TTS e STT (autore: Vittorio Albertoni)

Premessa

Le tecnologie attraverso le quali possiamo interagire con una macchina utilizzando la voce sono praticamente due: la sintesi vocale e il riconoscimento vocale.

Con la sintesi vocale otteniamo che un computer, ricevendo un input da tastiera o da file, produca un output verso un altoparlante o registrando su file audio, imitando la voce umana. E' ciò che si chiama TTS, che sta per Text To Speech.

Con il riconoscimento vocale otteniamo che un computer, ricevendo un input espresso con voce umana, produca un output consistente in un testo scritto su file, come risultato di una dettatura. E' ciò che si chiama STT, che sta per Speech To Text.

Come sotto-categoria di quest'ultima tecnologia abbiamo il caso della comprensione del segnale in voce umana per compiere un'azione, come potrebbe essere l'interruzione della riproduzione di un file audio dopo che abbiamo pronunciato la parola «stop».

In questo manualetto vediamo cosa ci offre il mondo del software libero per fare queste belle cose innanzi tutto lavorando con il sistema operativo Linux, ma non solo.

Indice

1 TTS Text To Speech

Probabilmente la traduzione di testo in parole recitate a voce è una funzione molto utile nel mobile ma poco utile a chi usa un personal computer. Non a caso i software migliori in questo campo derivano da applicazioni nate nell'industria automobilistica (per gli apparecchi telefonici in dotazione alle vetture) o in quella degli smartphone.

Peraltro non è certo con questi software che possiamo avere dal computer il valore aggiunto che ci può offrire un attore che ci legge una poesia o un romanzo: non siamo certo a questi livelli di perfezione.

Comunque esistono parecchi prodotti.

1.1 eSpeak

La citazione è d'obbligo in quanto si tratta del veterano di questi software, nato nel mondo Linux ma disponibile anche per Windows (con installato SAPI5) e OS X.

Chi usa Linux molto probabilmente lo trova installato con il sistema operativo o lo trova nel repository. Per tutti quanti basta andare all'indirizzo

http://espeak.sourceforge.net/download.html

E' il veterano ma è anche quello che fornisce i risultati peggiori come resa della voce umana, specialmente nel caso di frasi lunghe e articolate, dove manca di ogni espressività, non essendo sensibile alla punteggiatura.

Data la grande varietà di lingue riconosciute è molto utile per avere immediato riscontro della corretta pronuncia di parole singole. Se, per esempio, vogliamo conoscere la corretta pronuncia della parola «lieutenant», tanto per prendere una parola con pronuncia un po' strana, nell'inglese di Londra basta che scriviamo a terminale

```
espeak -v en-gb lieutenant
```

e se poi vogliamo sapere come la stessa parola si pronuncia a New York basta che scriviamo a terminale

```
espeak -v en-us lieutenant
```

Con il comando a terminale

```
espeak --voices
```

otteniamo l'elenco delle lingue riconosciute e delle relative sigle.

Se ciò che chiediamo di recitare al computer non è una sola parola ma una frase, dobbiamo racchiuderla tra apici, semplici ('....') o doppi ("...."").

Se vogliamo che il computer legga un intero file di testo già predisposto in italiano, il comando a terminale è

```
espeak -v it -f <file.txt>
```

dove, al posto di `<file.txt>` mettiamo il nome e l'estensione del file di testo da leggere. Ed è qui che ci accorgiamo di come espeak legga male.

Se vogliamo che la lettura venga registrata in un file audio, dobbiamo aggiungere l'opzione -w seguita dal nome che vogliamo dare al file audio. Ad esempio, il comando

```
espeak -v de -w prova.wav 'guten abend'
```

registra il saluto in lingua tedesca «guten abend» nel file audio prova.wav.

Infine, se vogliamo che la voce recitante non sia la voce maschile di default ma sia una voce femminile, basta che aggiungiamo alla sigla della lingua, con il segno +, la lettera f con il numero della variante; così il comando

```
espeak -v it+f5 ciao
```

restituisce un «ciao» pronunciato in italiano da una calda voce femminile (l'ultima delle 5 disponibili, che per me è la migliore).

