Speech Recognition HOWTO

Stephen Cook

scook@gear21.com

Diario delle Revisioni

Revisione v2.0 19 Aprile 2002 Revisionato da: scc
Modificata la licenza (ora GFDL) e aggiunta una nuova pubblicazione.
Revisione v1.2 5 Febbraio 2002 Revisionato da: scc
Aggiunti altri software commerciali (inviati da Mayur Patel).
Revisione v1.1 5 Ottobre 2001 Revisionato da: scc
Aggiunta informazione su Vocalis Speechware. Corrette/Aggiornate varie altre cose.
Revisione v1.0 20 Novembre 2000 Revisionato da: scc
Aggiunta informazione su L e H e HTK.
Revisione v0.5 13 Settembre 2000 Revisionato da: scc
Versione Iniziale.

Il riconoscimento del parlato (ASR, dall'inglese Automatic Speech Recognition) su sistemi Linux sta diventando una attività sempre più semplice. Sono disponibili oggi diversi pacchetti software indirizzati a utenti finali e a sviluppatori. In questo documento si analizzano gli aspetti di base del riconoscimento del parlato e si descrivono alcuni dei programmi disponibili. Traduzione a cura di Marco Cova e revisione a cura di Sandro Cardelli.

Sommario

1. Legal Notices

1.1. Copyright/License

1.2. Liberatoria

L'uso del presente materiale è unicamente a vostro rischio e pericolo. L'autore non si assume alcuna responsabilità per i contenuti di questo documento, incluso responsabilità per danni finanziari e fisici. In nessun modo l'autore si assumerà responsabilità per eventuali danni indiretti o consequenziali l'utilizzo di concetti, casi di esempio, informazioni contenuti nel presente documento.

1.3. Trademarks

Tutti i copyright contenuti in questo documento sono mantenuti dai rispettivi proprietari.

2. Forward

2.1. Informazioni su Questo Documento

Questo documento è indirizzato a utenti Linux con un livello di competenze basso o intermedio interessati ad acquisire delle conoscenze teoriche sul riconoscimento del parlato e a provare questa tecnologia nella pratica. Può, inoltre, risultare utile a sviluppatori che vogliano imparare gli aspetti di base della programmazione di riconoscimento del parlato.

La stesura di questo documento è iniziata in contemporanea alle mie ricerche di programmi e librerie di sviluppo per il riconoscimento del linguaggio disponibili per Linux. Il riconoscimento del parlato (ASR o semplicemente SR) su Linux ha appena iniziato a muovere i suoi primi passi. Spero che questo documento lo possa indirizzare nella giusta direzione e sia uno strumento utile sia per gli utenti che per gli sviluppatori di questa tecnologia.

Questo documento non vuole essere un'esposizione completa di tutte le tecniche di SR. Si concentra, invece, nello spirito degli HOWTO, sugli aspetti più pratici. É comunque disponibile una sezione Pubblicazioni dove il lettore interessato può trovare i riferimenti a libri o articoli su argomenti che non sono trattati qui. Questo documento non vuole nemmeno essere la parola definitiva sull'argomento ASR e Linux.

Per ottenere la versione più recente di questo documento, si può consultare l'archivio di LDP o controllare l'indirizzo: http://www.gear21.com/speech/index.html.

2.2. Riconoscimenti

Vorrei ringraziare le seguenti persone per il loro aiuto, revisioni e sostegno in favore di questo documento:

- Jessica Perry Hekman
- Geoff Wexler

2.3. Commenti/Aggiornamenti

Se volete fare commenti, suggerimenti, revisioni, aggiornamenti, o siete semplicemente interessati a discutere di ASR, inviatemi un'email all'indirizzo scook@gear21.com (mailto:scook@gear21.com).

2.4. ToDo

Questa è una lista delle cose ancora da fare:

- Aggiungere delle descrizioni nella sezione Pubblicazioni.
- Aggiungere altri libri alla sezione Pubblicazioni.
- Aggiungere altri link completi di descrizione.
- Migliorare la descrizione dei vari passi di un sistema ASR.
- Includere la descrizione di FFT e Filtri.
- Includere la descrizione dei principi dei DSP.

2.5. Storia delle Revisioni

v0.1 prima bozza preliminare - Agosto 2000

v0.5 bozza finale - Settembre 2000

3. Introduzione

3.1. Introduzione al Riconoscimento del Parlato

Il riconoscimento del parlato è il processo attraverso cui un computer (o un altro tipo di macchina) riconosce il linguaggio parlato. In altri termini, attraverso questo processo, si può parlare ad un computer e fare in modo che questo identifichi correttamente le parole pronunciate.

