Building high quality databases for minority languages such as Galician

Francisco Campillo*, Daniela Braga†, Ana Belén Mourín*, Carmen García-Mateo*, Pedro Silva†, Miguel Sales Dias†, Francisco Méndez*

* Signal Theory Group, University of Vigo, Spain, {campillo, carmen, fmendez}@gts.tsc.uvigo.es, a_b_mourin@hotmail.com
† Microsoft Portugal, {i-dbraga, i-pedros, Miguel.Dias}@microsoft.com

Abstract

This paper describes the result of a joint R&D project between Microsoft Portugal and the Signal Theory Group of the University of Vigo (Spain), where a set of language resources was developed with application to Text–to–Speech synthesis. First, a large Corpus of 10000 Galician sentences was designed and recorded by a professional female speaker. Second, a lexicon with phonetic and grammatical information of over 90000 entries was collected and reviewed manually by a linguist expert. And finally, these resources were used for a MOS (Mean Opinion Score) perceptual test to compare two state–of–the–art speech synthesizers of both groups, the one from Microsoft based on HMM, and the one from the University of Vigo based on unit selection.

1. Introduction

Text–to–Speech systems have improved their performance drastically in the last years, to the extent that every stage in their development has to be thoroughly taken into account. In this sense, the design of high quality speech databases is a key point for the performance of the synthesizer. Therefore, it is important to carefully design a good methodology for stages such as selection of the speaker, recording of the corpus and the selection itself of the sentences to be recorded.

As many other languages, Galician (González et al., 2008) suffers from a serious shortage of speech resources, making difficult its integration in the new world of human language technologies, where the tendency is to do all the interaction between human and machines by means of natural language capabilities. In this sense, the Signal Theory Group of the University of Vigo has been working on Speech technologies in Galician for more than ten years, and Microsoft has a widely developed methodology to build new languages in a short period of time. This way, the cooperation of both groups can help to add Galician to the group of languages with high quality synthetic speech.

This article is outlined as follows: Section 2. summarizes the procedure followed to select the best speaker for the recording, taking into account both objective and subjective criteria; Section 3. describes the design of the corpus to be recorded, a fundamental stage for the performance of the speech synthesizer; Section 4. presents the design of the lexicon, as well as the description of the information it gathers; Section 5. is dedicated to the technical details of the recording sessions; In Section 6. a brief overview of the systems of both groups is given; Section 7. shows the evaluation conducted on these systems, both with the speech corpus previously described, and the results. Finally, Section 8. is dedicated to the conclusions and future lines of research.

2. Voice talent selection

The selection of the speaker was carried out in three stages (Braga et al., 2007a; Braga et al., 2007b) combining both subjective and objective criteria.

First, twelve native female Galician professional speakers with Galician as mother tongue (with experience in radio or television), were selected from a set of candidates, and a short recording of each one was performed in order to develop an online perceptual preference test to choose the best five of them according to features such as pleasantness, intelligibility, correct articulation and expressiveness.

Second, the best five speakers according to this test participated in a second session of recording, about one hour length. The aim of this second test was to study the robustness of the voice, taking into account features such as the reading rhythm and the amplitude of the speech signal.

Finally, the best speaker was chosen combining the results from the two tests.

3. Design of the corpus

The corpus consists of 10000 Galician isolated sentences between 1 and 25 words length, extracted from a newspaper and belonging to different types: declarative, interrogative, exclamatory, ellipsis and lists of numbers. The corpus was designed to be a good sample of the Galician phonemic and syntactic characteristics. Therefore, a greedy algorithm was used for the selection of the sentences, taking into account different criteria. First, a good phonemic coverage, with a variety of phonemes in the different contexts, according to the Galician language distribution. And second, a variety of syntactic structures was also looked for, in order for the corpus to be useful as a prosodic corpus too. In this sense, sequences of phrases (Noun phrase, Verb phrase, Adjective phrase, Adverb phrase, different types of conjunctions, etc), considering pauses as delimiters, were used as the input for the greedy algorithm. As an example, Table 1 shows the percentage of appearence of the most common sequences of phrases in Galician, estimated in a large text corpora extracted from a newspaper.

Finally, these sentences were manually reviewed by a linguist expert.

