# From Marked Text to Mixed Speech and Sound

*Paolo Massimino*

Loquendo S.p.A.
Vocal Technology and Services

Paolo.Massimino@loquendo.com

**ABSTRACT**

Loquendo TTS is a commercial Multilanguage/Multivoice Text-To-Speech synthesizer, attaining great acoustic naturalness and linguistic accuracy. Currently available languages are: Catalan, Chinese, Dutch, British English, American English, French, German, Greek, Italian, Portuguese, Brazilian, Castilian, Argentine, Chilean, Mexican and Swedish. Loquendo TTS 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 Loquendo TTS integrated *audio mixer* allows mixing sound files and synthetic voice. It's possible to mix one or more sound files simultaneously, at the same time. Thanks to explicit tags embedded in the text, easy synchronization between audio files and speech is guaranteed even if the text is modified. Every sound effect is treated as an independent track, with independent timeline, volume and sample rate. Commands such Mix, Play, Stop, Pause, Resume, Loop and Fade allow users to have complete control on the audio sources.

In order to make easy the use of the integrated audio mixer, a multi-platform application is shipped with Loquendo TTS SDK: TTSDirector. Loquendo TTS Director is a Java multi-platform development tool intended for helping the user in the design of the application prompts. The text of the application prompt can be written in the edit box and interactively refined by means of a "listen & edit" procedure, allowing to tune the TTS behavior by means of the Loquendo TTS User Control Tags.

This paper gives details about the previous topics and can be used as basis for the workshop demonstration, concerning the use of the audio mixer, integrated into the Loquendo TTS, and other functionalities.

## 1. INTRODUCTION

Up to today, the TTS (Text-To-Speech) and the Audio Mixer have been separated functionalities, and separated application software. Usually, starting from a text, one ore more synthetic speech files are generated (concerning one or more voices), and in a second step, a separate audio mixer permits the integration of the synthetic speech with music or other signal files, adding various effects and controls if necessary. The Loquendo **TTS integrated *audio mixer*** allows mixing sound files and synthetic voice. It's possible to mix one or more sound files **simultaneously,** at the same time. Every sound file (audio source) is considered as an independent audio track, with independent volume, timeline and sample rate. The sample rate frequency of the audio sources is automatically converted according to the voice frequency used. In order to make easy the use of the integrated audio mixer, a multi-platform application is shipped with Loquendo TTS SDK: **TTSDirector**. With this application, the user can control various parameters and characteristics of the Text-To-Speech production, using a friendly menu interface, including all the audio mixer capabilities.

## 2. LOQUENDO TTS OVERVIEW

Loquendo TTS is a commercial Multilanguage/Multivoice Text-To-Speech synthesizer, attaining great acoustic naturalness and linguistic accuracy. Currently available languages are: Catalan, Chinese, Dutch, British English, American English, French, German, Greek, Italian, Portuguese, Brazilian, Castilian, Argentine, Chilean, Mexican and Swedish. Loquendo TTS 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 labeled speech signal: the larger the speech database, the higher the voice quality. On average, about 140 channels can be served by a Pentium IV 2 GHz, 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 8 KHz PCM μ-law and A-law (telephone quality). Also a compressed audio format is available, supporting several bit rates for optimum disk space versus audio quality optimizations, whenever needed. Loquendo TTS is available both as a .DLL (.SO) for Windows, Linux and Solaris, and as a static library. Since the entire system is written in ANSI-C, the Loquendo TTS 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 (SAPI) 4.0 and 5.0. 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 Loquendo TTS library. Speech output is dynamically configurable, i.e. different channels may have different audio destinations. Loquendo TTS 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 at 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, Loquendo TTS 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. Loquendo TTS 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.