Esiste un'ottima interfaccia grafica di questo software che, utilizzando le librerie GTK, gira sicuramente su Linux ma non su Windows e OS X, a meno che vi siano caricate quelle librerie, e si chiama **gespeaker**. La si trova nel repository della distro Linux che utilizziamo.

Essa, come si può vedere dalla seguente figura 1,

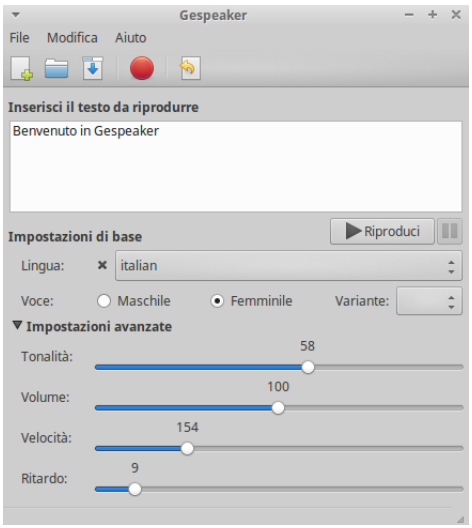


Figura 1: Finestra di lavoro di Gespeaker

semplifica notevolmente tutto.

1.2 Pico TTS e relativa estensione a LibreOffice Writer

Pico TTS è la versione open source del motore Svox, il più utilizzato dai sofisticati sistemi di navigazione satellitare delle migliori marche automobilistiche e che ha poi invaso il mondo degli smartphone.

A differenza di ciò che avviene con il vecchio glorioso eSpeak, qui la recitazione delle frasi è notevolmente migliore, è sensibile alla punteggiatura, nel senso che la virgola e il punto provocano lievi distacchi, il punto di domanda provoca inflessione diversa da quella provocata dal punto esclamativo, ecc.

E' disponibile per Linux in sei lingue (italiano, francese, tedesco, spagnolo, inglese gb e inglese usa).

Le relative librerie sono

libttspico0
libttspico-utils
libttspico-data

e sono disponibili per il sistema operativo Linux della famiglia Debian/Ubuntu e derivate, compreso Mint.

Si possono installare, con collegamento Internet attivo, con il comando a terminale

`sudo apt-get install libttspico0 libttspico-utils libttspico-data`.

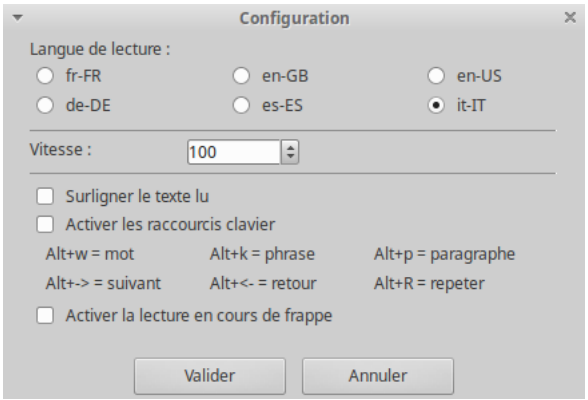
Fatto questo, si scarica l'estensione **picosvoxooo_v1_1.oxt** all'indirizzo

https://fr.osdn.net/projects/sfnet_aoo-extensions/downloads/17608/2/picosvoxooo_v1_1.oxt/

e la si installa aprendo LibreOffice Writer, menu STRUMENTI ▷ GESTIONE ESTENSIONI… ▷ AGGIUNGI...

Alla riapertura di Writer troveremo nella barra degli strumenti le seguenti icone

Cliccando sulla prima a sinistra apriamo la finestra delle impostazioni



nella quale dobbiamo innanzi tutto scegliere la lingua.
Nella finestrella VITESSE possiamo scegliere una velocità di lettura.
Infine possiamo attivare una o più delle altre opzioni che ci vengono proposte (sottolineare il testo letto, attivare combinazioni di tasti rapidi, fare in modo che vengano lette le singole parole man mano vengono scritte con la tastiera).
Le altre icone si riferiscono ai comandi di lettura e di navigazione nel testo: passando il mouse su ciascuna di esse ce ne viene descritta la funzione.