Prima di procedere, è necessario introdurre un po' di terminologia:

Enunciato

Un enunciato è la vocalizzazione (la pronuncia) di una o più parole che hanno un singolo significato per il computer. Possono essere enunciati una singola parola, alcune parole, un'intera frase, o anche più frasi.

Dipendenza da chi parla

I sistemi ASR possono essere dipendenti o indipendenti da chi parla. I sistemi dipendenti sono progettati per soddisfare le esigenze di uno specifico utente. Generalmente, presentano un'elevata accuratezza quando utilizzati da tale utente, ma hanno prestazioni meno buone se usati da utenti differenti. Assumono, infine, che l'utente non modifichi significativamente timbro e ritmo della parlata. Al contrario, i sistemi indipendenti sono progettati per essere usati da utenti diversi. I sistemi adattivi, di solito, funzionano inizialmente come sistemi indipendenti e poi, utilizzando tecniche di training, si adattano all'utente per migliorare l'accuratezza del riconoscimento.

Vocabolari

Vocabolari (o dizionari) sono liste di parole o enunciati che possono essere riconosciute dal sistema SR. Generalmente, vocabolari di dimensioni minori permettono un riconoscimento migliore da parte del computer, mentre vocabolari più estesi creano maggiori difficoltà di riconoscimento. A differenza dei normali dizionari, ciascun elemento presente nel dizionario di un sistema SR non deve necessariamente essere una singola parola. Tali elementi possono, infatti, consistere anche di una o più frasi. I dizionari più piccoli possono essere costituiti anche solo di uno o due enunciati riconosciuti (ad esempio, "Wake up"), mentre vocabolari molto grandi possono averne un centinaio di migliaia o anche più.

Accuratezza

Le capacità di un sistema SR si possono misurare calcolandone l'accuratezza, ovvero quanto bene è in grado di riconoscere enunciati. Questo include non solo la capacità di identificare un enunciato noto ma anche di determinare se un certo enunciato non è presente nel vocabolario. Sistemi ASR buoni possono presentare un'accuratezza del 98% o anche superiore! Il livello di accuratezza minimo accettabile per un sistema dipende dal particolare tipo di applicazione in cui è utilizzato.

Addestramento

Alcuni sistemi di riconoscimento del parlato possono adattarsi al particolare utente che li utilizza. Quando il sistema presenta questa capacità, può essere possibile effettuare una sessione di addestramento. Durante queste sessioni all'utente è richiesto di ripetere un certo numero di frasi comuni o standard per permettere al sistema di adattare gli algoritmi utilizzati al suo particolare modo di parlare. L'addestramento di un ASR generalmente ne migliora l'accuratezza di riconoscimento.

La possibilità di addestrare un sistema ASR può essere utilizzata da utenti che hanno difficoltà a parlare o a pronunciare determinate parole. Se l'utente è in grado di ripetere senza variazioni significative un certo enunciato, il sistema ASR, opportunamente addestrato, dovrebbe essere in grado di adattarsi e effettuare con successo il riconoscimento.

3.2. Tipologie di Riconoscitori del Parlato

I sistemi di riconoscimento del parlato possono essere suddivisi in alcune classi differenti sulla base del tipo di enunciati che sono in grado di riconoscere. Uno dei problemi principali degli ASR consiste nel determinare quando un utente inizia e finisce un enunciato. La maggior parte dei sistemi può essere inserito in diverse classi a seconda di quale tecnica utilizzano per risolvere questo problema.

Parole Isolate

Sistemi a parole isolate di solito richiedono che ciascun enunciato presenti un periodo di pausa, cioè assenza di segnale audio, su ENTRAMBI i lati della finestra di campionamento. Questo non significa che accettano solamente parole singole, ma che riconoscono un solo enunciato alla volta. Spesso questi sistemi hanno stati di "Ascolto/Non-ascolto", in cui richiedono all'utente di attendere tra la pronuncia di un enunciato e l'altro (e in queste pause il sistema elabora l'enunciato appena sentito). Sistemi a Enunciati Isolati può essere un nome migliore per questa classe.

Parole Connesse

Sistemi a parole connesse (o più correttamente a 'enunciati connessi') sono simili ai sistemi a parole isolate, ma permettono che enunciati isolati siano pronunciati all'unisono, con una pausa minimale tra l'uno e l'altro.

