SINTESI DELLA VOCE E AGENTI PARLANTI

Piero Cosi

ISTC-SFD - (ex IFD) CNR
Istituto di Scienze e Tecnologie della Cognizione
Sezione di Fonetica e Dialettologia
Consiglio Nazionale delle Ricerche
e-mail: cosi@csrf.pd.cnr.it
www: http://www.csrf.pd.cnr.it/

INTRODUZIONE

Chi non ricorda la voce di HAL 9000, il computer di bordo della navicella spaziale Discovery, "protagonista" del famoso film di Stanley Kubrik "2001 Odissea nello Spazio". Era il 1968 e, a distanza di 35 anni, si può forse dire che le previsioni contenute in un film di fantascienza, almeno per quanto riguarda la qualità della voce di un sintetizzatore vocale,si sono avverate.

La qualità dei migliori sintetizzatori vocali da testo scritto (*TTS, Text to Speech Synthesis*), intesi come quei sistemi in grado di pronunciare con una voce naturale artificialmente rigenerata, un qualsiasi testo attualmente disponibili, non solo sul mercato, ma anche nei laboratori di ricerca più avanzati, è sicuramente paragonabile a quella di HAL.

Si deve ricordare però che non tutti i problemi sono stati risolti. Confrontando, infatti, la capacità emotivo/espressiva di un attuale sintetizzatore vocale con quella di HAL ci si rende immediatamente conto del *gap* ancora non risolto e che ancora qualche anno dovrà passare prima di poter ottenere una sintesi affidabile anche da questo punto di vista.

Possiamo però senz'altro dire che la qualità della sintesi della voce ha assunto un livello tale da poter ormai essere utilizzata in moltissime applicazioni.

La lettura di messaggi, memorizzati in sistemi di posta vocale, e-mail e fax all'interno di una mail box unificata, accessibile attraverso standard e-mail-clients, tramite web o telefono, la lettura di pagine web, gli avvisi di particolari emergenze, i servizi clienti delle aziende telefoniche ad esempio per l'inoltro delle chiamate, la consultazione interattiva ed amichevole di fonti di informazione elettroniche, gli ausili di lettura per i portatori di disabilità visive e molti altri ancora sono solo alcuni dei possibili esempi di applicazione attualmente sperimentati con successo.

Bisogna inoltre ricordare che parlando di sistemi di sintesi o TTS. negli ultimi anni, non si considerano soltanto quei sistemi in grado di sintetizzare file audio, ma anche video, come ad esempio nella realizzazione delle cosiddette "Talking Heads" o agenti parlanti, sistemi in grado di simulare virtualmente una persona umana che parla.

UN PO' DI STORIA

Gli studi passati hanno portato alla conoscenza fondamentale della dinamica alla base della generazione della voce umana. Quando parliamo, un suono base, prodotto dal flusso d'aria generato dai polmoni e passante attraverso le corde vocali, viene modulato dalla cavità orale, dal naso e dalla bocca ed è la posizione delle diverse parti della lingua e la posizione delle labbra che sono responsabili dei diversi suoni componenti il segnale verbale, ed è questo meccanismo complesso che si deve essere imitato per realizzare sistemi di sintesi vocale. Possiamo senz'altro affermare che la storia di questa tecnologia ha inizio nel 1939 presso i Bell Laboratories dove venne presentato per la prima volta il VODER (Voice Operating DEmonstratoR), VODER era una sorta di strumento musicale dove una barra vibrante generava le frequenze fondamentali, variabili attraverso un meccanismo a pedale ed il suono prodotto veniva modulato utilizzando dei filtri acustici, controllati con le mani. La qualità della voce era ovviamente molto scadente ma un simile meccanismo era la prova della realizzabilità di una voce sintetica. A metà del ventesimo secolo negli Stati Uniti nei laboratori Haskins è stato poi presentato il Pattern Playback, uno strumento ottico/elettronico capace di sintetizzare suoni vocali a partire da una loro rappresentazione acustica. Vi sono anche esempi di sintesi vocale sin dall'antichità. Ad esempio, nel 1779, a San Pietroburgo, il professore russo Kratzenstein costruì dei risuonatori acustici capaci di produrre i suoni delle cinque vocali.

Tornando a tempi più recenti, la vera svolta nel campo della sintesi vocale fu l'arrivo della tecnologia digitale che, associata agli enormi progressi nello studio del meccanismo di produzione della voce, rese possibile intorno agli anni settanta, la realizzazione dei primi sistemi per la sintesi della voce da testo scritto.

