosodic labelling was manually supervised. Furthermore, the phoneme segmentation and epoch detection of two hours of speech of each baseline voice were manually supervised as well as one hour of the voice conversion voices.

The database is accompanied with extensive documentation and a lexicon including phonetic transcription, lemma and POS.

This paper is organized as follows. Section 2 summarizes the specifications of the corpora, in Section 3 shows the recording procedure and speakers' selection criteria. Section 4 shows the labelling procedure and Section 5 ends with some conclusions.

[1] http://www.lc-star.com

[2] http://www.tc-star.org

[3] http://www.ecess.org

2. Corpus Design

This section summarizes the specifications of the created LR in the above mentioned three main fields: (i) building speech synthesis systems, (ii) investigating specific research topics in intra-lingual and cross-lingual voice conversion and (iii) investigating in expressive speech synthesis.

For building a single voice in a given language for a state of the art speech synthesis system a total volume of 10h of speech is considered to be adequate. Assuming 0.4 sec duration in average per word 10 h of speech corresponds to the time needed to read a text corpus of about 90 000 running words.

The creation of the database for TTS is based on read speech. Selected speakers read texts from different sources. In order to achieve a good coverage on a variety of different domains, besides from the domains covered within TC-STAR project (parliamentary speech), the corpora was built with collected texts from novels, newspapers and magazines from a set of wide domains.

The amount of data is distributed on sub-corpora as shown in Table 1:

Sub-corpora	Size (# words)	Application
Parallel transcribed speech	9000	Baseline Voice conversion
General transcribed speech	36000	Baseline
Written text	27000	Baseline
Frequent phrases	8000	Baseline
Triphone coverage sentences	8000	Baseline
Mimic sentences	2000	Baseline Voice conversion
Expressive speech	2000	Expressive speech

Table 1. Definition of sub-corpora, size and application

Table 1 summarizes the three types of voices recorded for the intended applications:

Baseline voices:

Composed by a corpora of 90,000 words (about 10 h of speech). The baseline corpora include sentences and paragraphs taken from parliamentary transcribed speech and transcribed broadcast news (45,000 words), and sentences taken from several text sources. The purpose of the text sources is twofold: to enrich the vocabulary and to facilitate the selection of the sentences to achieve a good phonetic and prosodic coverage. The text sources are composed by contemporary novels and frequent phrases from a wide number of defined domains. Finally, a small corpus was manually designed to achieve the phonetic coverage.

Voice conversion:

This corpus contains 11000 words. Sentences were chosen from transcribed texts of the European and Spanish parliament. The corpus is used for both, intra-lingual and cross-lingual voice conversion. For cross-lingual voice conversion, the corpus is translated in two languages (UK English and Spanish). A small part of the corpus is called 'mimic sentences'; the speaker will try to imitate the prosodic patterns of a reference (or template) speaker.

Expressive speech:

The corpus is composed by sentences and paragraphs containing up to 2000 words from transcribed speech. The speaker reads the corpus imitating the original speaker.

Corpus generation

Sentences were taken from text corpora of more than 10 million words. Selection of sentences was done with the following criteria:

Sentences length:

Corpora contain short and long sentences and paragraphs.

Phonetic coverage:

Generating high quality TTS voices from corpora implies that the recorded corpora should have a good coverage of the basic TTS speech segments and their prosodic properties. It is evident that the higher the amount of recorded speech the better should become the coverage. However a compromise between coverage and effort in creating the LRs has to be taken into account.

In the Spanish synthesis database the chosen speech segments for phonetic coverage are triphones. Stress and unstressed triphones were considered distinct triphones. The minimum number of distinct triphones to be included in the database was defined as the number of distinct triphones necessary to cover the 95% of a corpus of more than 10 million words from texts taken from several on-line newspapers and web sites. A double check to prove the validation of the selected triphones was done with texts coming from parliamentary transcriptions (3 million words).