Parlato Continuo

Sistemi a discorso continuo rappresentano il passo successivo. Questi sistemi sono tra i più difficili da creare in quanto devono impiegare delle tecniche speciali per determinare i confini di un enunciato. Permettono all'utente di parlare in maniera quasi del tutto naturale. Sono sistemi di dettato al computer.

Parlato Spontaneo

Sembra che ci sia una varietà di possibili definizione di parlato spontaneo. Al livello più semplice, lo si può definire come parlato che sembra naturale e non preparato. Un sistema ASR in grado di riconoscere il parlato spontaneo deve gestire una serie di particolarità del linguaggio naturale, come la pronuncia continua di parole distinte, suoni come "um" e "ah", e anche leggere balbuzie.

Verifica/Identificazione della Voce

Alcuni sistemi ASR sono in grado di identificare specifici utenti. Questo documento non tratta sistemi di verifica o sicurezza basati sul riconoscimento vocale.

3.3. Usi e Applicazioni

In linea di principio, in ogni compito in cui è richiesto ad un utente di interfacciarsi col computer si può ricorrere a sistemi ASR. Tuttavia, le applicazioni seguenti sono quelle più comunemente utilizzate.

Dettato

Il dettato è senz'altro l'applicazione di maggior uso di sistemi ASR. Lo si usa per trascrizioni in campo medico, legale, economico, ma anche per fare del normale word processing. In alcuni casi si usano vocabolari speciali per incrementare l'accuratezza del sistema.

Comando e Controllo

Si definiscono sistemi di Comando e Controllo (C&C) gli ASR che sono progettati per eseguire particolari funzioni e azioni sul sistema. Enunciati come "Esegui Netscape" e "Esegui un nuovo xterm" ne rappresentano un esempio.

Telefonia

Alcuni sistemi PBX/Mail vocale offrono ai chiamanti la possibilità di dire il nome dei comandi che vogliono eseguire anziché richiedere la pressione dei corrispondenti pulsanti.

Sistemi Wearable

Dal momento che l'input nei dispositivi wearable è piuttosto limitato, poterli comandare via voce è una possibilità attraente e naturale.

Campo Medico/Disabilità

Molte persone hanno difficoltà ad usare la tastiera a causa di lesioni indotte da stress fisici ripetuti (RSI, dall'inglese Repetitive Strain Injuries), distrofia muscolare e altre cause ancora. Per esempio, quanti hanno difficoltà uditive potrebbe connettere un sistema ASR al loro telefono per convertire la voce del chiamante in formato testuale.

Applicazioni Embedded

Alcuni telefoni cellulari includono funzionalità di riconoscimento del parlato C&C che riconoscono enunciati come "Chiama Casa". Questo potrebbe essere un campo importante per lo sviluppo di ASR e Linux. Perché non posso ancora parlare alla mia televisione?

4. Hardware

4.1. Schede Audio

Dal momento che il parlato ha richieste di banda relativamente modeste, quasi tutte le schede audio di qualità medio-alta a 16 bit sono sufficienti per il compito di riconoscimento del parlato. Bisogna abilitare il supporto del suono nel kernel e si deve disporre dei driver corretti per la propria scheda audio. Per maggiori informazioni sulle schede audio, è possibile fare riferimento a "The Linux Sound HOWTO" disponibile all'indirizzo http://www.LinuxDoc.org/. L'argomento della qualità delle schede audio spesso è causa di discussioni animate sul loro impatto su rumore e accuratezza del riconoscimento.

Sono raccomandate schede audio con la conversione A/D (da analogico a digitale) più 'pulita'. Tuttavia, nella maggior parte dei casi, la chiarezza del suono digitale campionato dipende più dalla qualità del microfono e dalla presenza di rumore che dalle caratteristiche della scheda audio. Il rumore elettrico causato da monitor, slot PCI, dischi fissi, eccetera, di solito ha un impatto del tutto trascurabile rispetto al rumore sonoro causato da ventole di raffreddamento di computer, sedie scricchiolanti o un respiro pesante.

Alcuni programmi ASR potrebbero richiedere una scheda audio specifica. Tipicamente, è una buona idea tenersi lontano da quei programmi che impongono specifici requisiti in fatto di hardware, dal momento che questo limita notevolmente le possibiltà di effettuare cambiamenti futuri. Si dovranno valutare attentamente i benefici e gli svantaggi offerti da software che richiedono hardware specifico per funzionare correttamente.