SUI DIVERSI SISTEMI DI SINTESI VOCALE

Esistono molte strategie fra loro differenti per sintetizzare il parlato, ma in termini generali si dividono essenzialmente in due grandi categorie denominate ***system-models*** **e** ***signal-models***.

- ***Modello del Sistema di Produzione (system-model)***
  (***sintesi articolatoria***)

Il segnale acustico è il risultato della modellizzazione e simulazione del meccanismo fisico di produzione del suono. Questo approccio è anche conosciuto come *sintesi articolatoria*. La sintesi articolatoria si prefigge di generare il segnale vocale mediante una corretta modellizzazione dell'apparato orale umano. Questo metodo di sintesi utilizza in pratica dei modelli computazionali biomeccanici per la riproduzione del parlato simulando il comportamento degli articolatori interessati nella fonazione e le corde vocali. I modelli degli articolatori sono guidati nel tempo in modo da riprodurre le configurazioni caratteristiche di ogni fonema utilizzando delle regole che riflettono i vincoli dinamici imposti dalle articolazioni. Per generare il segnale vocale, la forma del condotto orale, definita dalla posizione degli articolatori, viene convertita in una funzione di trasferimento, che utilizza come ingresso un segnale di eccitazione generato tramite un modello delle corde vocali. Il problema è quindi ricondotto alla determinazione dei punti di articolazione caratteristici di ogni fonema e delle transizioni tra fonemi. Per determinarli sono spesso utilizzati dati presi da radiografie o risonanze magnetiche dinamiche. Nonostante il suo notevole valore scientifico questo tipo di sintesi non ha ricevuto grande attenzione, a causa della scarsa competitività in termini di qualità con altri sistemi di sintesi e della elevata complessità indispensabile per ottenere buoni risultati in termini di naturalezza.

- ***Modello del Segnale (signal-model)***

Con tale approccio si vuole rappresentare il suono che arriva al nostro apparato uditivo, senza fare un esplicito riferimento al meccanismo articolatorio che genera il suono stesso, ma esclusivamente al meccanismo fisico/acustico responsabile della produzione della voce intesa come onda sonora di pressione. In questo approccio sono rappresentati a loro volta i metodi di ***sintesi per formanti*** e di ***sintesi per concatenazione***.

- ***Sintesi per formanti***

Anche se si basa su una elaborazione del segnale che viene prodotto dall'apparato fono-articolatorio, la sintesi per formanti, in realtà non ignora del tutto il meccanismo di fonazione umana. Infatti si basa sulla teoria sorgente-filtro della fonazione, assai ben descritto ad esempio da Gunnar Fant nel suo famoso libro "Speech Sounds and Features" (1973, MIT Press). La sintesi per formanti utilizza, infatti, un modello del condotto vocale realizzato mediante un filtro composto da un numero limitato di risonanze (con 4 si riesce ad esempio ottenere una voce di buona qualità), con frequenza, ampiezza, e banda di risonanza variabili. Modelli più elaborati utilizzano ulteriori risonanze e antirisonanze per i suoni nasali, con associato anche del rumore ad alta frequenza utile ad esempio nella simulazione delle consonanti fricative e occlusive. Il segnale d'ingresso è sempre generato tramite un modello più o meno approssimato delle corde vocali. Una sorgente molto stilizzata ma funzionante consiste in un treno di impulsi per i tratti di voce vocalizzati o in un rumore bianco per le parti non vocalizzate (si veda Figura 1). Questo modello utilizza delle notevoli semplificazioni rispetto alla realtà. Ad esempio il presupposto che la sorgente di eccitazione sia completamente indipendente dal filtro è assolutamente improbabile.

