

Advanced Speech Synthesis: Emotion and Affect

SpeechTEK|2006
The Voice Solutions Showcase

Monica Bisacca, Loquendo
August 7, 2006



Empower

Why is Loquendo working on expressive TTS?

1. It is universally considered to be the new frontier in speech synthesis research.
2. To overcome one drawback of TTS: its dull and impersonal flavor.
3. The market expects commercial TTS systems to provide emotional speech and lifelike voices.
4. ...it's fun!

Research on Prosody

- This involves three different areas:
 - a. **Signal Processing**: algorithms for altering the acoustic prosodic parameters of the speech signal without losing speech quality.
 - b. **Prosody Modeling**: acoustic/perceptual analysis of prosodic parameters, extracting their typical patterns corresponding to different linguistic/expressive functions.
 - c. **Text Analysis**: finding textual cues to prosody, finding out what is the expressive intention of a text.

Intention and Emotion

- A realistic ambition for TTS is to enhance its effectiveness in communication by **adapting its speaking style to the intention** of the message:
 - to welcome the listener, I would choose a lively, friendly tone.
 - to alert people in case of danger, I would use a forceful tone.

Such prosodic patterns follow certain conventions and are explicitly adopted by the speaker. **An inventory of such patterns can be considered as an extension** of the set of patterns for ‘normal’, ‘linguistic’ prosody.

- On the other hand, the involuntary effects of **emotions** on speech – e.g. panting for panic, sobbing for sadness, rising pitch for excitement - are much more difficult to study and reproduce; **they are highly subjective**, depending on human psychology and physiology, and **sometimes contradictory**.

Two approaches to Expressive TTS

1. To design a **general method of assigning a given expressive intention** to a text, independently of its content.

It is an ongoing and challenging task, involving research on signal processing, speech acoustics and human communication.

2. **Enriching synthetic messages with expressive phrases and sounds**, which convey expressive intentions and enhance the emotional color.

It's a commercially available solution: Loquendo released Expressive TTS in October 2004.

First Approach: Synthesizing Speech with Emotions

1. Data Collection: an **emotional-speech DB** is recorded in three emotional styles.
2. Data Analysis: **rules extrapolated** from comparison with a prosodically neutral style.
3. Synthesis Parameters: **target intonation** (pitch profiles), **speech rate** (duration) and **intensity** (energy) constraints have been obtained by applying the rules extrapolated from the analysis stage.
4. Emotional Speech Synthesis: **prosody** **'transplantation' techniques** have been applied to the output of the corpus-based speech synthesis system.

Emotional Speech DB Composition

- 4 styles: **Neutral, Angry, Happy** and **Sad**
- 25 phonetically balanced sentences for each style.
- Sentences were composed of 10-15 words.
- Sentences had no emotional content.
- The speaker had to simulate each emotional style, while a director was present during the recording sessions to control pronunciation and prosody and to avoid over emphatic performances

Text example: *The competitor has made twenty five offers, closing only five contracts"*

Data Analysis

A syllable was chosen as the reference acoustic unit. In order to extract prosodic features, the analysis of this corpus was organized into three main steps. For each utterance, the analysis consisted of:

1. Syllable labeling and alignment
2. Fundamental frequency extraction
3. Energy calculation

Syllables were classified into 4 categories depending on their position in the sentence and whether they were stressed or not (lexical stress):

1. FS the first stressed syllable of the sentence or after a speech pause
2. S stressed syllable
3. LS the last stressed syllable of the sentence or before a speech pause
4. U unstressed syllable

Data Analysis (cont.)

