Aromo: Argentine Spanish TTS System

Conference Paper · November 2012

CITATIONS
6

READS
511

3 authors:

Humberto Maximiliano Torres
National Scientific and Technical Research Council
49 PUBLICATIONS **186** CITATIONS

SEE PROFILE

Jorge Gurlekian
Universidad de Buenos Aires
93 PUBLICATIONS **487** CITATIONS

SEE PROFILE

Christian Gustavo Cossio-Mercado
Universidad de Buenos Aires
9 PUBLICATIONS **24** CITATIONS

SEE PROFILE

Aromo: Argentine Spanish TTS System

Humberto M. Torres, Jorge A. Gurlekian, and Christian Cossio-Mercado

Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA
Av. Córdoba 2351, 9 Piso Sala 2. Ciudad Autónoma de Buenos Aires, Argentina
hmtorres@conicet.gov.ar,jag@fmed.uba.ar,cgcossio@rucatech.com.ar

Abstract. This paper introduces Aromo text-to-speech system for Argentine Spanish, which was designed for telephony applications and is based on unit selection and concatenation. The system operates as a client-server engine that supports MRCP, SIP and SSML technologies. Perceptual evaluation results show that Aromo's voice achieve high performances in both naturalness and intelligibility.

Keywords: Text-to-Speech; Argentine Spanish TTS; Unit-selection Synthesis

1 Introduction

The Laboratorio de Investigaciones Sensoriales has created Aromo, the first system for converting text into speech designed entirely for the Spanish spoken in Argentina. Its female voice is the first one for this variant of Spanish. It was designed to work as a TTS engine for telephony applications, with a client-server structure that supports MRCP [11], SIP [16] and SSML [17] protocols. It was written in C/C++ under Windows operating system, and works natively in 32 and 64 bits.

2 TTS System Overview

Aromo architecture is modular as other typical TTS systems [3,2,15]. It consists of three main modules: text processing, prosody generation, and unit selection and concatenation. Below, each of the modules will be briefly presented.

2.1 Text Processing

Input text is processed in several sequential stages: first, to put it in orthographic expanded form, by text normalization and tokenization module; then, to extract the sound sequence to be pronounced, by grapheme-to-phoneme conversion module; and finally to extract the text features to be used as input when predicting prosody, by text feature extraction module.

Text normalization covers the following steps: 1) number expansion according to Spanish language rules; 2) translation of abbreviations and acronyms according to a user-defined lexicon or spelling rules; 3) checking of all words in order

to define if they can be uttered by the system and, if not, notification of need of rewriting; and 4) splitting of input text into sentences delimited by punctuation and synthesis control marks.

Grapheme to phoneme conversion module comprises a set of rules that map letters and their corresponding sound according their context [7].

Prosody prediction and unit selection modules require information from the input text as, e.g., sentence type, morphosyntactic structure and location of syllables with lexical stress. Text feature extraction module is responsible for extracting all the features that will be used in the subsequent steps of the process.

Part-of-Speech (POS) tagging is performed using a model trained with a corpus of sentences from newspapers specifically built for this purpose, using a Maximum Entropy tagger [12]. The used tagset [4] is hierarchical —our system uses the first two levels of each tag— and follows EAGLES recommendations [10]. The resulting POS Tagger is available as a web server registered as system service, and its performance level is greater than 98% [19].

2.2 Prosody Prediction

Our system predicts three prosodic features: pauses, segmental duration and fundamental frequency contour (F0).

Pauses are inserted as marked by orthographic symbols or markup tags but also using a prediction model that identify pauses not marked in the text based on POS tags and a set of rules [18].

Duration of each phoneme and pauses are predicted based on the information extracted from the text, using an artificial neural network (ANN) [21] that receives, for example: identity, articulation and sound type of neighboring phonemes; POS tags; lexical stressed condition of syllable; position of syllable into the sentence.

We used the Fujisaki intonation model to predict F0 contour [6], introducing strong linguistic constraints in order to reduce the number of free parameters of the model: one phrase command per intonational phrase located near the beginning of each phrase; and one accent command per content word, located near to the stressed syllable. The parameters of the model are predicted through Classification and Regression Trees [20,22], using the following features as input: location and length of the intonative phrase; identity, location and length of the stressed vowel; identity of the context phonemes; POS tag; distances to the previous and next stressed vowel that has an accent command associated; and parameters of the previous accent or phrase command.

2.3 Unit Selection and Concatenation

Data provided by text processing and prosody prediction modules define a sequence of target phones. With that information the system looks for the optimal sequence of speech units using a generic Viterbi search, minimizing both an objective cost which indicates the distance to the desired unit sequence and a

concatenation cost that takes into account the possible artifacts generated by the paste of the units. In order to quantify these costs, we used physical parameters such as energy, F0 and durations, as well as context parameters such as stress condition, phonetic context, and position within the word and phrase.

In our system we evaluate 37 parameters for calculating total cost. With all these measures, we perform a linear combination that reflects the fitness of each unit to be selected. We decided to use non-linear weight functions so as to avoid generating perceptible artifacts, which greatly degrade the resulting speech quality, and calibrated weight values manually.