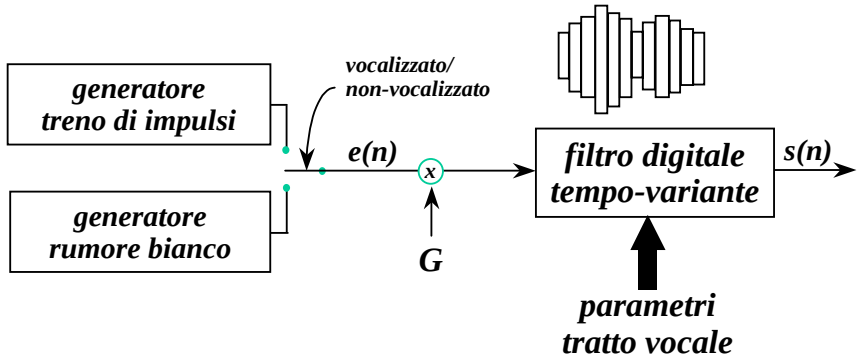


Figura 1. Modello di sintesi filtro-sorgente

Con questo metodo, al fine di sintetizzare una frase, per ogni fonema e per ogni sua transizione, bisogna determinare i parametri di controllo dei filtri e della sorgente di eccitazione variabili nel tempo. Questo tipo di sintesi genera un parlato altamente intelleggibile ma non completamente naturale; presenta, comunque, il vantaggio di una bassa richiesta di risorse di memoria e di calcolo.

- ***sintesi per concatenazione***

Questo tipo di sintesi unisce, modificandoli con appropriati algoritmi, piccoli frammenti (*unità elementari*) di segnale vocale, al fine di sintetizzare un intera frase. Questi metodi evitano le difficoltà di simulare l'atto di fonazione umana mediante specifici modelli, tuttavia introducono atri problemi, quali ad esempio la difficoltà di concatenazione omogenea delle unità acustiche registrate in diversi contesti e la modifica della prosodia intesa come variazione di intonazione e durata. A tal fine vengono utilizzate specifiche tecniche di elaborazione del segnale (*signal processing)*: fra le più comuni ricordiamo quelle denominate ***Predizione Lineare*** e ***PSOLA***, la prima basata sulla teoria del modello sorgente-filtro precedentemente introdotta, la seconda esclusivamente su tecniche di elaborazione del segnale, al di fuori quindi di un modello del fenomeno di produzione della voce.
Per questa modalità di sintesi possiamo fare un'ulteriore suddivisione delle strategie in base alle unità fondamentali utilizzate per la concatenazione. Si possono infatti distinguere la sintesi per ***difoni*** (generalmente definiti come la porzione del segnale vocale che và da metà di un fonema alla metà del fonema[1] successivo), ***trifoni***, ***metà-sillabe***, ecc. fino ad arrivare all'estensione di ***unità variabili*** utilizzate nei sistemi di sintesi più recenti che utilizzano algoritmi denominati "***Automatic Unit Selection***".Questo tipo di sintesi dunque concatena le unità selezionate da un database vocale e, dopo una decodifica opzionale, invia in uscita il segnale vocale risultante. Poiché i sistemi di questo tipo usano frammenti di un discorso registrato risultano più naturali.

SISTEMI DI SINTESI PER CONCATENAZIONE

Questi sistemi consentono di ottenere una sintesi da testo di assoluta generalità, combinando frammenti di voce molto piccoli. Le unità elementari sicuramente più utilizzate sono i difoni, precedentemente introdotti. Per consentire la sintesi, sono necessari i difoni corrispondenti a tutte le coppie di fonemi di una determinata lingua. Generalmente i sistemi di questo tipo utilizzano poco più di un migliaio di difoni, ricavati da parole, in genere sequenze di sillabe senza significato, registrate da un parlatore umano con intonazione monotona. Queste unità vengono poi concatenate per formare le frasi desiderate su cui agiscono sofisticati algoritmi in grado di variarne la durata e la frequenza fondamentale in modo da ottenere i valori più adatti al testo.

Un diagramma a blocchi di un tipico sistema di sintesi per concatenazione, che, nella sua parte di analisi testuale può considerarsi comune a tutti gli altri metodi, è illustrato in Figura 2.

