Description of the AhoTTS Conversion System for the Basque Language

Inma Hernaez, Eva Navas, Juan Luis Murugarren, Borja Etxebarria

University of the Basque Country

inma, eva, juanluis, borja @bips.bi.ehu.es

1 Abstract[1]

The AhoTTS system (the *aho* part pronounced as the 'ao' in *Mao*) is a modular TTS conversion system that can be used either as a developing tool or as an API. The architecture of the system is multilingual but all the modules are presently developed for the Basque language.

2 General description

The AhoTTS is a modular text to speech conversion system, so it enjoys the many advantages of this structure described in [1]. The system is specially designed to develop independently all the modules that integrate a TTS system, allowing developers to work on different modules of the system at the same time.

The system uses four main processing modules: the text processor, the linguistic processor, the unit selection module and the synthesis engine. In addition, two databases are used: the dictionary, used by the first two modules and the synthesis units database, used by the last two modules (see Fig. 1).

There are two versions of the system. In the first version the different modules of the system can be executed separately. This version is used for developing new prosodic models and algorithms. The second one provides an API to use the TTS system from any application.

In the modular version, developed in Linux and available also in Windows as a console application, each module has text files as input and output, except for the last module that produces a wave file. Each module uses as input the output of the preceding module, adding some new information to the data stream.

The API is compiled directly from the same C/C++ source files to different hardware and software platforms. Currently supported platforms are Windows/VC5.0, DOS/djgpp, Solaris/GCC and Linux/GCC. The API is easy to use and provides ways to set up the TTS system. In this way, the system may be configured to use different languages and databases, various prosodic models, as well as different possible synthesis engines. The dictionary and synthesis units database can be shared if multiple calls from different synthesis objects are needed.

The input text must be in the ISO-8859-1 format. Escape sequences can be inserted to control manually the system: currently they are used to insert breaks in the desired points of the text.

3 Text & Linguistic processing

The text processor is capable of expanding numbers, acronyms, dates, times, abbreviations etc. into directly readable characters. It also segments the input text into utterances, sentences and words, and organises these structures as a hierarchically ordered list of items. Each item has linked information such as type of word or sentence, and the data read from the dictionary.

The linguistic module does all the processing needed to transform the input word list into a phone list. Its structure is shown in Fig. 2.

The first step of the linguistic processing is grouping words into 'phrases', according to the information provided by the dictionary and a small set of syntactic rules. Then words are expanded into characters.

[1] Part of this work has been supported by MCYT under Project TIC2000-1005-C03-03

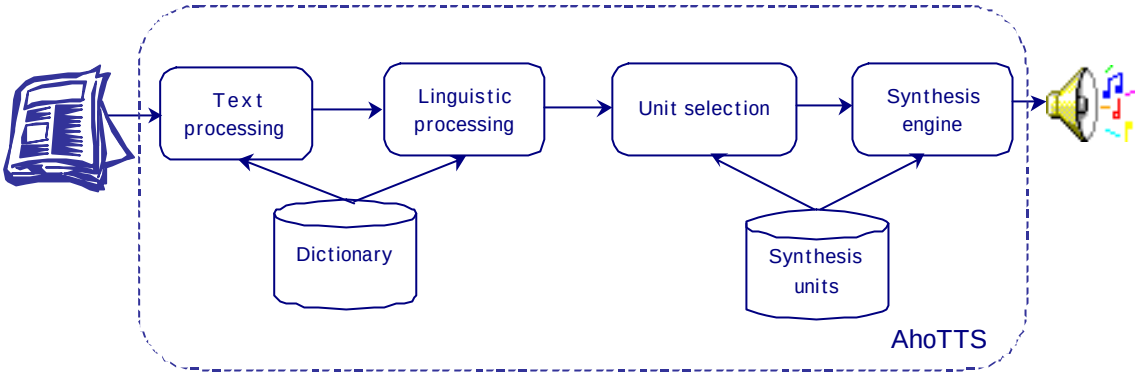

Fig. 1: Modules of AhoTTS system.

Non orthographic breaks are also inserted if needed, and then the phonetic transcription is done by rule using SAMPA phones. Next, after segmentation into syllables, the accents are placed by rule [2].

Prosodic information is then added to each phone. First of all, segmental duration is calculated. AhoTTS has two different models for performing this task: a statistical model that uses CART [3] techniques to decide phoneme duration and a model that sets duration by rule, based on [4].

Once duration is set, F0 values are calculated for each phone. A variable number of F0 values can be supplied for each phone by the F0 model. Currently two F0 models are available. A simple model [5] assigns peaks to the accented syllables and draws valleys to the unstressed syllables. The second model, more elaborated, follows Fujisaki's method [6] and the parameters of the model are derived using CARTs [7].