For each sentence of the database, parameters were extracted both at utterance and syllable level:

Utterance parameters:

- Maximum pitch
- Minimum pitch
- Maximum energy

Syllable parameters:

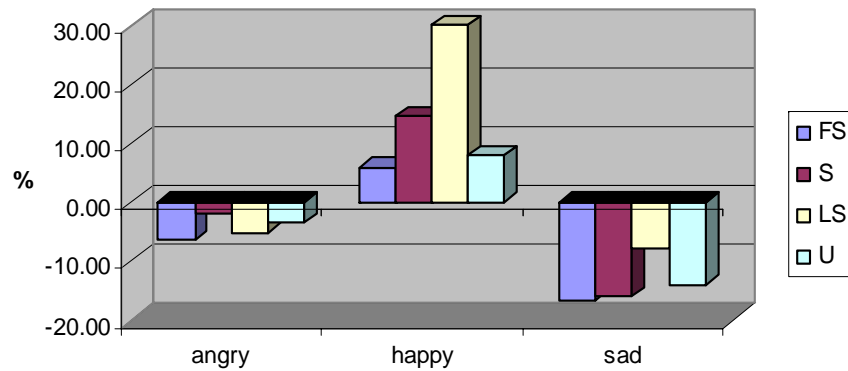
- Duration
- Pitch range
- Mean pitch
- Energy

Data were analyzed and classified in order to trace, for each speaker and for each emotional style, the average variation of these parameters with respect to the neutral style parameters.

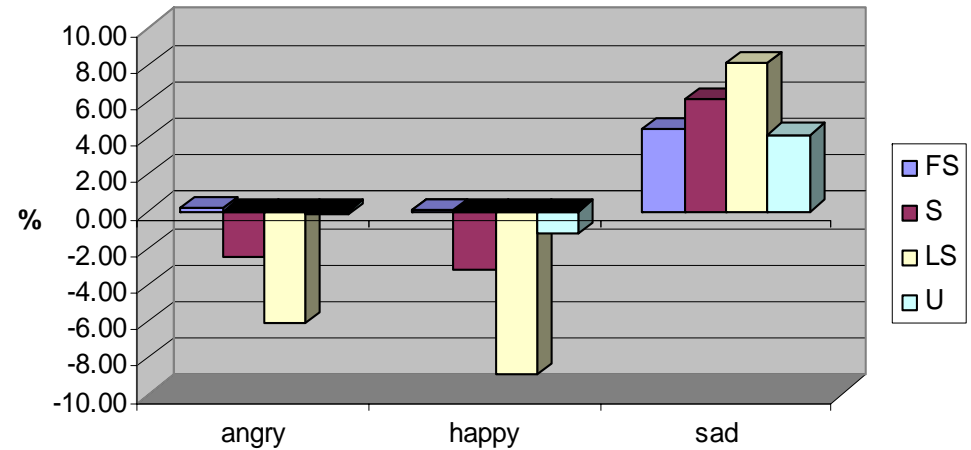
$$c_{<style>}^i = \frac{p_{<style>}^i}{p_{neutral}^i}$$