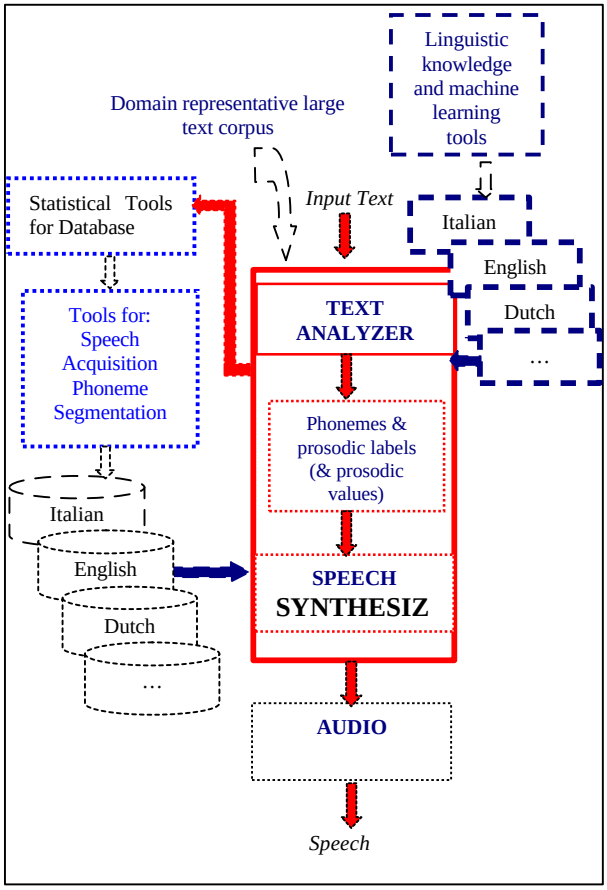


Figure 1. The Loquendo TTS system architecture

The run-time system is composed of two main modules: the Text Analyzer, 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, lexical and rules developed by experts or obtained by machine learning tools [1]. 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 labeling. 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.

## 3. LOQUENDO TTS MIXER CAPABILITIES

A recent feature of Loquendo TTS (starting from the release 6.3, dated mid of 2004) is the audio mixer. The *audio mixer* is integrated inside the TTS; in this way, using simple command tags embedded in the text, users can reproduce audio files and music synchronized with the words pronounced by synthetic voices. Commands such Mix, Play, Stop, Pause, Resume, Loop and Fade allow users to have complete control on the audio sources. Thanks to explicit tags among the words, easy synchronization between audio files and speech is guaranteed even if the text is modified. Every sound effect is treated as an independent track, with independent timeline, volume and sample rate. No more offline audio re-sampling is required: the sample rate frequency of the audio sources is automatically converted according to the voice frequency used. As part of a synthetic speech system, the Loquendo Mixer Audio module allows the realization of extremely high quality, multilingual speech applications.
The audio mixer is initialised at the first occurrence of a "\audio" or "\audio(…)" tag. A set of commands is available inside the "\audio(…)" tag, in order to select between several functionalities:

- \audio(play=<filename>): play of a signal file at the specified position in the text.
- \audio(mix=<filename>): play of a signal file at the specified position in the text, but synthetic speech and <filename> audio will be mixed together.
- \audio(name=<track name>): set a mnemonic name to the current track.
- \audio(volume=<range(0-200)>): set the volume of the current audio track.
- \audio(pause[=filename]): pause the current audio track.
- \audio(resume[=filename]): resume the current audio track.
- \audio(pauseall): pause all the audio tracks.
- \audio(resumeall): resume all the paused audio tracks.
- \audio(stop[=filename]): stop the last audio track.
- \audio(stopall): stop all the audio tracks.
- \audio(path=<path>): specify a common path where the audio files are stored.
- \audio(track=<filename.wav>): specify which track is considered as the current track.
- \audio(mix2play[=filename]): switches the current track from mix mode to play mode.
- \audio(fadein=<msec>): set a 'fade in' effect for the current track.
- \audio(fadeout=<msec>): set a 'fade out' effect for the current track.
- \audio(recstart=<track name>) + \audio(recstop): record speech that can be used in another part of the text.
- \audio(close): close the mixer.