Linguistic processing ends assigning by rule a power level to each phone.

Both the text processor and the linguistic module use the dictionary to perform their tasks. The dictionary contains mainly acronyms and words with information of different nature about them, but as Basque is a suffixed language and words are constructed by accumulation of suffixes, character groups smaller than words not necessarily equal to morphemes or suffixes are also stored.

Each entry has as much linked coded information as necessary. For example, an acronym usually needs only the expanded expression, but a word might contain syntactic information, the position of the accent, indication about breaks... Up to 14 different groups of properties can be associated to every dictionary entry.

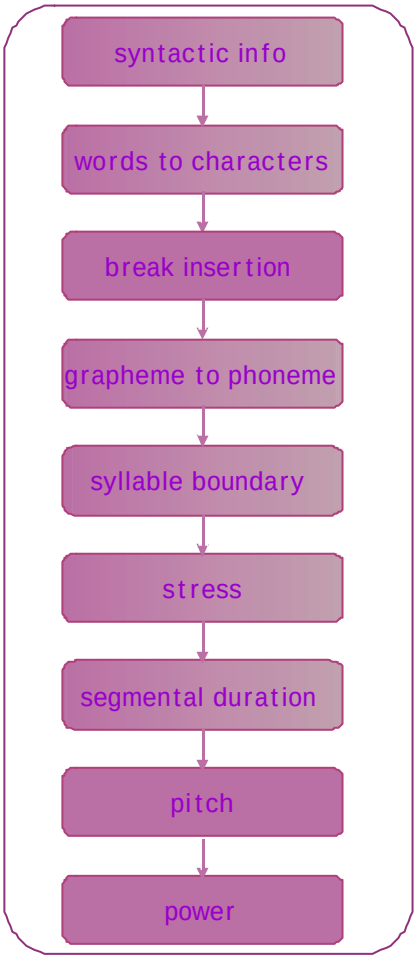

Fig. 2: Structure of the linguistic processor in AhoTTS.

The dictionary is created initially as an ASCII file. Afterwards it might be compiled to a binary file to speed up the search during the synthesis process. The search is performed in both versions if available, first in the ASCII version

and then, if the entry has not been found, in the binary one. This allows the user to write a small ASCII dictionary with customised information that will override the binary dictionary provided with AhoTTS.

4 Unit selection & synthesis engines

From the phone list, the synthesis unit list is constructed using the available units in the database and a set of simple rules. The prosodic information is appropriately moved from every phone to the corresponding synthesis unit. At present, the unit inventory consists mainly on diphones and triphones and a few larger units, although units smaller than the phone are also used for some unvoiced sounds.

The synthesis engines use overlap and adding of the stored waveform units [8]. The waveforms units in the database may have been previously processed to improve the concatenation algorithm [9][10] which modifies the duration and fundamental frequency of the stored units according to the prosodic information supplied. Spectral smoothing at the concatenation points may be applied if desired.

Additionally, at the moment of creating the database, limitations of the prosodic modifications may be introduced for every sound. For example, modification of the duration of a burst in a plosive sound is not allowed.

5 System configuration

A number of parameters can be used to set up the system and control the operation mode of the synthesiser. According to the modules that use these parameters we can group them as follows:

a) Parameters used only by the text and linguistic processing modules:

- **Lang:** Although the system has been developed mainly for the Basque language, the architecture is language independent, and some processes have also been developed for Spanish. The chosen language cannot be changed inside a synthesis process. To use another language, a new object of synthesis should be created. Legal values for this parameter are EU for Basque and ES for Spanish.

- **HDicc:** The dictionary to be used. Different dictionaries may be used if different languages are used, for example. Only one dictionary can be used for each process of synthesis, but several processes of synthesis can share the same dictionary.

- **PthModel, DurModel, PauModel:** specify the pitch, duration and breaks model to use.

b) Parameters used only by the unit selection and synthesis modules:

- **DphDB:** Database to be used in the synthesis process. Only one database can be used for each process of synthesis, but several processes of synthesis can share the same database.

- **Method:** Synthesis engine to be used. Valid values are TDPSOLA and MBROLA.

- **IntrinsicLen:** Intrinsic length of the units in the database may be hold during the synthesis process, thus bypassing the prosodic rules. The allowed values for this parameter are YES (to hold the database length) and NO.

- **TDSpecSmooth:** It is used to perform spectral smoothing at the concatenation point of the units. Its values are YES (to use spectral smoothing) and NO.