Syllable Variation Coefficients

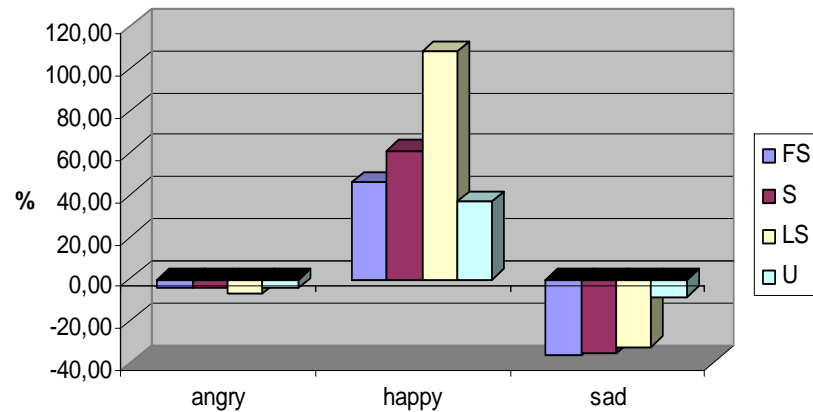
Mean F0



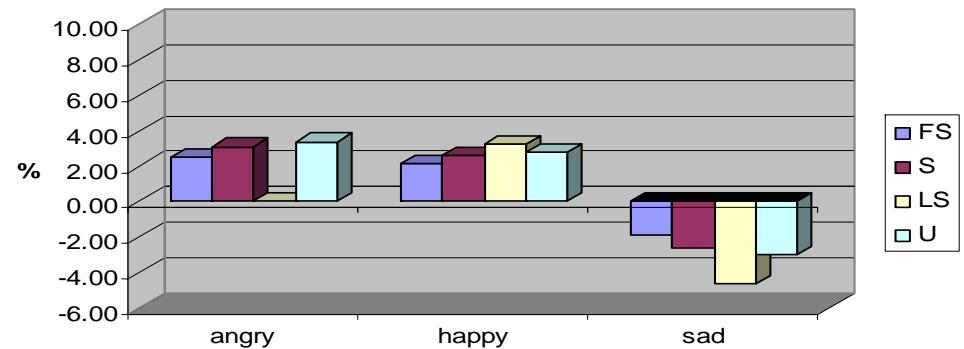
Syllable duration



F0 Range



RMS Energy



Synthesis target parameters

Pitch

Emotional style target pitch values are set according to the variation coefficients except in case of pre-pause patterns that are modified more gently.

Duration

Syllable target durations are recalculated on the basis of the corresponding variation coefficients.

Different scaling factors are then applied to speech pauses, setting them longer for sad styles and shorter for more passionate styles.