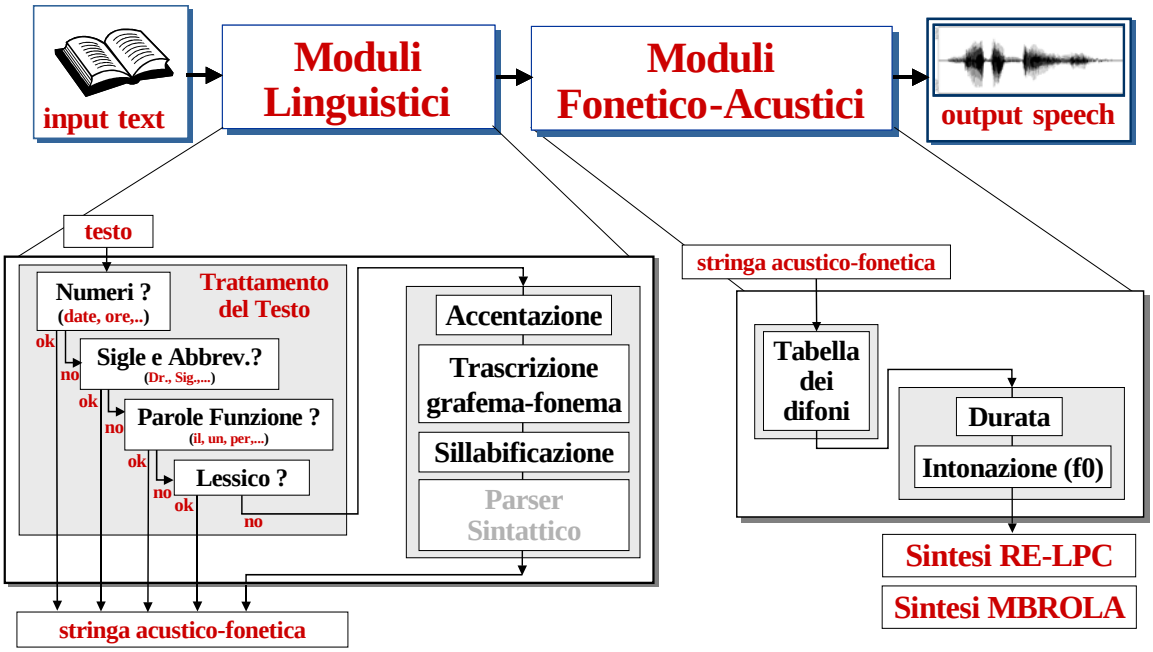


Figura 2. Architettura generale di un sistema di sintesi da testo scritto per concatenazione di difoni.

[1] Con il termine fonema si intende il più piccolo suono che compone la parola come ad esempio una vocale /a/ o una consonante /s/.

Il primo blocco (*Moduli Linguistici*), comune essenzialmente a tutte le tipologie dei sistemi di sintesi, è il modulo di analisi che acquisisce il messaggio testuale in ASCII e lo converte in una serie di simboli fonetici e targets metrici (frequenza fondamentale, durata, ampiezza). Tale modulo consiste di una serie di '*sotto-moduli*' con funzioni distinte, ma in molti casi collegate: il testo di input è prima analizzato e i numeri, i simboli non alfabetici e le abbreviazioni sono espanse in parole (per esempio l'abbreviazione 'Ing.' è trascritta come 'Ingegnere', '222' come 'duecentoventidue'). Tutte le altre parole, se non sono parole-funzione o parole già presenti in un lessico di riferimento, vengono accentate, trascritte foneticamente, sillabificate e opzionalmente analizzate mediante un parser sintattico che, riconoscendo parte del discorso per ogni parola nella frase, è utilizzato per 'etichettare' il testo. Il suo compito inoltre è togliere l'ambiguità a parti costituenti la frase per generare una corretta stringa di suoni, ad esempio per disambiguare parole scritte allo stesso modo ma con significato o accento diverso (ad esempio: "viàggio" vs "viaggiò").

Il secondo blocco (*Moduli Fonetico-Acustici*) assembla le unità in base alla lista di targets ed è principalmente responsabile della qualità acustica e della naturalezza della sintesi. Le unità selezionate sono infine inserite in un sintetizzatore in grado di generare le forme d'onda del segnale vocale.

In particolare l'analisi prosodico-intonativa, intesa come determinazione della durata e dell'intonazione (frequenza fondamentale) in corrispondenza delle unità da sintetizzare, è sicuramente la parte più importante di ogni sistema di sintesi ed viene elaborata, o per regole esplicite, caratterizzate e studiate in dettaglio per ogni lingua, oppure mediante un approccio statistico (ad esempio basato su CART, *Classification and Regression Trees*) in grado di apprendere da un corpus di esempi le caratteristiche prosodiche di una determinata lingua. Entrambe i metodi sono ovviamente ottimizzati a diversi livelli principalmente a seconda della bontà dell'analisi testuale, fonetica e sintattica precedentemente illustrata.

SISTEMI DI SINTESI A SELEZIONE DI UNITA' ("*UNIT SELECTION*")

Negli ultimi anni si stanno imponendo i sistemi *corpus-based* o *unit selection* la cui caratteristica fondamentale è quella di non aver bisogno di limitare il numero, e la dimensione dei frammenti da concatenare. Questo tipo di sintesi è caratterizzato dalla memorizzazione, la selezione e la concatenazione di segmenti di discorso di dimensioni variabile. Questi segmenti vengono estratti, mediante specifici algoritmi basati su tecniche statistiche, da grandi corpora di materiale vocale pre-registrato, naturale e fluente.