4. Lexicon

The lexicon comprises the most frequent words in Galician, and includes information about phonetic transcription, syl-

Sequence of phrases	Percentage (‰)
NP	98.9‰
PrepP	52.0‰
NP + PrepP	30.1‰
AdvP	26.1‰
PrepP + PrepP	21.0‰
NP + PrepP + PrepP	10.1‰
VerP + NP + PrepP	6.2‰
NP + Conj(and) + NP	5.0‰
NP + AdjP + PrepP	4.5‰

Table 1: Sequences of phrases and their percentage of appearance

labification, stress morphosyntax, and origin of the word in some cases.

Phonetic transcription was obtained first as the output of the University of Vigo synthesizer and then manually reviewed by a linguist expert, who mainly corrected errors related to mid vowel openness (/e/ and /o/ versus /E/ and /O/), which is a difficult problem in Galician since sometimes the distinction depends on etymology (González et al., 2008).

With regards to morphosyntax, the lexicon contains information about the part–of–speech (POS) tag (noun, adjective, verb, adverb, preposition, etc.), some subtypes (for instance, type of pronoun, type of conjunction, etc.) and number, gender, person or inflection where necessary. Same as with phonetic transcription, lexicon entries were firstly analyzed with the University of Vigo synthesizer to get all of the possible tags for each word (Seijo et al., 2004), and then reviewed by the same linguist.

5. Recording of the corpus

On an initial stage, different professional recording studios in the area of Galicia were contacted and visited, in order to select the most suitable one for recording a speech corpus. In the end, *Estudios Musicales Metrópolis*[1], in Vigo, was the chosen one. The microphone was a *Neumann*, and *Pro Tools HD3* was the software used for the recording. The edition and segmentation of the recordings was handled by the Brazilian company *Produlz*[2], with many years of experience in this area.

The work was organized in 9 sessions of $5-6$ hours each, taking short breaks every hour and every other time the speaker needed it. Three people attended the sessions to pay attention to technical recording issues, errors in the pronunciation of the sentences, and variations in the rhythm or amplitude of the realization along the recording. Any sentence considered to be wrong was discarded and rerecorded.

[1]http://metropolis.estudiosmusicales.es/

[2]http://produlz.com/

6. Description of the systems

This section describes the main features of the two synthesizers developed in both groups.

6.1. University of Vigo

The TTS system of the University of Vigo is a state of the art unit selection speech synthesizer that uses the demiphone as the basic unit for concatenation. In this kind of systems, synthetic speech is generated by means of the waveform concatenation of acoustic segments extracted from natural speech (Hunt and Black, 1996), under the assumption that synthetic speech will be indistinguishable from natural speech as long as the segments are used in contexts similar to the ones they were extracted from. The best sequence of units is chosen by dynamic programming, with a target cost function that measures the differences between the target unit and a candidate unit from the database, and a concatenation cost function that measures the distortion related to joining two units from the database.

Intonation modeling is performed by means of unit selection (Campillo and Banga, 2006), with the accent group (defined as a sequence of unaccented words ending in an accented one) as basic unit. Regarding duration, different linear regression models are trained for each phoneme class.

Figure 1 is a block diagram of the unit selection stage of the synthesizer: several intonation contours are generated in the intonation unit selection stage, and for each of them a best sequence of demiphones is chosen by means of traditional acoustic unit selection. In the end, the best combination of intonation contours and acoustic units is chosen.

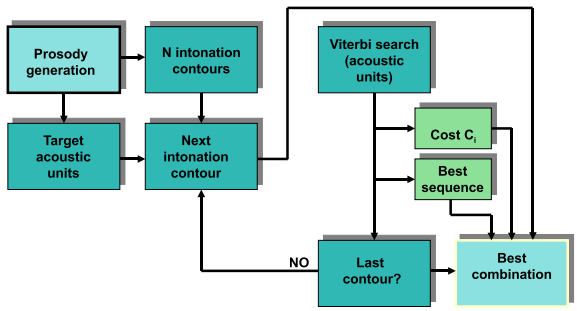

Figure 1: Flow graph of the unit selection stage of the University of Vigo synthesizer

Finally, the waveforms of the speech units resulting from unit selection are concatenated using a TD-PSOLA (Moulines and Charpentier, 1990) based method. Units with a prosody close enough to the target one are not modified, in order to preserve as much as possible the quality of the original recording.