- **MrkMode, Mrkfiles:** Allow the insertion of marks inside the speech wave. The different label files that can be created are phone label file (Ph), diphone label file (Dph) and OLA label file (OLA).

c) Common parameters:

- **PthMean, PthDev:** Control the average pitch and pitch range of synthesised speech. They can be changed once a full utterance has been synthesised, i.e., they can not be modified in the middle of an utterance.

- **DurMean:** Sets the average speed of the output speech: the duration specified by the prosody module can be modified to speed up or slow down the output speech.

- **PowMean:** Controls the volume of the sound signal.

6 Present and Future Works

The AhoTTS system is under continuous construction. A demonstration of the system is available at http://bips.bi.ehu.es/tts.

At the moment, the following tasks are being carried on:

- A SABLE interface [11] is being added for the system to support SABLE labels at the input text.
- Two new models of intonation are being developed [12][13].
- An existing formant synthesiser [14] is being integrated as a new synthesis engine.
- A synthesis engine based on HNS [15] modelling is already developed and is being integrated into the system.
- A corpus-based synthesis system [16] is being developed.

7 Acknowledgements

Many people, mainly under graduated students, have participated in the development of the AhoTTS system, and it is not possible to mention them all here. We are grateful to all of them.

We are also grateful to the financial support got over the years from the University of The Basque Country and the Basque Government.

8 References

[1] R. Sproat, J. Olive (1996).
A modular architecture for multi-lingual texto-to-speech
Progress in Speech Synthesis, Springer, New York, 1996.

[2] I. Hernáez (2000)
Euskararko testu-ahots bihurgailua (Text to Speech Conversion system for Basque language) (in Basque)
Txillardegi, lagun-giroan, Txipi Ormaetxea ed., Udako Euskal Unibertsitatea-Bilbo-2000, pp.177-192.

[3] L. Breiman, J.H. Friedman, R.A. Olsen, C. J. Stone (1984)
Classification and Regression Trees
Chapman&Hall, 1984.

[4] D. H. Klatt (1976)
Linguistics uses of segmental duration in English: Acoustic and perceptual evidence
J. Acoust Soc. Am. 59:1209-1221,1976.

[5] I. Hernáez, J.C. Olabe, A. Cuesta, R. Gandarias, P. Etxeberria (1994)
Improving naturalness in a Text-to-Speech Conversion System for the Basque Language
7th Mediterranean Electrotechnical Conference, pp. 61-64, Antalya Turkiye.

[6] H. Fujisaki, K. Hirose (1984)
Analysis of voice fundamental frequency contours for declarative sentences of Japanese
Journal of Acoustic Society of Japan. vol. 5 4 pp. 233-242, 1984.

[7] E. Navas, I, Hernaez, A. Armenta, B. Etxebarria, J. Salaberria (2000)
Modelling Basque intonation using Fujisaki's model and CARTs
State of the art in Speech Synthesis digest, 3/1-3/6, London 2000.

[8] F. Charpentier, E. Moulines (1989).
Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones.
EUROSPEECH'89, Paris, France

[9] T. Dutoit, H. Leich (1993).
MBR-PSOLA : Text-to-speech synthesis based on an MBE re-synthesis of the segments database
Speech Communication, December 1993.

[10] B. Etxebarria, I. Hernáez, I. Madariaga, E. Navas, J.C. Rodríguez, R. Gándara (1999)
Improving quality in a Speech Synthesizer based on the MBROLA algorithm
EUROSPEECH-99, Budapest, Hungary.

[11] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo (1998).
SABLE: A Standard for TTS Markup
ICSLP98, pp. 1719-1724 , 1998.

[12] P. Taylor (2000)
Analysis and Synthesis of intonation using the Tilt model
Journal of the Acoustical Society of America. vol. 107 3, pp. 1697-1714, 2000.

[13] E. López Gonzalo, L. A. Hernández Gómez (1995)
Automatic data-driven prosodic modelling for text to speech
Eurospeech 95, Madrid, Spain.

[14] I. Hernáez, J. C. Olabe, P. Etxeberria, B. Etxebarria, A. Cuesta (1994)
AHOZKA: Un sistema de conversión de texto a voz para el euskara
SEPLN, Boletín nº 14, 1994.

[15] J. Laroche, Y. Stylianou, E. Moulines (1993)
HNS: Speech modification based on a harmonic + noise model.
Proc. IEEE ICASSP-93, Minneapolis, pp. 550--553, Apr 1993.

[16] A. J. Hunt, A. W. Black. (1996)
Unit selection in a concatenative speech synthesis system using a large speech database. In ICASSP'96, Atlanta, 1996