This large set of command permits to build sophisticated audio prompts. For instance, in the following example, a background

of music will be played together with the speech of the text, thanks to the “command tag”; at the same time, the volume is set to a selected value (70%):

```
\audio(mix=c:/Segnali/intro2.wav;volume=70)
Loquendo presents the new audio mixer
capability.
```

The following example is similar to the previous one, but the *fade-out* eliminate the sharp stop of the audio at the end of the playing:

```
\audio(mix=c:/Segnali/intro2.wav;volume=70)
Loquendo presents the new audio mixer
capability.
\audio(fadeout=1000;mix2play) .
```

It is possible to control the audio stream using the *pause* and *resume* commands:

```
\audio(mix=c:/Segnali/intro.wav;volume=35)
You can easily control the audio stream.
Example: you may use  \audio(pause) the pause
tag  to pause the music.
And the \audio(resume) resume tag if you wish to
start it again.
\audio(fadeout=4000;mix2play)
```

The *audio mixer* supports 16 bit sound files, mono and stereo, with arbitrary sample rate frequency.
“.wav” files are supported and played. “.mp3”, “.wma”, “.asf”, “.ogg”, “.avi”, “.mpg” will be supported in future versions. “.raw” , “.pcm” and any other extension files are played as raw files.

## 4. LOQUENDO TTS DIRECTOR

Loquendo TTS Director is a Java multi-platform development tool intended for helping the user in the design of his application prompts. In Figure 2 is shown a snapshot of the TTSDirector graphical interface.

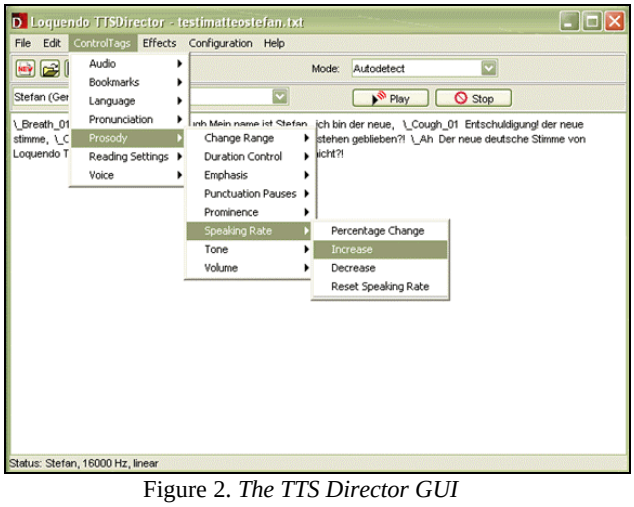


Figure 2. *The TTS Director GUI*

The text of the application prompt can be written in the edit box and interactively refined by means of a "listen & edit" procedure, allowing to tune the TTS behavior by means of the Loquendo TTS User Control Tags. A detailed menu helps choosing the proper tags. The tuned prompt can be saved as a text or as an audio file. The allowed encoding for the input text are (Western European) ISO Latin 1, that is ISO-8859-1, and UNICODE UTF8 and UTF16. TTSDirector needs the Java Runtime Environment (JRE) version 1.4.2 (at least). Two combos allow selecting, respectively, the default **TTS voice** (that may be changed via control tags in the texts) and the **Mode** (Multi-line, Paragraph, SSML). In a similar way, font type and font dimension can be changed by means of other two combos.

The buttons **Play** and **Stop** allow synthesizing the edited text with Loquendo TTS.

The **File** menu allows opening and saving the edited prompts, both in text and audio formats.

The **Edit** menu allows Cut & Paste in the edit window (also available via left mouse button).

