ACTOR®: A MULTILINGUAL UNIT-SELECTION SPEECH SYNTHESIS SYSTEM

Silvia Quazza, Laura Donetti, Loreta Moisa, Pier Luigi Salza

Loquendo S.p.A.
Via Nole 55, Torino, Italy
silvia.quazza@loquendo.it laura.donetti@loquendo.it loreta.moisa@loquendo.it pierluigi.salza@loquendo.it

ABSTRACT

The ACTOR® Text-To-Speech (TTS) synthesis system, developed at Loquendo S.p.A., is here described. The system employs a unit-selection concatenative synthesis technique, relying on labeled acoustic databases providing phonetic and prosodic coverage of the intended language/domain and on an original algorithm for run-time selection of the acoustic units to be concatenated. This technique yields high-naturalness and human sounding voices. ACTOR® is a multi-voice and multi-language system, exploiting different kinds of language dependent knowledge (grammatical, phonetic and prosodic, as well as acoustic) with the support of several development tools (statistical tools for database design, machine learning algorithms, tools for speech signal analysis and phonetic alignment, etc.).

1. INTRODUCTION

Among the traditional text-to-speech techniques, concatenative synthesis seems at present to have won the race for natural sounding artificial speech. Based on the concatenation of speech segments directly extracted from the natural voice of a speaker, it embeds acoustic-phonetic knowledge into the acoustic units themselves. It makes it easier to implement new languages and it is more likely to preserve the natural quality of the original voice. So it provides a good answer to the growing demand for human-sounding multi-language text-to-speech systems, coming from the spreading of voice technology applications in man-machine interaction [1]. Diphones were the most widely used acoustic units till a few years ago, basing their success on their high combinatorial power: a relatively small number of diphones allows the generation of any message in a given language. The drawbacks were the number of junctions and the need to heavily modify the prosodic parameters, resulting in a "robotic" voice. Traditional speech synthesizers were limited by computational difficulties, largely overcome by nowadays computers. Moreover, robust automatic speech analysis and labelling tools are now available, making it feasible to extend the concept of concatenative synthesis towards what is called *corpus-based* or *unit-selection* synthesis. The new technique obtains highly natural-sounding speech by concatenating units longer than diphones and available in many prosodic variants, according to the idea of reducing the number of junctions and the need of prosodic modification [2]. The key factors in this approach are the *phonetic and prosodic coverage* of the intended domain and the run-time *selection criteria* for the acoustic units.

2. HISTORY

The recent developments of speech synthesis research at Loquendo (a spin off of the former Speech Technology Department at CSELT) can be set in this framework. ELOQUENS® was the first commercial CSELT TTS system for Italian. The system, released in '93, was based on diphones [3]. Demands for high quality synthesis output in the applications developed at CSELT imposed to improve on the synthesizer. The successive system, released in '97, was a *specialized* but functionally equivalent version of ELOQUENS® meant to replace it in the Automatic Reverse Directory Service [4]. The system was based on a predefined dictionary of acoustic units larger than diphones to be concatenated for synthesizing only isolated words in the restricted domain of telephone directory. Unrestricted-text general-purpose multi-voice multi-language human-sounding synthesis was then requested for going on to the whole market of voice access services. Hence the evolution moved towards a *unit selection* technique. The acoustic modules of the system were entirely redesigned around the idea of corpus-based, high-naturalness concatenative synthesis. The following strategy was adopted for the newly developed system: a) to rely on labeled acoustic databases from which to extract the longest and best fitting units, without pre-defining their size and nature, and b) to apply prosodic adjustments only where necessary. The new system was firstly presented in [5] and then called ACTOR®. Its architecture is multilingual and multi-voice, relying on separate acoustic databases for the different voices. Appropriate phonetic coverage methods are applied in the acoustic database design. Automatic processing tools perform signal analysis and labelling and the synthesis algorithm exploits the speech database at its best. Since its first release, the system has improved, together with its development environment, and several languages and voices have been implemented.