Questa strategia di sintesi mira, non più a modificare gli attributi metrici, come durata del suono o frequenza fondamentale di piccole unità fondamentali di eguale durata, ma a modificare il segnale originale solo quando è indispensabile, ottenendo risultati ottimi per quanto concerne la naturalezza timbrica della voce sintetica. I frammenti acustici diventano quindi più lunghi, anche sequenze di molti fonemi, parole o addirittura frasi intere, in modo da ridurre i punti di giunzione. Queste unità sono inoltre disponibili in più esemplari, corrispondenti ad esempio a contesti e ad intonazioni diverse. La dimensione del dizionario acustico può, infatti, raggiungere una dimensione anche 50 volte superiore a quella dei sistemi a difoni. Questa, che in passato era una difficoltà insormontabile, è stata ampiamente superata con l'avvento degli attuali computer dotati di enorme capacità di calcolo e di memoria.

LE APPLICAZIONI

Sono numerose le possibili applicazioni dei sistemi di sintesi da testo scritto "naturali" e di qualità paragonabile a quella umana. La diffusione capillare dell'utilizzo del computer sarà infatti senz'altro facilitata da un'interfaccia con cui si possa interagire con tutte le fonti di informazione in linguaggio naturale e non più secondo modalità non a tutti congeniali.

Fra le molteplici applicazioni si possono ricordare:

- la lettura di messaggi, memorizzati in sistemi di posta vocale, e-mail e fax all'interno di una mail box unificata, accessibile, tramite web o telefono;
- la lettura di pagine web;
- gli avvisi di particolari emergenze
- i servizi clienti delle aziende telefoniche, ad esempio per l'inoltro delle chiamate;
- la consultazione interattiva ed amichevole di fonti di informazione elettroniche;
- gli ausili di lettura per i portatori di disabilità visive come ad esempio i lettori di schermo (*Screen Reader*) che altro non sono che accessori del computer per riprodurre in voce qualsiasi cosa appaia sullo schermo, oppure i lettori di libri in grado di leggere autonomamente testi a stampa
- i corsi avanzati per l'apprendimento
- i portali vocali;

COSA MANCA?

La caratterizzazione di un segnale vocale in un dato stato emotivo deve essere definita tramite la misura dei correlati acustici ad esso associati, che a loro volta derivano dai vincoli fisiologici. Per esempio, quando una persona è in uno stato di paura o gioia, il battito del cuore e la pressione del sangue aumentano, la bocca diventa secca e ci sono occasionali tremori muscolari. La voce aumenta di intensità, di velocità, e nello spettro vi sono forti componenti in alta frequenza. I principali correlati acustici delle emozioni, studiati in letteratura sono: f0, durata, intensità, e una serie di caratteristiche del timbro quali la distribuzione dell'energia spettrale, il rapporto segnale-disturbo (*HNR, harmonic-to-noise ratio*) e alcuni indici di qualità della voce (*voice quality*).

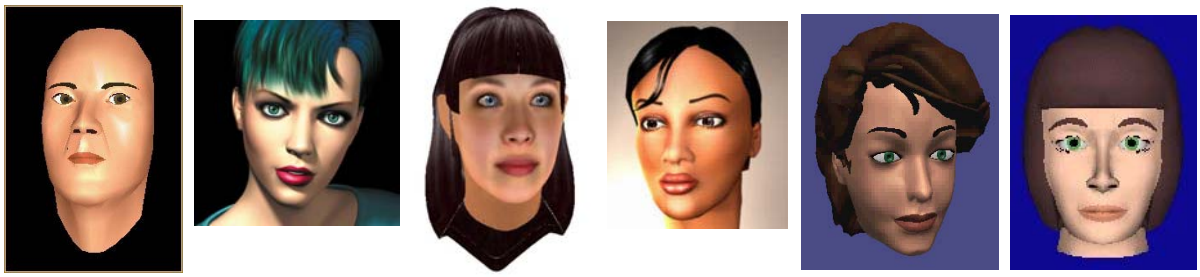
Quest'ultima proprietà distingue le modalità con cui viene prodotto il segnale glottale (voce aspirata, soffiata, tesa, ecc.. Pochissimi sistemi di sintesi includono queste diverse modalità espressive e sicuramente nessuno di quelli attualmente commercializzati: se si deve quindi leggere una fiaba ad un bambino o le "notizie ansa" in un servizio informativo le modalità espressive sono identiche. Pur tuttavia vi sono esempi in letteratura che hanno studiato questo problema cercando di elaborare alcuni modelli computazionali per rendere conto di queste caratteristiche espressive nei futuri sistemi di sintesi.

