

Festival Speaks Italian!

*Piero Cosi**, *Fabio Tesser***, *Roberto Gretter*** and *Cinzia Avesani**
with Introduction by *Mike Macon****

*IFD-CNR

Istituto di Fonetica e Dialettologia – Consiglio Nazionale delle Ricerche
e-mail: fcosi.avesani@csrf.pd.cnr.it - www: <http://www.csrf.pd.cnr.it/>

**ITC-IRST

Istituto Trentino di Cultura - Istituto per la Ricerca Scientifica e Tecnologica
e-mail: fgretter.tesser@irst.itc.it - www: <http://www.itc.it/enIRST/>

***OGI-CSLU

Oregon Graduate Institute for Science and Technology
Center for Spoken Language Understanding
e-mail: macon@ece.ogi.edu - www: <http://www.ece.ogi.edu/~macon/>

Abstract

Finally Festival speaks Italian. In this work, the development of the first Italian version of the Festival TTS system is described. One male and one female voice for three different speech engines are considered: the Festival-specific residual LPC synthesizer, the OGI residual LPC Plug-In for Festival and the MBROLA synthesizer. The new Italian voices will be freely available for download for non-commercial purposes together with specific software modules at <http://nts.csrf.pd.cnr.it/IFD/Pages/Italian-TTS.htm>.

This paper is devotedly dedicated to the memory of Mike Macon, whose recent passing on was really a shock to all of his friends.

1. Introduction

This introduction is extracted from Mike Macon's "Overview - CSLU Speech Generation Projects" [1].

Speech generation is generally accomplished by one of the following three methods:

- **General-purpose concatenative synthesis.**

The system translates incoming text onto phoneme labels, stress and emphasis tags, and phrase break tags. This information is used to compute a target prosodic pattern (i.e., phoneme durations and pitch contour). Finally, signal processing methods retrieve acoustic units (fragments of speech corresponding to short phoneme sequences such as diphones) from a stored inventory, modify the units so that they match the target prosody, and glue and smooth (concatenate) them together to form an output utterance.

- **Corpus-based synthesis.**

Similar to general-purpose concatenative synthesis, except that the inventory consists of a large corpus of labeled speech, and that, instead of modifying the stored speech to match the target prosody, the corpus is searched for speech phoneme sequences whose prosodic patterns match the target prosody.

- **Phrase-splicing.**

Stored prompts, sentence frames, and stored items used in the slots of these frames, are glued together.

The strengths and weaknesses of these methods are complementary. As for speech **quality** and **scope**, general-purpose concatenative synthesis is able to handle any input sentence but generally produces mediocre quality. Corpus based synthesis can produce very high quality, but only if its speech corpus contains the right phoneme sequences with the right prosody for a given input sentence. If the corpus contains the right phonemes but with the wrong prosody, the end result may locally (i.e., within the range of a phoneme sequence that was available in the corpus) sound quite good, but the utterance as a whole may have a bizarre sing-song quality with confusing accelerations and decelerations. And, obviously, phrase splicing methods produce completely natural speech, but can only say the pre-stored phrases or combinations of sentence frames and slot items; naturalness can be a problem if the slot items are not carefully matched to the sentence frames in terms of prosody.

An additional issue to consider is the **amount of work** required to build a system. The cost of generating a corpus or an acoustic unit inventory is significant, because besides making the speech recordings, each recording has to be analyzed microscopically by hand to determine phoneme boundaries, phoneme labels, and other tags. Such time consuming analysis is not necessary for phrase splicing methods. On the other hand, applications involving names may be prohibitive for phrase splicing methods (In the US, there are 1.5 million distinct last names!).

A final consideration is **size**. Although the prices of memory and disk space are continually dropping, being able to have more channels on a given hardware platform translates directly into increased profits, and there is also an increasing interest in using speech synthesis on handheld devices. Thus, size still matters. Concatenative synthesis has the edge on size. Moreover, its quality limitations are less of a problem because the acoustic capabilities of handheld devices are themselves limited.

In other words, each of these methods has problems with quality, scope, the amount of resources required, or size.

2. Speech Synthesis Engines

In this work on Italian we restrict our field of interest to

general-purpose concatenative synthesisⁱ. In particular, three different synthesis engines have been taken into consideration: the Festival residual LPC synthesizer [2], the OGI residual-LPC synthesizer [3], and the Mbrola synthesizer [4].