3. ACTOR® GENERAL FEATURES

ACTOR® is a commercial Multilanguage/Multivoice Text-To-Speech synthesizer, attaining great acoustic naturalness and linguistic accuracy.
Currently available languages are: Italian, Castilian and Mexican Spanish, British and American English, German,

French and Brazilian Portuguese. Argentine Spanish and Greek are under development.

ACTOR® is a flexible engine, based on multi-language external knowledge-bases, efficient and platform-independent. It performs text-to-speech conversion as a real-time "software-only" process. The number of channels that can be simultaneously served depends on CPU power and database size of the chosen voice. Different voices are available, consisting of labelled speech signal: the larger the speech database, the higher the voice quality. On average, about 30 channels can be served by a Pentium II 400 MHz, and a speech database requires about 200 Mb disk space, in its 16 KHz Linear PCM coding (tape quality). Less disk-space-demanding supported formats are 11025 Hz Linear PCM and 8 KHz PCM μ-law and A-law (telephone quality).

ACTOR® is available both as a DLL for Windows and as a static library for Solaris. Since the entire system is written in ANSI-C, the ACTOR® library may be virtually portable to any architecture supporting this language, including DSP boards. A set of legacy APIs allows the control of every aspect of the TTS process. The engine is also compliant with Microsoft Speech SDK 4.0 (SAPI). Synthesized speech can be output to a multimedia audio board, a telephone card or a file. The developer may implement his own "custom audio destinations" (such as a LAN, or a legacy audio board) which can be interfaced with ACTOR® library. Speech output is dynamically configurable, i.e. different channels may have different audio destinations.

ACTOR® supports the Voice XML mark-up language (versions 1.0 and 2.0) and accepts both ANSI and UNICODE text formats. Flexibility is one of its relevant features. Voice, language, audio format, user lexicon, etc., can be set run-time, on a per-channel basis. API's and control tags allow to modify speech parameters such as speaking rate, pitch range and volume, and to control reading styles and pronunciation (word-by-word, spelling, dates, phonetic input, etc.).

In order to tailor speech output on the intended application, ACTOR® provides advanced user lexicons with context-grammars and phonetic transcriptions for managing user-exceptions, and exploits the corpus-based technique to allow domain-dependent acoustic add-ons to the base voices.

4. ARCHITECTURE

ACTOR® is conceived as a multilingual system where language-dependent knowledge is kept as far as possible separate from core algorithms. Together with its development tools, it can be viewed as a dynamic system, allowing incremental implementation of new voices and languages. The overall system architecture is shown in Figure 1, where the central block represents the run-time system, with its knowledge-bases and its data flow from input text to output speech, and dashed blocks represent the development environment.

The run-time system is composed of two main modules:

the Text Analyser, converting the input text into a detailed phonetic and prosodic representation;

the Speech Synthesizer, converting the abstract phonetic/prosodic stream into signal samples which are then played by the Audio Destination.

Text Analysis relies on language-dependent knowledge, lexica and rules developed by experts or obtained by machine learning tools. Speech Synthesis obtains its raw material from acoustic dictionaries, each representing a different voice. In turn, acoustic dictionaries are developed in a sophisticated environment, where their contents are first designed so as to provide high phonetic/prosodic coverage of the intended language (or application domain) and then implemented by speech recording and labelling. The Text Analysis module itself is part of such environment, as the first step necessary to compute the statistical distribution of phonetic/prosodic sequences in the intended domain.

The next paragraphs will give a more detailed description of ACTOR®'s run-time modules and development tools.