The **ControlTags** menu provides a structured access to the available Loquendo TTS Control Tags. The Tags are grouped according to their categories so that it is easy to choose the intended one. The categories are: voice, language, prosody, pronunciation, spelling and pausing, **audio mixer**, and the TTSDirector interactive menu enables the right choice to be made easily. The selected control is automatically inserted in the edit box, at the caret position. The selected control is automatically inserted in the edit box, at the caret position (the “caret” is a flashing line, block, or bitmap in the client area of a window or in a control that accepts keyboard input). It indicates the place at which text or graphics are inserted. In case the control needs further specification by the user, a wizard is presented, or in general, this is marked by a yellow text in the edit box, asking for the needed details. E.g.:

```
\voice=<insert a valid voice name>
```

The **Effects** menu is a guide to the advanced features of "expressive cues" and "plugin lexicons". In case the selected voice is provided with such special add-ons, this menu allows selecting the desired effect. The repertoire of *Expressive Cues* consists of a set of pre-recorded formulas, comprising conventional figures of speech, like greetings and exclamations ("hello!", "oh no!", 'I'm sorry!"), interjections ("Oh!", "Well!", "Hum"..) and paralinguistic events (e.g. breath, cough, laughter, etc.), which suggest expressive intention (to confirm, doubt, exclaim, thank, etc.). The use of such formulas can make vocal messages lifelike and expressive. The Effects menu allows selecting the proper formulas among those available for the active voice. The linguistic formulas are listed in the ***SpeechActs*** submenu, according to intuitive linguistic categories. The paralinguistic events are accessible from the ***Extras*** submenu. The selected expression is directly inserted in the edit box. Every “SpeechAct” or “Extra” is played when the mouse pointer pass on the loudspeaker icon, in order to have a faster select of the proper Expressive Cue.

The **Plugin** submenu allows activating/deactivating the plugin lexicons available for the current voice. The selected plugin lexicon (see the relative paragraph in this Guide) is activated on the edited text from the caret position onward, until explicit de-activation.

The **Configuration** menu allows setting some acoustic and prosodic parameters for the Loquendo TTS voices: sampling frequency and coding, pitch, speaking rate and volume.

More edit instances (panes with a tab) can be opened and saved in a single TTSDirector session, in order to build and test several voice prompts at the same time. The “New” button or the “CTRL-t” key can be used to switch between the instances. Separate Cut-Copy-Paste popup menus are available for every instance, and can be activated a click of the right button of the mouse in the editor area. A similar click of the right button on the editor’s tab activate a Save-Save as-Close popup menu, and can be used to save the data present in the relative editor instance.

## 5. CONCLUSIONS

Combining speech and sound, in order to exploit them into auditory user interfaces, is a complex task. Usually, the integration of speech and sound is performed after the Text-To-Speech phase, but this approach has several problems, especially in the synchronization and automation of the voice prompts (by means of scripts). An innovative solution, realized in Loquendo TTS, is the integration, inside the Text-To-Speech, of the audio mixer. In this way, standard TTS markup is enhanced by means of new tags concerning the audio mixer, and it is easy to maintain the timing between speech and sound events. Moreover, a tool (Loquendo TTSDirector) has been developed, which gives a friendly access to all the speech and audio mixer capabilities, and let the user to build, and save, new and complex voice prompts in a few steps. By means of the integration of the audio mixer capabilities with the set of Text-To-Speech control tags, and their embedded utilization in the text with the help of TTSDirector, the mixed audio files and synthetic speech prompts production is by far easier to create, check and change!

## 6. REFERENCES

[1] 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

[2] 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.

[3] P.L. Salza, “Phonetic transcription rules for text-tospeech synthesis of Italian”. Phonetica, 47, pp.66-83, 1990.

[4] M. Balestri, “A coded dictionary for stress assignment rules in Italian”. Proc. EUROSPEECH ’91, Genova, 1991.

[5] Sproat R., “Corpus-Base Methods and Hand-Built Methods”, Proceedings of ICSLP, Beying, 2000.