2.1. Festival residual LPC synthesizer

Festival is a general multi-lingual speech synthesis system developed at the CSTR (Center for Speech Technology Research, Edinburgh, Scotland, UK) [5] offering a full text to speech system with various APIs (Application Program Interfaces), as well an environment for development and research of speech synthesis techniques. It is written in C++ with a Scheme-based command interpreter [6] for general control. Festival is a *general-purpose concatenative* text-to-speech (TTS) system that uses the residual-LPC synthesis technique, and is able to transcribe unrestricted text to speech. Festival is a multi-lingual system suitable for research, development and general use. It is freely available for research and educational use. CSTR has recently expanded its range of languages: there are now available TTS systems for English, Spanish and Welsh and, finally.....**Festival speaks Italian!**

2.2. OGI residual-LPC synthesizer (Festival plug-in)

This OGI residual-LPC synthesizer, which has to be considered as a plug-in for Festival, has been developed at OGI (Oregon Graduate Institute of Science and Technology, Portland, OR) [7], and provides a new signal processing engine, and new voices, not included in the Festival distribution. Specifically, new pitchmark and LPC analysis algorithms, together with some scheme scripts that enable the creation of new voices in the OGIresLPC synthesizer are included. It is freely available for research and educational use. OGI has expanded its range of languages: there are TTS systems for English, Spanish and Welsh and, finally.....**OGI residual-LPC speaks Italian!**

2.3. Mbrola

Mbrola is a speech synthesizer based on the concatenation of diphones, coded as speech samples with 16 bits (linear), developed at the Faculté Polytechnique de Mons, TCTS Lab. in Mons, Belgium [8]. It takes a list of phonemes as input, together with prosodic information, and produces speech output. Therefore it is NOT a Text-To-Speech (TTS) synthesizer, since it does not accept raw text as input. Brazilian Portuguese, Breton, British English, Dutch, French, German, Romanian, Spanish, Swedish are already available as full software multilingual speech synthesizers (i.e., the DSP part of a TTS system), and finally**MBROLA speaks Italian!**

3. Italian TTS development

As for Italian, we decided to concentrate first on the general-purpose concatenative synthesis, in order to speed up the process of creating new voices in Italian and to verify the quality of an Italian TTS created with this technique. In particular we concentrated on diphones. As a general rule, a diphone is considered as the portion of speech relative to the sequence of two phonemes starting from the middle of the first phoneme and ending at the middle of the following one. Due to specific constraints imposed by the Festival's

architecture, geminated consonants and affricates are exceptions to the general definition of "diphone".

3.1. Italian Diphone Database

A recording of a new Italian synthesis database with a male speaker (P.C.) has been executed at IFD, while a similar recording with a female speaker (L.P.) has been executed at ITC-IRST. The laryngograph signal (electro-glottal graph - EGG) has been recorded too for better pitch extraction. The speaker reads a set of carefully designed nonsense or true Italian words embedded in syntactically-correct but semantically-incorrect phonetically rich sentences which have been constructed to elicit particular phonetic effects. This technique, facilitating the monotone pronunciation of the sentence, ensures that the collected database only contains the required phonetic variability. Various scripts for automatic segmentation, diphone extraction and LPC analysis have been developed to speed up the creation of a new voice. The database has been formatted in the Festival residual-LPC, OGI residual-LPC Plug-In for Festival and Mbrola formats for a total of 3 male and 2 female new voices:

pc_diphone, pc_OGI_diphone, pc_mbrola and lp_diphone, lp_ogi_diphone.

As illustrated in Figure 1, a Sennheiser MKH 40 P48 microphone, connected with a PC and with the right channel of a DAT Sony DTC 1000 ES, and an Elettroglottograph connected to the left channel of the DAT have been used for recording.

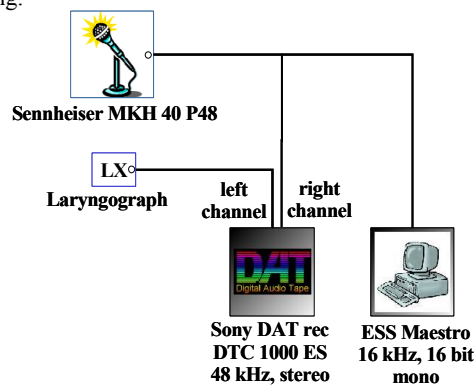


Figure 1. Recording Environment.