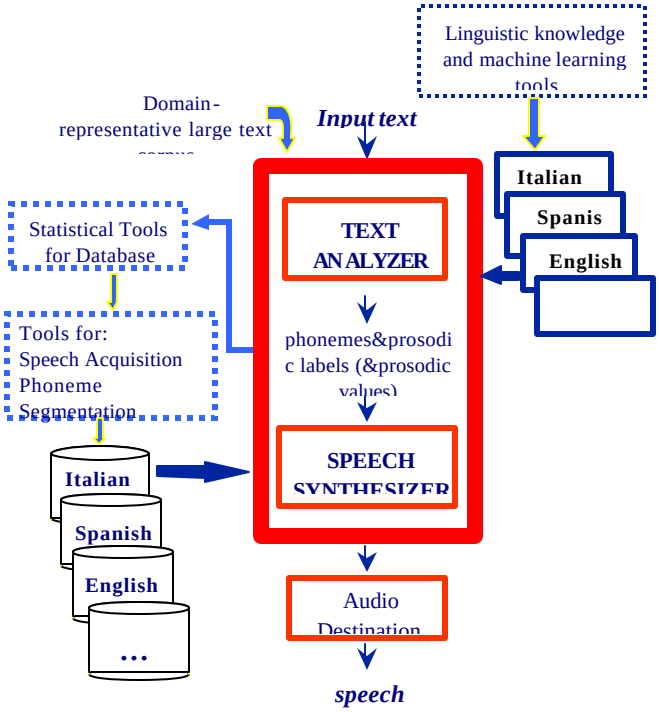

Figure 1. ACTOR® system architecture

5. THE RUN-TIME SYSTEM

The conversion of unrestricted text into speech requires many intermediate steps. Written text is under-specified and the exact sequence of sounds with the proper rhythm and intonation should be first defined by *text-analysis*, before the actual *speech-synthesis* can work. ACTOR®'s text-analysis modules convert the input text into a stream of phonemes, each associated with a prosodic label and with values of

duration and pitch. Such abstract description of the desired speech output is then input to the speech synthesis modules, which select from the acoustic database the best fitting speech portions and combine them into a smooth speech signal.

5.1. Text Analyser

Text Analysis can be viewed as a sequence of tagging and conversion tasks where the input text is interpreted, decoded and transformed into an explicit description of the corresponding spoken message.

5.1.1. Text formatting

Text formatting executes a preliminary surface normalization of the text trying to standardise its form, delimiting sentences and words and converting numbers, abbreviations, acronyms and conventional expressions (internet addresses, mathematical symbols, dates, currencies, etc.) into their fully expanded graphemic form.
These tasks are intertwined (e.g. delimiting sentences requires abbreviation spotting, to tell dots from full stops) and highly language-dependent. Languages do not agree even in the most general conventions concerning character sets and use of punctuation. For instance, in German ordinal numbers are marked with a ".", in Greek semicolons are used as question marks, commas as abbreviation marks and the entire alphabet is different from the Latin alphabet.
ACTOR® performs text-formatting by means of language-dependent procedures and tables. User lexica are available where character strings can be mapped (by context sensitive rules) into their desired expansions. This facility is helpful in tailoring ACTOR® to the application, where the specific domain may require ad- hoc expansions (names, acronyms, codes) and may solve ambiguities (e.g. "Dr.", "drive" or "doctor").

5.1.2. Grammatical tagging and prosodic parsing

Intonation and rhythm, and in some cases also stress and phonetic transcription, depend on the roles of words in the sentence, i.e. on sentence structure. This is why ACTOR® performs some grammatical analysis of the text based on a classification of words according to Parts of Speech (POS). At present, such analysis attains different degrees of accuracy for the different languages.
A minimal solution relies on a lexicon of relevant words (function words and frequent or critical words) enhanced by a set of contextual rules. On these basis, some words can be de-accented and some phonetic problems can be solved (e.g. the *liaison* in French, homographs like "read" in English or "ancora" in Italian, etc.).
A more sophisticated solution attempts a full grammatical tagging of the input text, applying an automatic learning algorithm (Word Classifier) explicitly designed for lexical classification [6]. The algorithm is trained on large lexicons of classified inflected forms and is based on the hypothesis that words with similar graphemic form belong to the same class. The approach has proven effective both in grammatical classification and in stress assignment (see par.5.1.3).

Once words have been assigned their POS, the prosodic structure of the sentence can be guessed. Function words are de-accented and minimal syntactic blocks (e.g. Noun+Verb, Pronoun+Auxiliary+Verb) are identified and gathered into syntactic-prosodic phrases by rhythmical rules [7]. On this basis, breath pauses and intonation contours can be properly generated.