The diphone coverage of the selected corpora was additionally checked.

Prosodic coverage:

A good prosodic coverage is needed to allow to create prosodic models and to provide speech segments suited to be used in all the prosodic contexts to be synthesized. For this purpose, the corpora design included the coverage of supra-segmental prosodic events (e.g. phrase breaks, phrasal and sentence accent and intonation contour, etc). Given the limitations in the total size of the database, the prosodic coverage, (i.e. coverage of supra-segmental prosodic events) was defined for diphones instead of triphones. A "significant" diphone set was defined (Febrer Godayol, A. 2001).as the list of distinct diphones (stress and unstressed) with a significant frequency of occurrence ($fd>10^{-4}$) in the parliamentary texts.

In this work, prosodic coverage for Spanish is based on diphones and their position in a sentence. For this purpose we defined

- Voiced diphones: one phoneme is voiced
- Unvoiced diphones: Both diphones are unvoiced

Position of diphones in sentences:

- Initial : From the beginning of the sentence till the first stressed diphone (included)
- Prepausal: From the last stressed diphone till the end of the sentence. In the Voiced diphones case, three possible sentence endings have been distinguished (point, coma, and question mark)
- Middle: from the first up to the last stressed diphones (not included)

The selected corpora contains a representative set of interrogative and exclamation sentences

As a result, the final baseline corpus contains more than 1200 sentences (from which 297 are questions) and paragraphs, more than 90,000 running words and 7,000 distinct triphones (3600 triphones are stressed and 3400 unstressed). The triphones of the Spanish synthesis database reach about the 98% of the total number of triphones in an independent corpus of 10 million words.

3. Recordings and Selection of Speakers

Recordings were carried out in a silent room SNR_A>40dBA with a reverberation measure RT60< 0,3 sec at 96 KHz sampling rate and 24 bit/sample. The recording software is NannyRecord, an in-house recording tool. Two microphones (a large membrane microphone and a close-talk microphone) plus the laryngograph signal were recorded simultaneously.

For the TC-STAR project requirements i.e., baseline system, inter-lingual and cross-lingual voice conversion and expressive speech was necessary to record the voices shown in Table 2.

# speakers	Kind of voice
1	Baseline voice male
1	Baseline voice female
2	Cross-language conversion voice male
2	Cross-language conversion voice female
1	Template voice
2	Mimic male voice
2	Mimic female voice
2	Expressive speech male voice
2	Expressive speech female voice

Table 2. Voices to be recorded and associated research requirement.

The baseline male speaker uttered the template voice. The four cross-language speakers uttered the mimic and expressive voices. Six speakers is the final number of recorded speakers for the complete database.

The following procedure was applied for the speakers' selection. For the baseline, two speakers, one male and one female were selected from a set of 10 professional speakers. In order to choose the baseline speakers, each one recorded one hour of speech. Signals were phonetically segmented by forced alignment and a speech synthesis voice was built automatically. A listening test was carried out by 10 subjects. Listeners had to score, for each speaker:

1. Pleasantness of their voice
2. Quality of the laryngograph signal
3. Quality of speech manipulated using TD-PSOLA.
4. Quality of synthesized signal. The sentences were chosen to control the number of concatenation points.

Bilingual (UK English and Spanish) speakers were chosen based on their accent and their capability to mimic sentences. Selection was carried out by native speakers of each language. At recording time, recordings were supervised by an expert linguist and an operator. The recordings took place along several days to do not tire the speakers and keep the quality of their speech consistent, even and uniform throughout all sessions along the recordings. A set of instructions were given to the speakers before the sessions start concerning speed, tone, intonation and style. For the baseline speakers, the dominant expressivity is that chosen by the speaker compatible with a professional translator speaking in neutral manner, for the expressive speech their expression had to be compatible with parliamentary sessions.