Once obtained the optimal sequence of units we proceed to its retrieval from the database and posterior concatenation in order to generate, without modifications, the requested speech.

3 Emilia's Voice Building

The development of a voice requires the adaptation of currently available prosody prediction models and the creation of a database for implementing the unit selection algorithm.

3.1 Building Corpus and Database

When designing textual corpora for a database to be used on a TTS system it is necessary to take into account coverage of phonemes as well as variations in prosody, sentence types and syntactic structures. Our corpus was created taking into account these requisites in a sequential manner. First, we guaranteed the presence of 97% of all Spanish syllables, in both stress conditions and all possible syllabic positions within a word [8]. Then, we added different types of declarative and interrogative sentences with different lengths and syntactic structures. Finally, we checked the phonetic coverage. As a result of this process the corpus resulted on 1851 sentences.

This corpus, of approximately 2.5hs, was recorded by a professional female announcer native of Buenos Aires, instructed to read the sentences with natural tonal variations. Recordings were made in a sound proof chamber, with an AKG dynamic microphone and a sampling of 16Khz/16bit.

The entire corpus was semi-automatically tagged into six tiers: sentences, words, syllables, phonemes, diphonemes and part-of-speech. First we made a first approach by automated methods, and then followed a manual correction done by trained speech therapists. Both phones and diphones transcriptions were automatically aligned with the corresponding speech signal by using two speaker-dependent Hidden Markov Models (HMMs) implemented with HTK speech recognition toolkit [9]. For phonemes we used standard three-state HMMs, and for diphonemes we used four-states model, taking as boundary point the frame in which the transition between the intermediate states occurs.

Sound files and labeling information are used to build the voice database. All information necessary to perform the unit selection process and data for

subsequent concatenation are both created and stored in the data server. This task is done automatically with a tool specially developed for this purpose.

3.2 Building Prosody Models

Although it is not mandatory to adapt prosody prediction models when designing a new voice, it is an recommendable task in order to achieve a good synthesis quality.

To adapt the phone duration models, first we automatically extract the input features from the data corpus and then train an ANN for each phoneme as a predictor of duration, optimizing network parameters configuration for best performance for each phoneme using stratified cross validation. Finally, ANNs are packaged into a file that will be loaded into the data server for future usage.

Similarly, to reestimate the rules that predict the F0 contour model, we use a set of routines that automatically extract the information from the corpus data and then creates a pseudocode with the rules that are to be used for prediction. Finally, we encode and recompile this module.

Prediction pauses model does not need to be updated when creating a new voice.

4 Voice Evaluation

We used three subjective evaluation methods for assessing our TTS system quality, as reported in [5]: ITU P.85 Recommendation, Semantically Unpredictable Sentences (SUS) and Mean Opinion Score (MOS).

ITU.P85 recommendation [14] was created to assess subjective performance for overall quality of speech voice output devices. Topics evaluated are: % of word recognition, overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness —using an ordinal scale of five categories— and acceptance/non-acceptance.

SUS test [1] is focused on evaluating the intelligibility of systems. SUS sentences are syntactically correct but have no meaning and present a very low word predictability. Its measure is the percentage of well recognized words per sentence.

In MOS evaluation [13] listeners use a fixed scale from 1 to 10 to evaluate overall quality. This is a general purpose test and provides average scores for natural and artificial speech.

The stimuli were created according to the recommendations given by three methods. A total of twenty listeners, 23-45 years old, participated in the experiment. Ten of them were Speech Language Therapists (SLT) and other ten were non-expert (non-SLT) without auditory impairments. Figure 1 shows the performance of Emilia voice for the three methods tested and the two groups of listeners.