5.1.3. Lexical stress assignment

In some languages the position of lexical stress is fixed (e.g. in French, on the last syllable of the word), but in many others stress can fall on whatever syllable of the word (e.g. in English, compo`nent, co`nfident; in Italian, inautenticita`, a`uguraglielo). In some languages (e.g. Spanish) proper orthography requires to mark stress explicitly, but also in these cases some stress prediction technique should be applied at least in those informal contexts (e.g. e-mail) where orthographic stress is generally left out.
For languages where stress position is variable, ACTOR® applies one of the following two approaches. For American and British English, lexical stress position is predicted together with phonetic transcription, by an automatic learning algorithm (see par.5.1.4). For other languages (Italian, Spanish, Brazilian Portuguese, German) the above mentioned Word Classifier [6, 8] is applied, trained on large lexicons of orthographic words classified according to their lexical stress position.

5.1.4. Grapheme to phoneme conversion

As languages differ in how closely they represent sounds by means of letters, they may require different solutions to map graphemes into phonemes. Although a one-to-one mapping never occurs, there are languages where contextual grapheme-to-phoneme rules can be stated, admitting exceptions which can be explicitly listed. ACTOR® adopts this approach for romance languages and for German. Word-level transcription rules scan graphemes left-to-right and rewrite them into phonemes, according to grapheme and phoneme context, position in the word and morpheme boundaries (some basic affix detection is performed, especially for German). The rule-based approach allows a good transcription control where it is easy to find out mistakes and to correct them. On the other hand, it needs deep knowledge of the language and can be applied only to languages where grapheme-to-phoneme mapping is regular enough.
For more irregular languages, such as English, obtaining the correct phoneme string from written text is not a trivial task, and ACTOR® faces this problem using a language-independent learning scheme based on CART (*Classification and Regression Trees*), trained on large phonetic lexicons of inflected forms [9]. This approach allows to carry out two tasks in a single step: lexical stress assignment and phonetic transcription altogether. One of the advantages of the CART approach is the reduction of the time needed for developing a new language, as it overcomes the difficulty of acquiring a deep insight into the language-specific phonetic features. Of

course, the approach is feasible only if a reliable and large phonetic lexicon is available.

5.1.5. Prosodic target

In order to complete the specification of the target speech, the segmental description must be integrated with suprasegmental features. ACTOR® specifies for each phoneme its desired prosodic realization, at two different levels. As in traditional diphone TTS systems, each phoneme is assigned a duration and a pitch value, according to a language-dependent intonation pattern. In addition, a prosodic label is assigned, marking the phoneme position in the sentence prosodic structure. Such label acts as an abstract summary of prosodic features, where for instance a label "last accent in a declarative sentence" means "longer duration, falling pitch contour". When possible, only this label is considered by the Synthesizer, relying on the fact that concatenating long phoneme sequences taken from the proper sentence location yields plausible prosodic patterns even without signal modification. In case prosody-fitting units can't be found in the database (e.g. a question must be output and the database contains only declarative sentences) the artificially computed values of duration and pitch are imposed on the output signal (see par.5.2.3).

Both labels and prosodic values are computed according to the sentence prosodic structure, taking into account accents, phrase boundaries and sentence modality. To obtain durations a CART is applied [10], while pitch is computed by rule, basing on language-dependent patterns drawn according to an IPO-like intonation model [11, 12].

5.2. Speech Synthesizer

Once the target phonetic/prosodic sequence has been specified, ACTOR®'s speech synthesis modules try to obtain the best speech signal matching the target, relying on the speech database. All the language and voice-specific knowledge is confined in the database, both in its speech content and in its phonetic/prosodic labelling. The synthesis engine performs two main steps: the selection of the longest phoneme sequences fitting the target and their actual concatenation into a smooth speech signal, prosodically modified when necessary.