I primi esperimenti hanno utilizzato la sintesi per formanti, principalmente perché questi sistemi permettono un ricco controllo del segnale. Purtroppo però la qualità del segnale prodotto con tali strategie spesso non è soddisfacente per valutare in dettaglio l'influenza emotiva dell'uscita vocale. Utilizzando invece metodi di sintesi concatenativa, i parametri di controllo solitamente sono solo la frequenza fondamentale e la durata. Con tali strategie si possono adottare due possibili soluzioni a questo problema. Ad esempio mediante l'utilizzo di un corpus di unità acustiche per ogni emozione dal quale selezionare le unità da concatenare oppure utilizzando esclusivamente tecniche di elaborazione del segnale al fine di variare i correlati acustici emotivi legati al timbro della voce direttamente sulla forma d'onda del segnale vocale stesso.

Nonostante gli sforzi compiuti in questo filone di studio siamo però ancora distanti da un'effettiva commercializzazione di un prodotto in grado di risolvere e queste difficoltà.

IL FUTURO

A parte le difficoltà di una sintesi emotiva ed espressiva ancora non adeguatamente affrontata, il futuro della sintesi vocale risiede anche nelle nuove tecnologie di animazione facciale ad essa associata che stanno portando negli ultimi anni alla progettazione e alla realizzazione di agenti parlanti (*Talking Agents*) in grado di rendere estremamente più appetibili moltissime applicazioni interattive (si veda Figura 3) di cui le potenzialità offerte dalle nuove tecnologie di comunicazione dell'informazione fornite dai telefonini di nuova generazione, basati



sulla tecnologia UMTS, sono solo un semplice e chiaro esempio.

Figura 3.Illustrazione di alcune "*facce parlanti*" apparse recentemente "alla ribalta":

- Baldi (UCSC Perceptual Sciences Laboratory,mambo.ucsc.edu),
- Ananova (www.ananova.com),
- Lucia (ISTC-SPFD CNR, www.csrf.pd.cnr.it/Lucia/index.htm),
- Anja (Telecom Lab Italia, multimedia.telecomitalialab.com/virtual_life.htm
- Greta (Catherine Pelachaud, www.iut.univ-paris8.fr/~pelachaud/)
- Sarah (DSP.Lab Dist Genova, www.dsp.dist.unige.it/~pok/RESEARCH/index.htm)

GLI ENTI

Attualmente sono disponibili vari sistemi TTS in italiano ed in Tabella 1 sono indicati solo alcuni degli Enti attualmente impegnati nello sviluppo di queste tecnologie per l'italiano.

		Modalità di sintesi		
ENTE	**www**	**Formanti**	**Difoni**	**Unità Variabili**
Audiologic	www.audiologic.it		X	
CNR IFAC "Nello Carrara" (Diphon2, Parla)	www.ifac.cnr.it			
CNR ISTC-SPFD (Padova) (Festival)	www.csrf.pd.cnr.it/tts/It-TTS.htm		X	
ElanSpeech	www.elanspeech.com/		X	
IBM (ViaVoice TTS)	www-3.ibm.com		X	X
BaBel Technologies S.A. (INFOVOX)	www.infovox.se	X		
Loquendo (Eloquence, Actor)	www.loquendo.it		X	X
RealSpeak	www.realspeak.com		X	X

Tabella 1. Elenco "incompleto" di alcuni Enti coinvolti nello sviluppo di sistemi di sintesi TTS in italiano in varie modalità.

Bibliografia

- ABLA, vision: http://www.abla.it/it/vision/index.html
- ABLA, "*LE TECNOLOGIE VOCALI*": http://www.abla.it/it/multimedia/tec_vocali_ita.pdf
- Silvia Quazza, "*Macchine parlanti*: da sogno antico a realtà", http://www.futurecentre.telecomitalia.it/mondoict-articolo.asp?sezione=4&id=2
- Ennas Emilia e Ornato Grazia, "VoiceXML: la voce in rete", http://lips.dist.unige.it/articoli/VoiceXML/PrimaPagina.htm , http://lips.dist.unige.it/articoli/VoiceXML/Sintesi%20TTS.htm