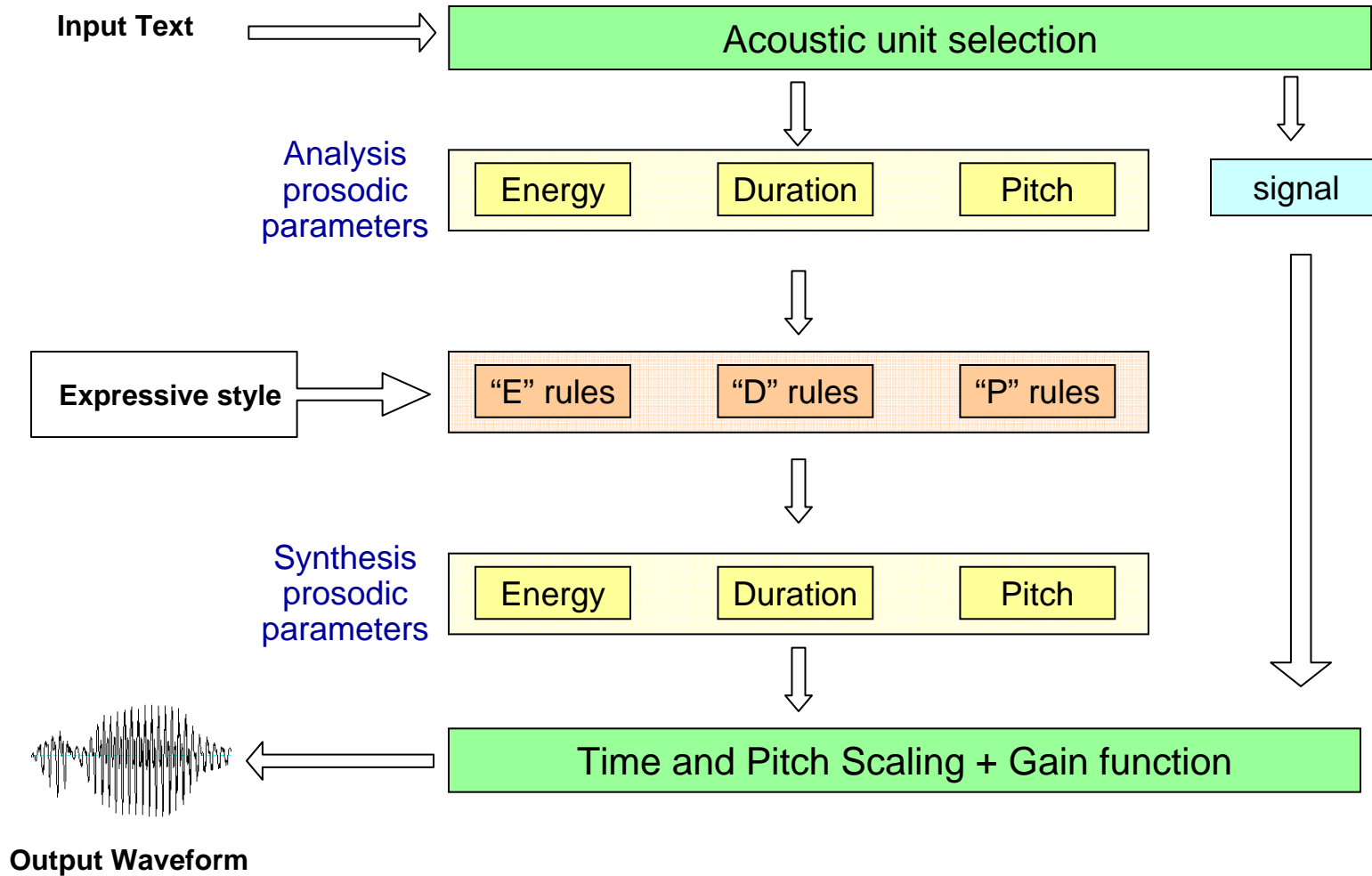
Energy

For each syllable, an energy target value is calculated according to the model parameters.