5.2.1. The acoustic database

Acoustic units are not predefined, rather they are dynamically extracted from a rich database of labeled speech, where naturally pronounced sentences are recorded. No formal constraints are imposed on the database content, except that it should include at least one sample for each phoneme of the language. Of course, the richest the database, the best the output speech.

The recorded speech is carefully analysed and labelled with all the phonetic and prosodic information necessary to select the speech segments which best match the input text. The minimal unit in such analysis is the demi-phone, defined as the signal portion delimited by a phone boundary and a diphone boundary.

Each recorded sentence is divided into the constituent phrases, and then represented in a database structure by using the following features:

- speech filename
- phonetic transcription
- phoneme prosodic label
- demi-phone boundaries indicators (time)
- demi-phone average pitch value (Hz)
- demi-phone pitch markers (time)

5.2.2. The unit selection algorithm

The unit-selection algorithm compares the input phonetic/prosodic description (what must be synthesised and how) with the acoustic database information (what speech material is available), looking for the longest demi-phone sequences in the database matching the input specification. The match is evaluated by means of a multiple parameter score computed for each input demi-phone.

The base criterion takes into account the demi-phone phonetic/prosodic context: the match between the input phoneme sequence and the candidate sequences contained in the database is analysed using a bell-shaped window centred on the focus demi-phone and ending at the first different phoneme. The degree of similarity between the first unmatched phoneme and the corresponding input phoneme is considered as well. All demi-phones in the acoustic database are evaluated according to this parameter and a score is calculated for each of them. Phoneme quality has priority over prosodic type, in the sense that if the required phoneme is not found in the acoustic database no acoustic signal will be given for that phoneme, while if the phoneme exists but with the wrong label, it is taken anyway.

If the design of the database has been accurate enough to ensure statistical coverage of phonetic/prosodic sequences of the language, more than one matching context will be found. In order to select the best one, a further criterion looks for the unit that ensures the smoothest transition with the previously selected one, basing on F0 and duration average values.

A last criterion, called concatenation factor, is then applied to encourage selection of long units and concatenation at easy-to-join boundaries.

The final score of each demi-phone corresponds then to the weighted sum of all the scores given by each criterion.

5.2.3. Concatenation and prosodic adjustment

After the selected unit (demi-phone sequence) has been extracted from the database, it should be combined with the preceding one. If it is well-suited to the target prosodic context and the distance between its F0 and duration values and those of the previous unit are smaller than a minimum threshold, a pure waveform concatenation takes place, without any modification.

Yet, in some cases the selected units do not match the target prosodic labels and their prosody must be changed. Also when residual F0 discontinuities are present at unit junctions the signal obtained by the unit concatenation should be adjusted. ACTOR® performs signal modification by means

of CSELTSEQUENS® a proprietary time-domain pitch synchronous algorithm. In the former case, the target prosodic values (F0 and duration) computed by the phonetic module are imposed on the original signal. In the latter case, only a compensation algorithm (based on a scaling factor) applies: the original shape of the intonation curve of each unit is preserved, and submitted to a percentage frequency scaling which smoothes F0 jumps and realigns out-of-range units.

6. THE DEVELOPMENT TOOLS

As stated above (see par.4 and Fig.1), ACTOR® is an open system designed so as to be easily enhanced with new languages, voices and application tailoring. Its quality greatly depends on its development environment, which is integrated with the system itself and which allows an efficient and reliable creation of knowledge bases.

6.1. Linguistic knowledge acquisition

The development of a new language in ACTOR® requires the acquisition of a rich basis of linguistic knowledge, including lexica and text corpora. Knowledge sources are varied, from newspaper collections to Internet texts, from computer–oriented lexical databases to electronic versions of dictionaries obtained by direct agreement with the publisher. Usually different sources should be integrated, carefully checked and manually enhanced.

A number of tools help in the creation and management of linguistic knowledge bases, allowing evaluation, archiving, format conversion, comparison, etc.

In addition, in order to extract rules from lexica, two different automatic learning environments have been arranged: an implementation of the classical CART technique and a Word Classifier [6]. As above mentioned (see par.5.1), these tools have been applied to grapheme-to-phoneme conversion, POS tagging and lexical stress assignment.