The recording session has been completely automated by a RAD script (Cslu Speech Toolkit [9]) enabling the direct acquisition of each sentence as a Windows PCM *.wav (16kHz, 16bits) file, opportunely named, onto computer disks. "Target diphones" are embedded in "target words" within the nonsense sentences as indicated in the following examples:

- "i venti **cia**pàpa **zap**àpa sono stanchi"
- "i gatti **ò**papa **à**papa sono belli"
- "il **tung**steno papap**ò** papòcc**ia** è ghiacciato"

For each sentence there are two main target words, but various consonantal clusters or vocalic sequences were also embedded in the carrier sentences in order to insure the possibility to use in the future more complex units of different length. In order to reasonably normalize the diphone intensity a within-word normalization has been executed on all speech waveforms. At present almost 1300 diphones have been created (CV, CC, VV...) but around 2000 will probably be the final number of diphones, when consonantal clusters or vocalic sequences will be included (clusters, rV, VVV,...).

Diphones have been coded in terms of 16 LPC coefficients together with their corresponding residual signal. Both the Festival Edinburgh Speech Tools and the OGIresLPC plug-In for Festival packages have been considered for LPC analysis; coding is pitch-synchronous in both cases. Diphones were also coded in terms of speech waveform (PCM linearly coded, 16kHz-16bits) and successively transformed in Mbrola format. LPC coefficients were computed synchronously with pitch for each voiced portion of speech, while for the unvoiced portions, pitch marks are placed at regularly spaced intervals (10ms). Diphone segmentation was executed automatically by adapting a “high-performance” Italian phonetic general-purpose speech recognition system developed at IFD [10] and trained on APASCI corpus [11] using the CSLU Speech Toolkit [12].

3.2. Natural Language Processing Text/Linguistic Analysis

A true general Natural Language Processing module (NLP) for TTS purposes [13] should obviously be capable of producing a correct phonetic and prosodic transcription of input text. Accurate phonetic transcription can only be achieved if a morphological analysis module is available, which enables the identification of all possible Part-Of-Speech (POS) categories for each word on the basis of their spelling. A contextual analysis module is needed to disambiguate among multiple POS categorizations, and a syntactic-prosodic parser is needed to find the text structure that more closely relates to its expected prosodic realization. Semantic and pragmatic information should be included too, but since very few data are currently available to consider that, TTS systems merely concentrate based on syntax to achieve natural prosody. A pre-processing module, to organize the input text sentences into manageable lists of words should always be present. Numbers, abbreviations, etc. should be identified and transformed into full text when needed. Finally, a Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech, concludes the TTS chain.

At the time of this writing, as described in Figure 2, various modules have already been created for Italian, but a true NLP system is still to be completed. A first module implements a simple grammar to convert strings of numbers into word sequences. Numbers are expanded at the word and phoneme level distinguishing among time, dates, telephone numbers, etc.. Non-numerical data are divided into “function-words” and “content-words”. In the prosodic modules, “function words” are treated differently from lexical words; once identified, they are divided by their grammatical group and by their specific sub-functional group (e.g.: definite ARTICLES: il, lo, la, i, gli, le; indefinite ARTICLES: un, uno, una; time ADVERBS: ieri, oggi, dopo, poi, ecc.). At this point all the words are phonemically transcribed by the use of a wide lexicon, compiled in Festival format to speed up search procedures; if they are not present in the lexicon, they are phonemically transcribed by the use of explicit stress-assignment, letter-to-sound and syllabification rules. The Lexicon, compiled in Festival format, comprises 500k stressed Italian forms, SAMPA [14] phonemically transcribed, divided in syllables and labeled after their grammatical class (POS) such as:

("accertare" V ((a tS) 0) ((tS e r) 0) ((t a l) 1) ((r e) 0)).

All the rule-modules are written in Scheme [6], a Festival specific language, and, at present, they are all statistically

designed on the basis of a wide text corpus. No morphological analysis is yet implemented. By convention, all mono-or disyllabic function words are considered as unstressed and form a prosodic unit with the following word.. SAMPA symbolism was adopted, excluding the symbolic representation for stressed vowels, which are followed by symbol “1”.