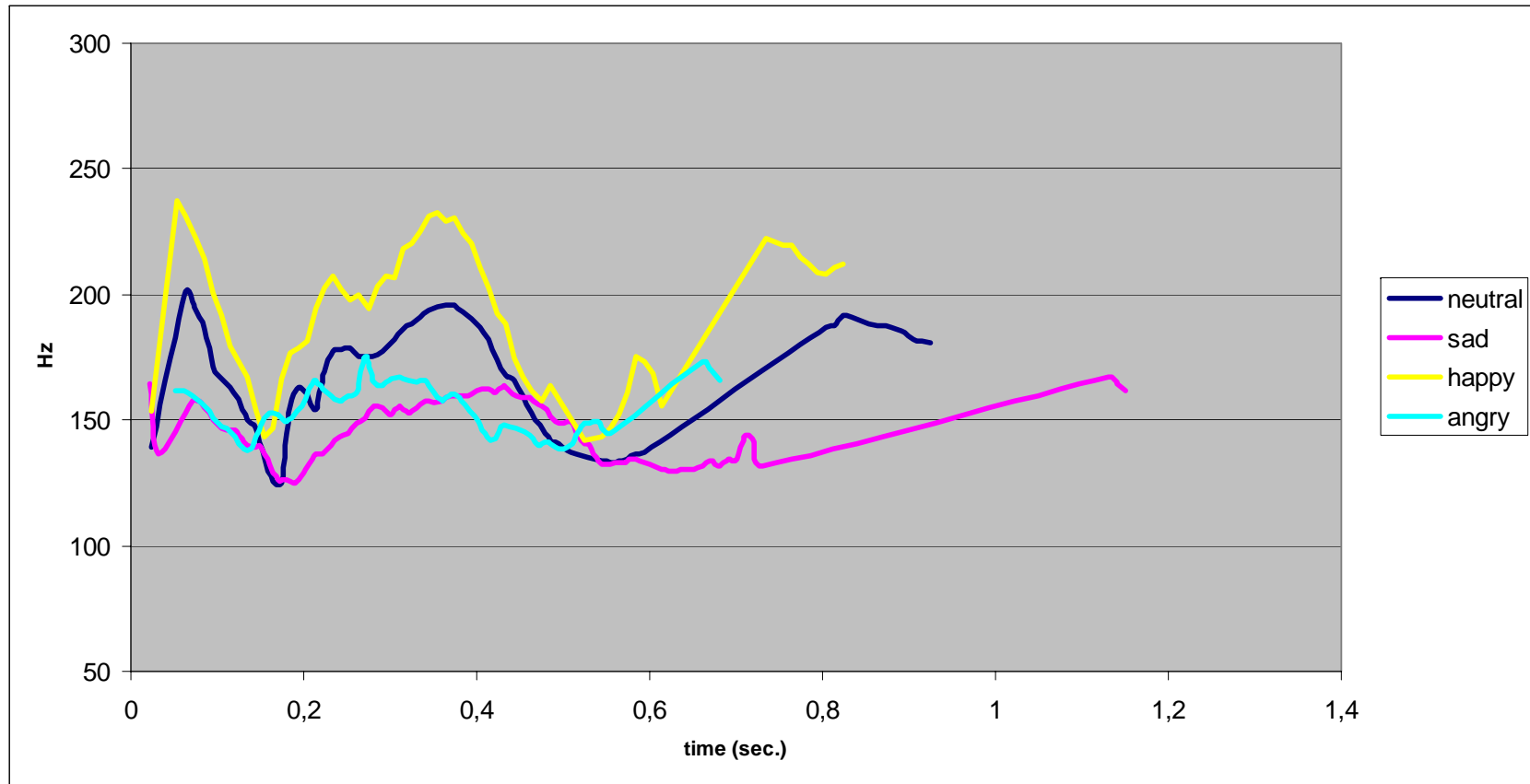
Signal manipulation

- A time domain Pitch Synchronous Overlap and Add technique is used to modify waveforms according to the new pitch and acoustic unit length values.
- An adaptive gain function is used to set the new energy values.
- In some cases, in synthesizing emotional speech samples, we had to tune the scaling coefficients to avoid disagreeable distortions. In particular, high activation styles have shown some critical aspects.

Synthesis scheme



Example: pitch contour (Susan)



Susan neutral



Susan sad



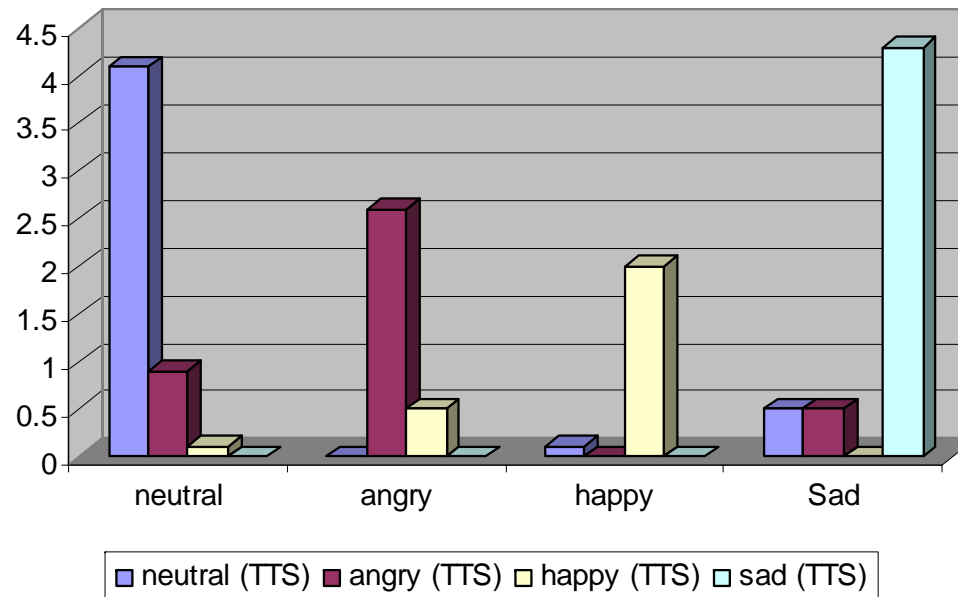
Susan happy



Susan angry

Evaluation results

- Texts without emotional content were used to synthesize three emotional samples plus the neutral one.
- Volunteers were asked to listen to the four samples in random order and to evaluate how much *sad*, *angry*, *happy* or *neutral* each stimuli sounded.
- They were allowed to listen to the sentences more than once if necessary.
- Rating range was from 0 to 5.



Considerations

- The described framework has proven **computationally efficient**, since most of the necessary information is stored in the TTS database.
- Results of perceptual tests show that **styles are well recognized** even if the **acoustical quality, in some cases, degrades** (perceivable distortions are produced).
- Actually, to improve this prototype:
 - spectral parameters should be considered in describing **articulatory alterations**
 - the prosody modification algorithm should be improved to **avoid distortions or artefacts**.