Ovviamente non è necessario che il testo da leggere sia prodotto da LibreOffice Writer. Basta aprire con LibreOffice Writer un qualsiasi testo prodotto prima e altrove per poterlo leggere.

Con questa applicazione non possiamo memorizzare la lettura in file audio. Possiamo rimediare semplicemente attivando un registratore di suoni mentre il computer legge.

1.3 Google Cloud Speech API

API sta per Application Programming Interface e Google Cloud Speech API sta ad indicare che Google mette a disposizione on-line per gli sviluppatori una interfaccia che consenta loro di creare applicazioni utilizzando la propria tecnologia di sintesi vocale.

Il tutto comporta due problemi: innanzi tutto bisogna saper programmare per utilizzare questa opportunità e i programmi così costruiti potranno girare solo essendo collegati a Internet.
Il secondo problema penso sia facilmente superabile.
Quanto al primo, il linguaggio di programmazione Python ha un modulo che con poco sforzo consente a chiunque di farsi un programmino per fare leggere del testo al computer utilizzando la tecnologia Google.

Il modulo gtts di Python

gtts sta per Google Text To Speech.

Chi ci tenga ad una presentazione un po' strutturata di Python e del mondo che gira attorno a Python può utilmente leggere il mio articolo «Python per tutti» del febbraio 2017, archiviato sul mio blog *vittal.it*, e il relativo allegato in formato PDF «mondo_python». Consiglio la lettura di questo allegato a chi voglia sviluppare l'esercizio che propongo in ambiente Mac OS X o Windows. Per chi usa Linux basta seguire le istruzioni che fornisco in questa sede. Ovviamente si tratta di premesse inutili per coloro che già conoscono Python.

Prima cosa che serve è avere installato Python sul computer. Chi usa Linux ha molto probabilmente già installato entrambe le versioni di Python (la 2 e la 3) insieme al sistema operativo. Chi usa Mac OS X molto probabilmente ha installato la sola versione 2. Chi usa Windows, se proprio vuole fare queste cose in un ambiente ostile a chi vuole capire l'informatica, deve installarselo e prepararsi a tribolare un po' di più. Tutto comunque, installer e istruzioni per l'installazione, si trova all'indirizzo

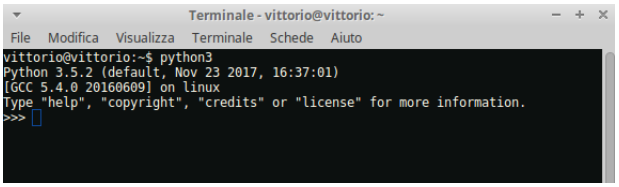
https://www.python.it/download/

Per l'esercizio che propongo dobbiamo disporre di Python 3, la versione attuale e del futuro.

Ora dobbiamo disporre di un programmino (si chiama pip) che ci aiuti ad installare i moduli che arricchiscono la dotazione di base di Python: ci serve per procurarci il modulo gtts. In ambiente Linux Debian/Ubuntu lo facciamo, collegati a Internet, con il comando a terminale
`sudo apt-get install python3-pip`[1] .
Nel mio documento prima richiamato si trovano le istruzioni per gli altri sistemi operativi.

A questo punto installiamo il modulo con il comando a terminale
`sudo pip3 install gtts`[2] .

Ora siamo finalmente in grado di affrontare il nostro esperimento, che qui propongo di eseguire in maniera interattiva nella shell di Python, dopo averla aperta con il comando a terminale
`python3`



[1] Nelle versioni più recenti di Linux non si usa più `apt-get` ma semplicemente `apt` e il comando completo diventa `sudo apt install python3-pip`.

[2] Ai novizi di Linux ed ai curiosi che usano altri sistemi operativi rammento che sudo sta per super user do e sta ad indicare un'azione da super utente, quale quella di installare un nuovo programma, prima di svolgere la quale il computer chiede l'immissione di una password. E' questa una delle basi della grande sicurezza del sistema Linux.