6.2. Microsoft

The front-end of the system is dictionary-based, being composed by a lexicon with approximately 93 thousand words, tagged with phonetic transcriptions, stress marks and syllable boundaries, and with POS information. The stress and syllable marking was automatically assigned using linguistic rule–based algorithms developed by University of

Vigo (González et al., 2008). The front–end is also composed by the text analysis, which involves the sentence separator and word splitter modules and includes a couple of other files, such as phone set and features and the POS tags set. It also includes the TN (Text Normalization) rules, the homograph ambiguity (also polyphony) resolution algorithm, a stochastic–based LTS (Letter–to–Sound) converter to predict phonetic transcriptions for out–of–vocabulary words and the prosody models, which are data–driven using a prosody tagged corpus of 2000 sentences. In this stage of the system, the prosody models were not enabled yet because the prosody tagged corpus is still not complete. The voice font building is also a very complex and demanding process that requires the following steps: script selection, as described in Section 3., recording process, edition and quality control as described in Section 5., wave process, automatic alignment and quality validation, font compiling and conversion of the original recorded waves to 8 Khz, 8 bits sample rate, since the main goal is to apply it in telephony applications. The front–end outputs phonetic transcriptions that are subsequently input of the HMM–based TTS engine (Tokuda et al., 2000) or back–end, which then outputs synthetic voice. Figure 2 illustrates the system workflow.

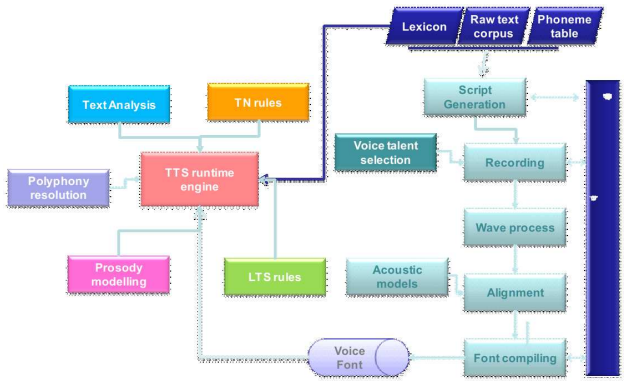

Figure 2: Flow graph of the Microsoft system

7. Evaluation

The recorded voice was integrated in the synthesizers of both groups, and a MOS (Mean Opinion Score) test was conducted, which is a very interesting comparison itself, given that HTS and unit selection are the most widely used synthesis techniques nowadays.

7.1. Description of the test

A pairwise comparison test was conducted, where evaluators were asked to listen to the samples of each sentence (presented in random order to avoid sorting preference effects) and score them in a 1–5 scale, as shown in Table 2. In order to facilitate the evaluation, listeners were allowed to listen to the sentences as many times as they wished. Evaluators were asked just about their preference, not about specific aspects of the waveform. In order to confirm the validity of the listeners' choices, each test included one sentence where the two synthetic realizations were identical. Hence, any listener finding a difference in this control sentence was discarded for the computation of the final results.

The test sentences were manually designed by a linguist expert, and consisted mainly of isolated sentences between four and twenty words length, and belonging to different types: declaratives, questions, ellipsis, etc. From this set, a subset of 40 sentences was randomly selected for the test, from which different random subsets of 20 sentences were presented to each listener, in order to make the individual tests less arduous.
Finally, it should be noted that the University of Vigo samples had to be downsampled to $8\ KHz$, given that the Microsoft system worked at that sampling frequency.

Score	Meaning
1	"A" system much better
2	" A" system better
3	Equal
4	"B" system better
5	"B" system much better

Table 2: Scores for the MOS test

7.2. Results and discussion

The group of listeners was formed of 33 people from the academic world, both with and without experience in synthetic speech. Only three listeners were rejected as a result of failing to recognize that two realizations were in fact the same. Therefore, the test comprised 570 valid evaluations (30 listeners multiplied by 19 different realizations of sentences).
Proportion tests for each score in Table 2 were conducted, being "System A" the unit selection synthesizer of the University of Vigo (see Section 6.1.), and "System B" the HTS synthesizer from Microsoft (see Section 6.2.). Table 3 shows the 99% confidence intervals for each test, as well as the corresponding $p - values$. Figure 3 depicts a pie chart with the mean values of each proportion test.