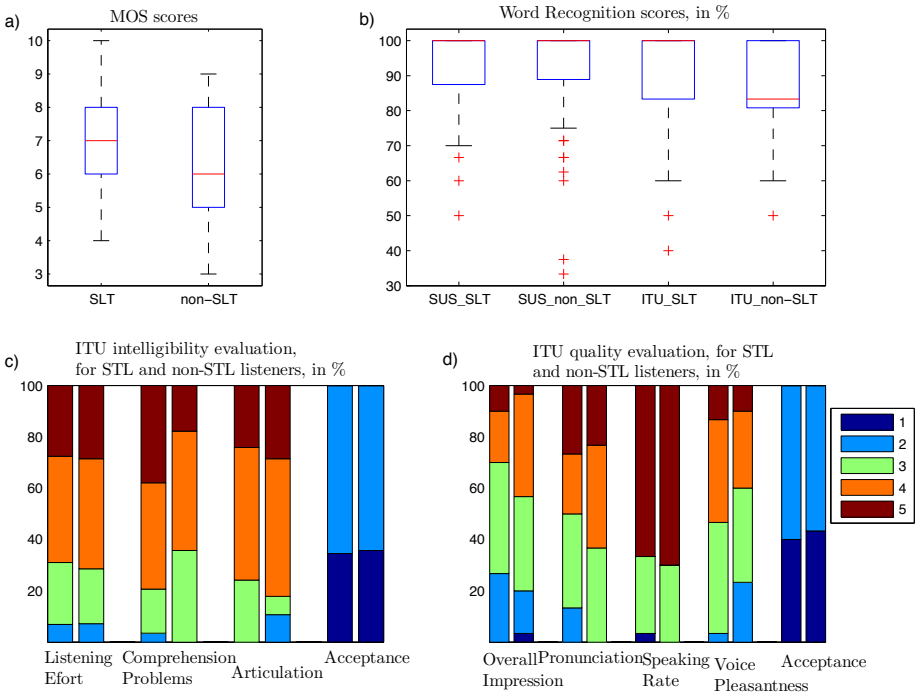

Fig. 1. Perceptual evaluation results: a) MOS quality scores; b) SUS and ITU word recognition scores, in %; c) four topic ITU inteligibility scores, in %; and d) five topic ITU quality scores, in %. The graphs discriminate the results obtained with Speech Language Therapists (SLT) or non-expert (non-SLT). Adapted from [5]

5 Conclusions and Future Work

In this paper we presented a brief description of Aromo TTS system for the Argentine Spanish. We have shown a general outline of its components and how they interact. We also included details of the process of creation of the voice and the results obtained in speech quality perceptual tests. Our system achieved a high performance comparable to other well-known commercial systems [5].

We are now working on creating a new male voice, analogous to the female voice that is currently available. Additionally, we are rebuilding the voice of Emilia in order to obtain greater quality for offline use.

6 Acknowledgements

To the Ministry of Science and Technology and Productive Innovation of Argentina.

References

1. Benoît, C., Grice, M., Hazan, V.: The SUS test. A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication 18(4), 381–392 (June 1966)
2. Clark, R.A.J., Richmond, K., King, S.: Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication 49(4), 317–330 (2007)
3. Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Text, Speech and Language Technology, V. 3, Kluwer Academic (1997)
4. EAGLES: `http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html`
5. Gurlekian, J., Cossio-Mercado, C., Torres, H., Vaccari, M.E.: Subjective evaluation of a high quality. In: Proc. of IberSPEECH. p. accepted for publication. Madrid, Spain (November 2012)
6. Gurlekian, J.A., Torres, H.M., Colantoni, L.: Evaluación de las descripciones analítica y perceptual de la entonación de una base de datos de oraciones declarativas de foco amplio para el español hablado en buenos aires. Estudios de Fonética Experimental XIII, 275–302 (2004)
7. Gurlekian, J.A., Colantoni, L., Torres, H.M.: El alfabeto fonético SAMPA y el diseño de córpora fonéticamente balanceados. Fonoaudiológica 47(3), 58–70 (2001)
8. Gurlekian, J.A., Rodriguez, H., Colantoni, L., Torres, H.M.: Development of a prosodic database for an argentine spanish text to speech system. In: Bird, B., Liberman (eds.) Proc. of the IRCS Workshop on Linguistic Databases. pp. 99–104. SIAM, University of Pennsylvania, Philadelphia, USA (December 2001)
9. HTK: `http://www.htk.eng.cam.ac.uk`
10. Leech, G., Wilson, A.: EAGLES recommendations for the morphosyntactic annotation of corpora. Tech. rep., Expert Advisory Group on Lang. Eng. Stds. (1996)
11. MRCP: Media resource control protocol version 2 (mrcpv2) `http://tools.ietf.org/html/draft-ietf-speechsc-mrcpv2-27`
12. OpenNLP: `http://opennlp.apache.org`
13. P.800, I.T.R.: Methods for subjective determination of transmission quality. Tech. rep., ITU (1996)
14. P.85., I.T.R.: Method for subjective performance assessment of the quality of speech voice output devices. Tech. rep., ITU (1994)
15. Schröder, M., Trouvain, J.: The german text-to-speech synthesis system MARY. International Journal of Speech Technology 6, 365–377 (2003)
16. SIP: Session initiation protocol `http://tools.ietf.org/html/rfc3261`
17. SSML: Speech synthesis markup language `http://www.w3.org/TR/speech-synthesis/`
18. Torres, H., Gurlekian, J.: Automatic determination of phrase breaks for argentine spanish. In: Bel, B., Marlien, I. (eds.) Proc. of the Speech Prosody 2004 (SP-2004). pp. 553–556. Nara, Japan (March 2004)
19. Torres, H.M.: Etiquetado de clase de palabras. Tech. rep., LIS (2010)
20. Torres, H.M., Gurlekian, J.A.: Parameter estimation and prediction from text for a superpositional intonation model. In: Proc of the 20 Konferenz Elektronische Sprachsignalverarbeitung. pp. 238–247. TUDpress Verlag (September 2009)
21. Torres, H.M., Gurlekian, J.A.: Argentine spanish segmental duration prediction. In: Proc. of 13th Argentine Symposium on Technology, 41 JAIIO. pp. 156–167. La Plata, Argentina (August 2012)
22. Torres, H.M., Mixdorff, H., Gurlekian, J.A., Pfitzinger, H.: Two new estimation methods for a superpositional intonation model. In: Proc. of INTERSPEECH 2010. pp. 50–53. Makuhari, Japan (September 2010)