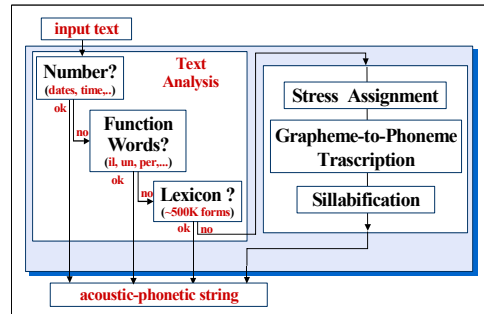


Figure 2. Text/Linguistic analysis for Italian.

3.3. Prosodic Analysis

As illustrated in Figure 3, starting from the phonemic string obtained as output of the Text/Linguistic modules, the correct diphones are extracted from the acoustic database and, for each unit, specific information relative to its mean duration and pitch is included. These data are successively sent to the real speech synthesis engine to generate the speech waveform: in particular, the Festival residual LPC, the OGI residual LPC Plug-In for Festival and the Mbrola engines have been considered for Italian. Prosodic analysis is for the moment the weak part of the system and will be strongly updated and improved in the future. Up to now, prosodic modules are too simple and rely essentially on punctuation marks and function words. Each phoneme is assigned a mean duration, which was statistically computed by analyzing a wide corpus of Italian sentences produced by various Italian television announcers RAI [15]. The duration of stressed vowels is augmented by 20% relative to the average vowel duration.

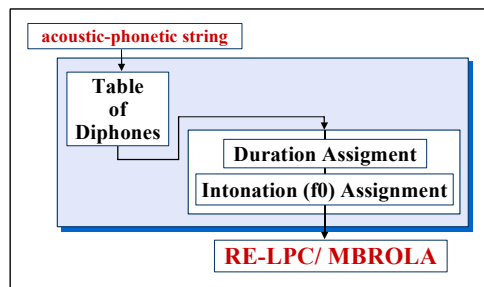


Figure 3. Prosodic Analysis for Italian TTS followed by the synthesis engine module (RE_LPC and MBROLA).

Pauses between words are divided in 2 categories: short pauses of 250ms, associated with some punctuation marks such as [' \ , ;] and long pauses of 750ms associated with main conclusive punctuation marks such as [? . : !]. As for intonation, declarative sentences (see Figure 5a) are segmented in intonational phrases each of which is assigned a baseline starting at 140Hz and ending at 60Hz (for a typical male voice). For any stressed syllable, the f0 contour is raised by approximately 10Hz over the baseline, while the last

syllable has a steeper inclination relative to the baseline. A resetting of the baseline is executed on the function words.

As for interrogative sentences, (see Fig. 5b), a falling-raising pattern is associated with the last word. A “Target Point” (TP) is assigned to the last stressed vowel, and is aligned at 3/4 of its duration: at that point the f_0 curve reaches a value corresponding to 80% of the baseline, falling from a value equal to the baseline assigned to the end of the preceding vowel. Starting from TP, f_0 raises up to f_{0max} with an inclination that spans over the post-tonic unstressed syllables. The last syllable is assigned a faster speed.

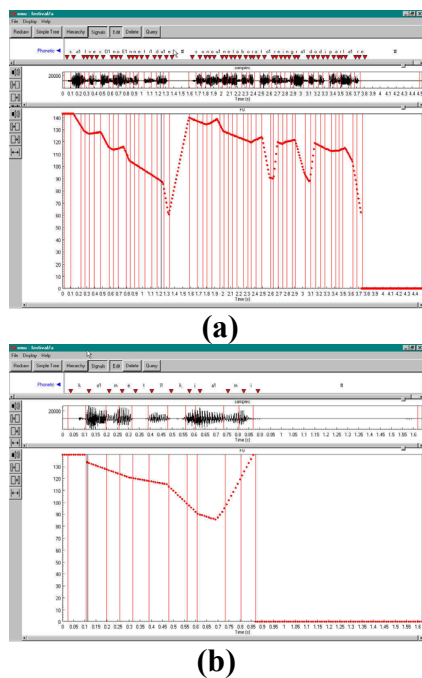


Figure 5. f_0 contour for two declarative sentences (a): “Salve, sono NT2. Sono un elaboratore in grado di parlare” (Hello, I am NT2. I am a computer able to speak); and for one interrogative sentence (b): “Come ti chiami?” (What’s your name?), synthesized with a male voice.