6.2. Acoustic Database design

Together with the unit selection algorithm, the acoustic database is the core of the corpus-based synthesis process. While the algorithm should exploit any speech material at its best, the output speech quality depends on how likely it is to find long and suitable units in the database. This is why great care is devoted to the design of dense texts to be recorded, covering the phonetic and prosodic contexts which are relevant for speech synthesis and most frequent in the intended domain (language/application). In order to obtain a natural and neutral reading by the speaker, texts must consist of meaningful sentences with regular structure. The two requirements, phonetic density and linguistic well-formedness, are often difficult to be kept together. A special semi-automatic environment has been created to help in this task, whose main steps are the following:

- phonetic/prosodic transcription of a representative text corpus by means of ACTOR®'s Text-Analyser
- computation of the statistical distribution of phonetic/prosodic sequences in the corpus
- extraction, by means of a greedy algorithm and of manual check, of a minimal subset of well-formed sentences ensuring the intended coverage.

This process may be applied both in the creation of general-purpose databases for a given language and in application-oriented acoustic specialization.

The size of the resulting database depends on the target quality and on the phonetic structure of the domain. At present, ACTOR®'s general-purpose databases amount to about 80,000 phonemes each, on average.

6.3. Acoustic Database analysis

The recorded speech material is then analysed and acoustic relevant parameters are computed.

An automatic phonetic alignment tool, based on context independent CDHMM (Continuous Density Hidden Markov Model) phone modelling, aligns each speech waveform with its corresponding phonetic transcription [13]. Its average performance is over 95% in speaker dependent mode within a tolerance of 20 ms. For new languages and voices, the phonetic aligner should be re-trained on aligned speech data.

A rule-based diphone segmentation module is then run based on three acoustic parameters (signal energy, spectral variation function and relative phone duration), two conditions ("equal to a value" and "belonging to a range of values") and two logical operators (AND/OR). A rule parser connected to the data structure of the phonetic aligner determines diphone boundaries by processing a set of about 200 acoustic/phonetic rules.

A proprietary pitch-period detection algorithm then applies, based on the search for maxima of a weighted modified autocorrelation function. On voiced portions of speech, pitch estimate errors are detected and corrected by a forward/backward adaptive procedure: pitch markers are assigned in correspondence of the nearest left zero crossings of waveform peaks. Voiced phone/diphone boundaries are aligned to the nearest pitch marker positions, whereas unvoiced phones are labelled with equally spaced intervals.

On the basis of such analyses, the speech database can be segmented into demi-phones, each associated with its duration in ms, signal rms, average F0 in Hz and average first derivative of F0 in Hz/sec (voiced only).

6.4. Diagnostic tools

Crucial requirements in the development of high-quality languages and voices are great accuracy in speech labelling and careful testing of ACTOR®'s output. To correct the errors occurring during the analysis of speech corpora and to debug the TTS functions, diagnostic tools (both automatic and manual) and interfaces have been created for accessing the different modules and knowledge.

In order to prepare the training material for the phonetic aligner and to check its output, a control phase is necessary, which is done in part by hand, and in part automatically. An automatic tool has been developed that detects the incorrect transcription of pauses and the mismatch between the voiced intervals and the phonetic symbols that should represent

them [14]. The manual check of segmentation boundaries is carried-out by acoustic phonetic experts and relevant mistakes are corrected, with the help of a graphic interface that allows signal visualisation and editing.

The diagnostic test of ACTOR®'s output is assisted by an interface allowing to access the intermediate processing steps. It includes the possibility to analyse the results of the text-to-speech conversion at different levels: text normalisation, prosodic analysis, grapheme-to-phoneme conversion, unit selection and unit concatenation. Errors, when occurring, are classified and reported in a document, with their context of occurrence and their severity degree.

For some languages, experiments aiming at evaluating the quality of the system have been carried out. We asked native speakers of the given language to express their opinion about intelligibility, naturalness and pleasantness of the voice.