Innanzi tutto importiamo il modulo, con il comando
`from gtts import gTTS`
poi utilizziamo questo modulo per costruire un oggetto, che chiamiamo «tts», contenente la recitazione con accento italiano della frase «Buon giorno, Vittorio»; lo facciamo con il seguente comando
`tts = gTTS(text='Buon giorno, Vittorio' lang='it')`
ed ora salviamo il contenuto dell'oggetto tts in un file audio, che chiamiamo «saluto» in formato mp3 (quello che produce il nostro modulo); lo facciamo con il comando
`tts.save('saluto.mp3')`.

Nella nostra directory personale troviamo il file saluto.mp3, pronto per essere ascoltato.

All'indirizzo *https://github.com/pndurette/gTTS* abbiamo tutte le istruzioni per utilizzare il modulo gTTS e l'elenco di tutti gli accenti linguistici gestiti con l'indicazione delle rispettive sigle di richiamo (it per l'italiano, de per il tedesco, fr per il francese, es per lo spagnolo, en-uk per l'inglese britannico, en-us per l'inglese americano e moltissime altre).

Nel nostro esercizio abbiamo assegnato al parametro `text` una stringa di testo scritta nel comando. Se volessimo far leggere un testo già scritto altrove dovremmo, utilizzando il linguaggio Python, aprire il file contenente lo scritto, leggerlo ed assegnarne il contenuto a una variabile stringa. Dopo di che l'assegnazione del parametro avverrebbe mettendo, dopo `text=`, il nome della variabile stringa senza apici. Ma ora andiamo oltre il piccolo esercizio dimostrativo sull'uso del modulo gtts.

Ovviamente, se invece di lavorare in modo interattivo nella shell scrivessimo le istruzioni in un file di testo prevedendo anche il richiamo di sistema per riprodurre il file audio e quant'altro, magari arricchendo il tutto con una interfaccia grafica, potremmo produrre noi qualche cosa di simile al Gespeaker illustrato nella figura 1 ed avremmo un nuovo Gespeaker con una voce ed un modo di leggere migliori.

E così ci siamo fatti anche una piccola immersione in quello che è di moda chiamare «coding».

2 STT Speech To Text

Se prescindiamo dal valore delle indicazioni che siamo abituati a sentire dal navigatore dell'auto, la pratica utilità della macchina parlante è molto relativa ed è sicuramente inferiore a quella della macchina che capisce il nostro linguaggio vocale.

In questo caso abbiamo il riconoscimento vocale al posto della sintesi vocale: STT al posto di TTS. Ora siamo noi che parliamo e la macchina capisce quello che diciamo. Con due possibili conseguenze: scrivere ciò che diciamo in un file di testo (dettatura) oppure eseguire un comando che impartiamo a voce (comando vocale).

2.1 Dettatura

In genere il software di dettatura è sempre stato prodotto e venduto da strutture profit oriented. Nel mondo del software libero si è lavorato e si lavora parecchio ma non c'è nulla di immediatamente fruibile senza complicazioni di configurazione e di praticità d'uso. Il sistema Linux ha a disposizione alcuni motori open source per il riconoscimento della voce: da Sphinx a Kaldi, da Julius alla novità ancora in fase di sviluppo Deep Speech di Mozilla. Chi voglia approfondire può trovare in rete informazioni su questi progetti. Non ritengo però vi sia nulla alla portata del dilettante, pur evoluto.

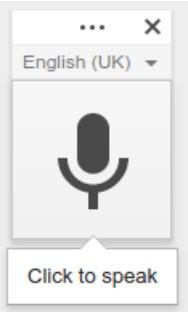
Ancora una volta ci potrebbe dare una mano Python con un modulo che si chiama SpeechRecognition e che si pone come interfaccia per utilizzare il motore Sphinx o la tecnologia Google Speech Recognition, che non possiamo considerare software libero ma che Google mette gentilmente a disposizione. Le complicazioni vanno tuttavia molto oltre quel poco che abbiamo visto nell'esercizio che ho proposto alla fine del precedente capitolo.

L'unica cosa pronta all'uso e che fornisce eccellenti risultati ci viene offerta sempre da Google, come servizio nell'ambito di Google Drive.