4. Labeling

The complete recordings were automatic labelled (transliteration, phonetic, prosodic, phoneme segmentation and epoch labelling). For each utterance (speech file) the database provides:

- the prompt text used to elicit the utterance,
- the orthographic annotation,
- the phonetic transcription,
- a rough annotation of symbolic prosody,
- the segmentation into the pitch marks, associated with the glottal closure.

Orthographic annotation

Orthographic annotation is a transliteration of what was actually said by the speaker without ambiguities at word level. Furthermore, if the signal of a given word is not suited for concatenative speech synthesis, the word is preceded by the symbol '*'. Although the speech produced has to match the prompt text, the orthographic transliteration reflects what the speaker actually said coping with minor deviations not detected during the recording phase

Phonetic transcription

The recordings are fully phonetically transcribed. The transcription has to be 100% supervised to annotate what the speaker really said, including elision, reduction or assimilation present in continuous speech. The phonetic transcription includes word and syllable boundaries. The 'pause' between words is included in the phoneme set and in the transcription. A pause is a silence with 'significant' duration.

Symbolic prosody annotation

Phrase breaks were annotated using two levels: minor break (intermediate intonational phrase) and major break (full intonational phrase).
Pitch accent (intonational prominence) was annotated using two levels: 'normal' and 'emphatic'.

Segmentation

Phonetic segmentation

All the signals are segmented automatically and/or manually. Two hours of the baseline voices were manually supervised and a 5% of the conversion voices were checked manually. The segmentation matches the manual phonetic transcription.
For each phoneme, the starting and ending time is provided. A 'middle' point can optionally be provided which indicates a reasonable point to split the speech segments in concatenative speech.
The supervised sentences were chosen with two criteria: All the parts of the corpus (written text, transcribed speech and designed sentences) have been supervised and the chosen sentences were those with lower score in the automatic process. This last criterion allows reviewing problematic cases either from an error in the segmentation or in the signal itself.

Word segmentation

All the expressive speeches are segmented either automatically and/or manually into words. For each word, the starting and ending time is provided. If reduction of parts of words is produced, a 'middle' time between words is provided as starting or ending point.

Pitch Marking

Speech signals of all the baseline and conversion voices are labeled with pitch marks. The pitch marking points are defined with reference to the maximum of signal (maximum is defined in close neighborhood of the positive slope of laryngograph signal).
Two hours of the baseline voices and one hour of the speech conversion voices were checked manually for each sub-corpus. Although the use of a laryngograph makes pitch detection algorithms very reliable some errors can still happen. The supervised sentences were chosen with two criteria: there are sentences from each part of the corpus (written text, transcribed speech and designed sentences) and the selected sentences were those with problematic values in the automatic pitch pattern

5. Conclusions

This paper described the TTS Spanish Language Resources generated in the TC-STAR project. The databases have been extensively tested in the framework of that project, and in the evaluation programs. The database closely follows the specifications (Bonafonte et al. 2005) and consequentially is one of the first steps in the standardization of the synthesis LR, The database will be available via ELRA.

6. Acknowledgement

The TC-STAR project is supported by the EC in the sixth Framework Programme under contract FP6-506738.

7. References

Bonafonte, A. Höge, H., Tropf , H., Moreno, A., van der Heuvel, H., Sündermann, D., Ziegenhain, U., Pérez, J., Kiss.I., (2005) "TTS Baselines and Specifications". Deliverable D8 of the EU project TC-STAR "Technology and corpora for Speech to Speech Translation" (FP6-506738)

Bonafonte, A. Höge, H., Kiss, I., Moreno, A., Ziegenhain, U.,Van den Heuvel, H., Hain, H.U., Wang, X.S., Garcia, M.N. (2006). TC-STAR: pecifications of Language Resources and Evaluation for Speech Synthesis. In : Proceedings LREC2006, Genoa, Italy.

Febrer Godayol, A. (2001) "Síntesi de la parla per concatenació basada en la selecció" Tesi doctoral Universitat Politècnica de Catalunya.