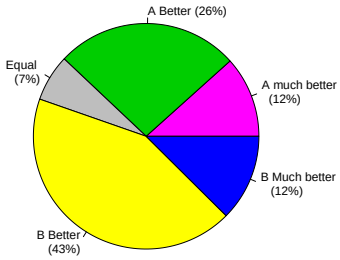

Figure 3: Results of the MOS test

As shown, results are clearly favourable for the Microsoft system, whose stage of development had reached the "intelligible" version, still somewhat distant from the final version. This version included basic text analysis (with word

Rating	Confidence interval (%)	$p - value$
"A" Much better	9.3 - 14.5	$6.67e^{-82}$
" A" Better	22.3 - 30.0	$3.04e^{-32}$
Equal	4.9 - 9.0	$6.72e^{-104}$
"B" Better	39.0 - 46.9	$4.40e^{-4}$
"B" Much Better	10.0 - 15.3	$1.33e^{-78}$

Table 3: Proportion tests for the results of the MOS test: 99% confidence intervals and $p - values$

breaking, sentence breaking, word parser, no compounding, no polyphony, no morphology, no vowel liaison and no prosody marking), in the front–end module and the HTS back–end module trained with a 10000 utterance voice–font.

The feedback from the evaluators pointed out that, although there was a system that clearly outperformed the other one regarding prosody and naturalness, the presence of artifacts made them prefer the more intelligible system. This result coincides with other comparisons: while Unit Selection generates more natural synthetic speech, HTS is more stable. However, a more detailed test than the one conducted here would be needed for a confirmation.

It should be emphasized that in the generation of the Microsoft samples no prosody estimation had been included yet, so there is still room for improvement. On the other hand, a detailed analysis of the artifacts of the unit selection system showed that they were caused by a problem with the pitch tracking algorithm: pitch marks were not always located at the same point of each period, which caused discontinuities of up to $30\ Hz$ at the concatenation points, regardless of the close f_0 at both sides.

8. Conclusions and future lines

This paper describes the result of a cooperation between Microsoft Portugal and the Signal Theory Group of the University of Vigo to develop high quality resources with application to speech synthesis for a minority language such as Galician.

The whole process to record a speech corpus was followed. An appropriate professional studio was chosen, and a voice talent selection was performed taking into account both objective and subjective features. 10000 sentences were selected in order to conform a good sample of the Galician language, regarding both phonemic and syntactic richness. In the recording sessions, three people took care of the different errors that could take place: technical, pronunciation, and variations in the rhythm or amplitude of the waveform. Every sentence with some of these errors was rerecorded.

A lexicon of more than 90000 entries was designed and manually reviewed by a linguist expert, with information about phonetic transcription, syllabification, stress, morphosyntax and origin of the word in some cases. Depending on the POS tag, other features such as number, gender, person or inflection were included.

Using these resources, a pairwise comparison test was conducted between two speech synthesizers belonging to the main current trends in this technology: unit selection and HTS. The results in Section 7.2. show a clear preference for the more stable synthetic speech of the HTS system. The comments of the evaluators remarked that they found the samples from the unit selection system more natural and human–like, but the presence of artifacts made them prefer the other ones. In both systems the next step seems to be clear: finalize the missing front–end features (compounding, polyphony, morphology, vowel liaison and prosody marking) in the HTS system, and design some method to detect and/or correct pitch marking and segmentation errors in the unit selection system. In any case, HTS technology seems to be more robust to this kind of errors, which is something very attractive when working with large corpora.

9. Acknowledgments

The work reported here was partially funded by MEC under the project TEC2009-14094-C04-04 "Búsqueda de Información en Contenidos Audiovisuales plurilingües" and Xunta de Galicia "Isidro Parga Pondal" research programme.

10. References

Daniela Braga, Luis Coelho, F.G. Resende, and Miguel Dias. 2007a. Subjective and Objective Assessment of TTS Voice Font Quality. In *XII International Conference Speech and Computer - SPECOM*, Moscow, October.

Daniela Braga, Luis Coelho, F.G. Resende, and Miguel Dias. 2007b. Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality. In *Advances in Speech Technology - International Workshop*, pages 129–138, Maribor, June.

F. Campillo and E. Rodríguez Banga. 2006. A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems. *Speech Communication*, 48:941–956.

Manuel González, Eduardo R. Banga, Francisco Campillo, Francisco Méndez, Leandro Rodríguez, and Gonzalo Iglesias. 2008. Specific features of the Galician language and implications for speech technology development. *Speech Communication*, 50(11–12):874–887.

A. Hunt and A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP*, volume 1, pages 373–376.

Eric Moulines and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467.

Lorena Seijo, Ana Martínez, Francisco Méndez, Francisco Campillo, and Eduardo R. Banga. 2004. A Galician textual Corpus for morphosyntactic tagging with application to text–to–speech synthesis. In *Proceedings of LREC*, pages 1759–1762, Lisbon.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of ICASSP*, volume 3, pages 1315–1318.