4.2. Microfoni

Un microfono di buona qualità è un componente fondamentale quando si usa un sistema ASR. Nella maggior parte dei casi, i normali microfoni da desktop non sono sufficienti: tendono a raccogliere troppo rumore dall'ambiente, rendendo, così, difficile il lavoro dell'ASR.

Nemmeno i microfoni che si tengono in mano sono la scelta migliore: possono essere scomodi visto che bisogna raccoglierli ogni volta. Tuttavia, sono efficaci a limitare la quantità di rumore che assorbono dall'ambiente e sono molto adatti ad applicazioni in cui l'utente che parla cambia frequentemente o quando non si deve parlare al sistema ASR molto frequentemente (nel cui caso usare un microfono a cuffia non è una buona scelta).

La scelta migliore e di gran lunga più comune consiste nell'usare un microfono a cuffia. Questo tipo di microfono, infatti, minimizza il rumore raccolto dall'ambiente ed è sempre vicino alla bocca. Sono disponibili modelli con e senza auricolari (mono o stereo). Io raccomando quelli stereo, ma è una questione di gusti personali.

Si possono trovare microfoni a cuffia per un prezzo che varia da 25 a 100 dollari. Un buon posto in cui iniziare a cercarli è http://www.headphones.com o http://www.speechcontrol.com.

Una nota veloce a proposito della regolazione: non dimenticate di alzare il volume del microfono. Lo si può fare con programmi come XMixer o OSS Mixer e bisognerebbe stare attenti a evitare interferenze. Se il sistema ASR include programmi di auto-regolazione, è bene usarli dal momento che sono ottimizzati per il particolare sistema di riconoscimento.

4.3. Computer/Processori

Applicazioni ASR risentono notevolmente della velocità di elaborazione del processore. Questa è una conseguenza del fatto che il processo di ASR effettua un gran numero di calcoli per il filtraggio digitale e l'elaborazione del segnale.

Come succede per ogni software CPU-intensive, più il processore è veloce meglio è. È possibile usare alcuni sistemi SR con un processore a 100MHz e 16MB di RAM. Tuttavia se il sistema richiede di poter effettuare delle elaborazioni velocemente (usa dizionari di grandi dimensioni, implementa schemi per il riconoscimento complessi, o ha una frequenza di campionamento elevata), si dovrebbe come minimo ricorrere ad un sistema a 400MHz e 128MB di RAM. In ogni caso, la maggior parte dei programmi elenca i requisiti minimi hardware richiesti.

Non si è ancora ricorsi all'uso di cluster (Beowulf o di altro tipo) per svolgere compiti di riconoscimento molto onerosi. Se siete a conoscenza di un simile progetto, già in atto o ancora in sviluppo, fatemelo sapere! scook@gear21.com (mailto:scook@gear21.com)

5. Software per il Riconoscimento del Parlato

5.1. Software Free

La maggior parte dei programmi free elencati qui è scaricabile dall'indirizzo:

http://sunsite.uio.no/pub/Linux/sound/apps/speech/

5.1.1. XVoice

XVoice è un sistema di dettato di tipo a parlato continuo che può essere usato con una serie di applicativi per XWindow. Permette all'utente di definire delle macro personalizzate. È un ottimo programma con un chiaro futuro. Una volta impostato, riconosce il parlato con una accuratezza adeguata.

XVoice richiede di scaricare e installare il programma di IBM ViaVoice per Linux (si veda la Sezione sui programmi commerciali). Prima di usare XVoice è necessario configurare correttamente ViaVoice. Inoltre, è richiesta la libreria Lesstif/Motif (libXm). È bene notare che, dal momento che questo programma interagisce con XWindow, si deve lasciare accessibile X sulla propria macchina. Conseguentemente, si deve prestare attenzione se il computer su cui lo si usa è connesso a internet o è usato da più utenti.

Questo software è principalmente destinato agli utenti. È disponibile in formato RPM.

HomePage: http://www.compapp.dcu.ie/~tdoris/Xvoice/ http://www.zachary.com/creemer/xvoice.html

Progetto: http://xvoice.sourceforge.net

Community: http://www.onelist.com/community/xvoice

5.1.2. CVoiceControl/kVoiceControl

CVoiceControl sta per Console Voice Control ed era originariamente stato progettato come KVoiceControl (KDE Voice Control). È un sistema di base di riconoscimento del parlato che permette ad un utente di eseguire applicazioni usando comandi vocali. CVoiceControl ha sostituito KVoiceControl.