Voice typing di Google

Per utilizzare il servizio occorre avere un account Google e accedere a Google Drive tramite il browser Google Chrome o, meglio ancora, visto che sono un fan del software libero, la sua versione open Chromium (tutti i browser vanno bene per accedere a Google Drive, ma il servizio di voice typing funziona solo con Chromium o Google Chrome).

Sul browser apriamo il menu delle app Google premendo il pulsante ⋮⋮⋮ in alto a destra e scegliamo l'app Drive premendo il punsante .
Nella nuova finestra clicchiamo sul pulsante blu in alto a destra, con la scritta New e scegliamo la voce Google Docs nel menu a discesa.
Ci si presenta così un editor di testo in tutto simile a Writer di LibreOffice e a Word di Microsoft Office.
Dalla barra del menu apriamo TOOLS e scegliamo VOICE TYPING…
Sulla sinistra dello schermo compare questa icona



che, nell'indicarci dove premere per avviare la possibilità di scrivere sotto dettatura, ci indica la lingua di default English (UK).
Premendo sul triangolino apriamo il menu in cui scegliere la lingua che intendiamo utilizzare e poi clicchiamo sul microfono.
L'icona si trasformerà in questa

e tutto ciò che diremo fino a quando sarà presente questa icona verrà scritto nella finestra dell'editor di testo.
Ovviamente dobbiamo parlare scandendo bene le parole, come dettando.
Per terminare la dettatura clicchiamo sul microfono rosso.

A patto che la registrazione sia buona e riprodotta con volume abbastanza elevato la dettatura può avvenire anche riproducendo un file audio con gli altoparlanti del computer.

In sede di dettatura vengono riconosciuti alcuni comandi per la punteggiatura e la formattazione del testo. Possiamo vedere di che cosa si tratta cliccando sul punto interrogativo che compare se passiamo il mouse sull'icona con il microfono nero.
La maggiore varietà di comandi funziona utilizzando la lingua inglese.
Se scegliamo la lingua italiana sono riconosciuti solo alcuni comandi per la punteggiatura (punto, virgola, punto esclamativo, punto interrogativo, nuova riga, nuovo paragrafo).

Il documento prodotto sotto dettatura in Google Drive viene archiviato nel cloud di Google. Qui lo possiamo conservare e da qui lo possiamo riformattare, rivedere, stampare, condividerlo con altri, ecc. Per produrne una copia nel formato della nostra attrezzatura d'ufficio (LibreOffice, Microsoft Office, ecc.) e sul nostro computer, dobbiamo copiarne il contenuto ed incollarlo nel nostro word processor preferito e da qui memorizzarlo dove vogliamo.

2.2 Comandi vocali

Il riconoscimento vocale può essere utile per impartire comandi al computer.

Nel mondo degli smartphone si può ormai fare praticamente tutto con la voce, ma non è così nel mondo dei personal computer.

A parte il costoso software commerciale Dragon, famoso anche come software per la dettatura, non esiste praticamente nulla di serio, almeno per chi parla italiano. Microsoft ha arricchito il proprio sistema Windows con la possibilità di impartire comandi vocali, ma il riconoscimento della nostra lingua non mi pare sia ancora contemplato.

Il software libero ci ha proposto alcune soluzioni, come **voice-commands** e **google2ubuntu**, che non funzionano sulle più recenti versioni di Linux.

Rimane, fino a nuovi accadimenti, **Simon**.

Lo troviamo nei repository delle varie distribuzioni Linux e lo possiamo caricare con il gestore dei programmi.

Si tratta di un software di utilizzo tutt'altro che semplice e, attraverso una complicata configurazione, teoricamente ci dovrebbe consentire di fare tutto, purché il tutto lo predisponiamo noi. Le uniche cose già pronte vanno bene per altre lingue, soprattutto il tedesco.

Se qualcuno si vuole cimentare, trova tutte le indicazioni all'indirizzo

https://simon.kde.org/

dove è possibile scaricare il software, anche per Windows e OS X.

Secondo me il gioco non vale la candela.