Second Approach: Enriching TTS with Expressive Cues

- This solution provided an **immediate response** to the market's requirement for expressivity in speech
- Not yet a fully emotional style for any phrase, but an enhancing repertoire of **Expressive Cues** consisting of a set of pre-recorded formulas (greetings, exclamations, paralinguistic events) which suggest **expressive intention** (to confirm, doubt, exclaim, thank, etc.).
- Enriched TTS continues to benefit from the **high quality and natural timbre** achieved with the Unit Selection technique.

Good afternoon ladies and gentlemen! Listen to this! _Throat
I am the American synthetic voice from Loquendo.
It gives me great pleasure to be here with you all. See you later! _Laugh





Loquendo Expressive TTS

- Loquendo TTS provides a **repertoire of "expressive cues"** allowing emotional pronunciation, creating extremely natural sounding speech.
- For every synthetic voice, **linguistic formulas are recorded in a natural and expressive way**, in such a way as to be compatible with neutral synthetic phrases.
- It is important to organize carefully the corpus of expressive phrases based on the characteristics of every language and, for each speech act, to **cover the most frequently used expressions**.

Figures of Speech

■ These are:

➤ **SpeechActs**, divided into intuitive linguistic categories, such as:

Announcements	(dear all! dear customer! ladies and gentlemen! ...)	
Apologies	(excuse me! I'm so sorry! oh my gosh I'm sorry! ...)	
Compliments	(good! congratulations! that's great news! ...)	
Disapproval	(I don't agree! I completely disagree! what a rip off! ..)	
Greetings	(all the best! bye bye! hello and welcome! ...)	
Refusals	(absolutely not! definitely not! I can't! ...)	
Surprise	(who would have believed it! surprise! ...)	
Thanks	(thank you! I'm very grateful! thanks for everything! ...)	

.....

➤ **Non-linguistic interjections**

(Oh, Aha, Er, Hmm, Doh, Oops...)




➤ **Paralinguistic events**

(e.g. coughing, laughter, breathing, etc.)



Methodology issues

- The same expression can be pronounced with **different layers of expressivity**, from neutral to emphatic, from sad to amazed. 
- **Paralinguistic events** are recorded in several versions, so avoiding the unnatural repetition of identical sounds
- The recorded material is inserted into the database with tags to identify different prosodic levels
- The user distinguishes between many different emotional styles by the simple use of punctuation marks (“!” “!!” “??” “?!”)

Conclusion

- **Synthesizing Speech with Emotions:**
 - Currently still work in progress
 - This is the preferred solution once the limits of the signal transformation algorithms have been resolved, and loss of acoustic quality overcome.
- **Enriching TTS with Expressive Cues:**
 - These linguistic formulas express the true meaning and character of a phrase
 - Without this added expressivity, everyday phrases such as “I’m so sorry!” have no credibility
 - Maintains quality of the Unit Selection Technique
 - Market has responded very positively
 - Truly lifelike TTS means there’s no need for costly, time consuming pre-recording, and enables a rapid deployment of vocal services that customers will love using.

Loquendo Today

- **Global company** formed in 2001 as a spin-off from the Telecom Italia R&D center with over **30 years experience in Speech Technologies**
- Complete set of **Multilingual speech technologies** on a wide range of devices
- **Full support of international standards** (VoiceXML, MRCP, VoIP)
- Ready for challenging future scenarios: **Multimodality, Security**
- Partnership as a key factor
- **Strong and growing presence** in Europe, North and Latin America
- HQ in Turin, Offices in US, Spain, Germany and France, and Worldwide Network of Agents
- “Best Innovation in Multi-Lingual Speech Synthesis” Prize AVIOS-SpeechTEK West 2005.

“Best Innovation in Multi-Lingual
Speech Synthesis” Prize
AVIOS-SpeechTEK West 2005.



“Best Innovation
in Speech Synthesis” Prize
AVIOS-SpeechTEK West
2006.



Empowering Your customers and employees with speech technologies

SpeechTEK | 2006
The Voice Solutions Showcase

Empower

Thank You

Loquendo
VOCAL TECHNOLOGY AND SERVICES