7. CONCLUSIONS

The overall structure and the base principles of the ACTOR® text-to-speech system have been described. The system belongs to the family of new-generation speech synthesis systems, conceived as multilingual and multivoice engines and based on unit-selection concatenative synthesis. The corpus-based technique, together with an advanced development environment, facilitates the implementation of new languages and voices, as well as the specialization on application domains. Flexibility is an important feature of the system, allowing to switch between voices, input interfaces, user lexica and specialized acoustic databases. The adopted synthesis technique provides unequalled speech output quality. Both timbre and prosody are mostly kept unaltered, resulting in a human sounding voice effect. As a drawback, preserving timbre naturalness reduces prosodic control over the synthesized speech. A challenge for the future will be to improve the synthesis technique so as to allow signal manipulation with no loss in acoustic quality, for changing prosodic styles according to application needs.

8. REFERENCES

[1] R. Billi, F. Canavesio, A. Ciaramella and L. Nebbia: "Interactive Voice Technology at Work: the CSELT Experience", *Proceedings of IVTTA '94*, Kyoto, September 1994.

[2] W.N. Campbell and A.W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis", in: J. van Santen et al. (eds.), *Progress in Speech Synthesis*, pp. 279-292, Springer New York, 1996.

[3] M. Balestri, S. Lazzaretto, P.L. Salza and S. Sandri, "The CSELT System for Italian Text-to-Speech Synthesis". *Proceedings of EUROSPEECH '93*, Berlin, Vol. 3, pp. 2091-2094.

[4] L. Nebbia, S. Quazza and P.L. Salza, "A Specialized Speech Synthesis Technique for Application to Automatic Reverse Directory Service", *Proceedings of IVTTA '98*, Torino, September 1998, pp. 223-228.

[5] E. Zovato, S. Sandri, "Two feature to check phonetic transcriptions in Text To Speech Systems", *Proceedings of EUROSPEECH 2001*, Aalborg, Vol. 3, pp. 2243-2246.

[6] S.Quazza, H.Van den Heuvel "Lexica in Text-to-Speech Systems", in *Lexicon development for speech and Language Processing*, F. Van Eynde and D. Gibbon eds., Kluwer Academic Publishers, Dordrecht, 2000

[7] B. Gili Fivela, S. Quazza, "Text-to-Prosody Parsing in an Italian Synthesizer. Recent Improvements", *Proceedings of EUROSPEECH '97*, Rhodes, Greece, September 1997, Vol. 2, pp. 987-990

[8] M. Balestri, "A coded dictionary for stress assignment rules in Italian", *Proceedings of EUROSPEECH '91*, Genova, September 1991, Vol. 3, pp. 1169-1171.

[9] F. Mana, P. Massimino, A. Pacchiotti, "Using Machine Learning Techniques for Grapheme to Phoneme Transcription", Vol. 3, pages 1915-1918, *Proceedings of EUROSPEECH 2001*, Aalborg, September 2001, Vol. 3, pp. 1915-1918

[10] F. Mana, S. Quazza, "*Text-To-Speech Oriented Automatic Learning of Italian Prosody*", *Proceedings of EUROSPEECH*, Madrid, 1995.

[11] 't Hart, J. et al, "A Perceptual Study of Intonation", Cambridge 1990

[12] S. Quazza, P.L. Salza, S. Sandri and A. Spini, "Prosodic Control in a Text-to-Speech System for Italian", *Proc. ESCA Workshop on Prosody*, Lund, 1993, pp. 78-81.

[13] B. Angelini, C. Barolo, D. Falavigna, M. Omologo and S. Sandri, "Automatic Diphone Extraction for an Italian Text-to-Speech Synthesis System", *Proceedings of EUROSPEECH '97*, Rhodes, Vol. 2, pp. 581-584.

[14] M. Balestri, A. Pacchiotti, S. Quazza, P. L. Salza and S. Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system", *Proceedings of EUROSPEECH '99*, Budapest, Vol. 5, pp. 2291-2294.