4. Conclusions

Finally FESTIVAL speaks Italian with three different synthesis engines: the Festival residual LPC synthesizer, the OGI residual-LPC synthesizer, and the Mbrola synthesizer.

The development of new voices was speeded up by the creation of various Scheme and Tcl/Tk [16] procedures and scripts and, in particular, those for the automatic segmentation of diphones by the use of a very high performance Italian phonetic recognition system [10], were quite effective and efficient. A new speaker, in fact, can record all the needed speech sentences in almost 2 weeks and, by the use of all the scripts and procedures, a new voice can be prepared in two days, even if a manual checking of the diphone corpus, especially for segmentation, is always necessary as a final step to refine small alignment errors and to improve the final performance of the system.

5. Acknowledgments

We would like to thank our friend Mike Macon who prematurely died in March this year, for his invaluable help especially in having made the OGILresLPC Plug-In for Festival

speaking in Italian. Among others, we would like also to thank: Alan Black, from Speech Group at Carnegie Mellon University, Thierry Dutoit, from Faculté Polytechnique de Mons, Baris Bozkurt, from MULTITEL-TCTS Lab, Paul Taylor, from Center for Speech Technology Research at University of Edinburgh, and John Paul Hosom from OGI, for having make this work possible.

6. References

- [1] Mike Macon, “Overview – CSLU Speech Generation Projects, www: <http://cslu.cse.ogi.edu/research.htm>
- [2] **FESTIVAL**: Alan W. Black (awb@cs.cmu.edu), Paul Taylor (Paul.Taylor@ed.ac.uk), Richard Caley, Rob Clark (robert@cstr.ed.ac.uk) - CSTR - Centre for Speech Technology - University of Edinburgh. www: <http://www.cstr.ed.ac.uk/projects/festival/>.
- [3] **OGILresLPC PlugIn for Festival**. www: <http://cslu.cse.ogi.edu/tts/download/index.html>.
- [4] **MBROLA**: The MBROLA Project. www: <http://tcts.fpms.ac.be/synthesis/>.
- [5] P.Taylor, A.W.Black & R.Caley, “The Architecture of the Festival Speech Synthesis”, *ESCA 98 Workshop*,
- [6] **SCHEME**, Computer Programming Language. www: <http://www-swiss.ai.mit.edu/~jaffer/Scheme.html>
- [7] M.Macon, A.Cronk and J.Wouters and A.Kain, “OGILresLPC: Diphone synthesiser using residual-excited linear prediction”, *num. CSE-97-007*, Department of Computer Science, OGI, Portland, OR, Sep, 1997.
- [8] T.Dutoit & H.Leich, “MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database”, *Speech Communication*, Elsevier Publisher, November 1993, vol. 13, n°3-4.
- [9] **CSLU-SPEECH-TOOLKIT**: www: <http://cslu.cse.ogi.edu/tools.htm>.
- [10] P. Cosi e J.P. Hosom, “High Performance “General Purpose” Phonetic Recognition for Italian”, *Proceedings of International Conference on Spoken Language Proc. (ICSLP-2000)*, Beijing, Cina, 16-20 October, 2000, Vol. II, pp. 527-530.
- [11] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter e M. Omologo, “A Baseline of a Speaker Independent Continuous Speech Recognizer of Italian”, *Proc. of EUROSPEECH 93*, Berlin, Germany, 1993.
- [12] M. Fauty, J. Pochmara e R.A. Cole, “An Interactive Environment for Speech Recognition Research”, *Proceedings of International Conference on Spoken Language Proc. (ICSLP-92)*, Banff, Alberta, October 1992, pp. 1543-1546.
- [13] T. Dutoit, *An Introduction to Text-To-Speech Synthesis*, Kluwer Academic Publishers, pp. 1996, 326.
- [14] A.J. Fourcin, G. Harland, W. Barry e V. Hazan, Eds., *Speech Input and Output Assessment, Multilingual Methods and Standards*, Ellis Horwood Books in Information Technology, 1989.
- [15] M. Federico, D. Giordani, and P. Coletti, “Development and evaluation of an Italian broadcast news corpus”, *Proc. 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.
- [16] **Tcl/Tk**: K. Ousterhout - ouster@sprite.berkeley.edu. WWW page: <http://sol.brunel.ac.uk/tcl/Tcl.html>.

ⁱ Part of this work has been sponsored by the European Project, IST-1999-10982, MPIRO: Multilingual Personalized Information Objects.