Il programma include una utility di configurazione per il microfono, un vocabolario "model editor" per l'aggiunta di nuovi comandi e enunciati, e il sistema di riconoscimento del parlato vero e proprio.

CVoiceControl è un ottimo punto di partenza per utenti esperti che vogliono iniziare ad esplorare i sistemi ASR. Non è il sistema più facile da usare, ma una volta che è stato correttamente addestrato può essere molto utile. Non dimenticate di leggere la documentazione prima di configurarlo.

Questo software è principalmente per utenti.

Homepage: http://www.kiecza.de/daniel/linux/index.html

Documenti: http://www.kiecza.de/daniel/linux/cvoicecontrol/index.html

5.1.3. Open Mind Speech

Iniziato alla fine del 1999, Open Mind Speech ha cambiato nome diverse volte: era VoiceControl, poi SpeechInput, e dopo FreeSpeech. Fa ora parte di "Open Mind Initiative". È un progetto open source. Al momento, non è completamente operativo ed è destinato principalmente a sviluppatori.

Questo software è principalmente per sviluppatori.

Homepage: http://freespeech.sourceforge.net

5.1.4. GVoice

GVoice è una libreria di ASR che usa ViaVoice di IBM per controllare applicazioni Gtk/GNOME. Include librerie per l'inizializzazione, il motore di riconoscimento, manipolazione di vocabolario e controllo del pannello. Tuttavia, lo sviluppo di GVoice non procede da oltre un anno.

Questo software è principalmente per sviluppatori.

Homepage: http://www.cse.ogi.edu/~omega/gnome/gvoice/

5.1.5. ISIP

L'Institute for Signal and Information Processing presso la Mississippi State University ha reso disponibile il proprio motore riconoscimento del parlato. Il toolkit include un front-end, un decoder e un modulo per l'addestramento. È un sistema funzionale.

Questo software è principalmente per sviluppatori.

Il toolkit (e altre informazioni su ISIP) è disponibile all'indirizzo: http://www.isip.msstate.edu/projects/speech/

5.1.6. CMU Sphinx

Sphinx è un progetto creato in origine a CMU. È stato recentemente rilasciato come software open source. È un programma piuttosto vasto che include un gran numero di strumenti e informazioni. È ancora in sviluppo, ma comprende moduli di addestramento, riconoscimento, modelli acustici e linguistici. La documentazione è scarsa.

Questo software è principalmente per sviluppatori.

Homepage: http://www.speech.cs.cmu.edu/sphinx/Sphinx.html

Source: http://download.sourceforge.net/cmusphinx/sphinx2-0.1a.tar.gz

5.1.7. Ears

Anche se Ears non è ancora del tutto sviluppato, è un buon punto di partenza per programmatori che vogliono esplorare i sistemi ASR.

Questo software è principalmente per sviluppatori.

Sito FTP: ftp://svr-ftp.eng.cam.ac.uk/comp.speech/recognition/

5.1.8. NICO ANN Toolkit

Il toolkit NICO Artificial Neural Network è una flessibile rete neuronale con feedback ottimizzata per applicazioni di riconoscimento del parlato.

Questo software è principalmente per sviluppatori.

Homepage: http://www.speech.kth.se/NICO/index.html

5.1.9. Software di Myers basato su Hidden Markov Model

Questo software scritto da Richard Myers implementa algoritmi HMM ed e scritto in C++. Fornisce un esempio e uno strumento di apprendimento dei modelli HMM descritti nel libro di L. Rabiner "Fundamentals of Speech Recognition".

Questo software è principalmente per sviluppatori.

Informazioni sono disponibili all'indirizzo: http://www.itl.atr.co.jp/comp.speech/Section6/Recognition/myers.hmm.html

5.1.10. Jialong è Speech Recognition Research Tool

Anche se non è stato originariamente scritto per Linux, questo strumento può essere compilato su Linux. Contiene tre diversi tipi di riconoscitori: DTW, HMM Dinamico, e un HMM a Densità Continua. È un programma per ricercatori e sviluppatori e non è un sistema ASR completamente funzionale. Tuttavia, contiene alcuni strumenti davvero utili.

Questo software è principalmente per sviluppatori.

Maggiori informazioni sono disponibili all'indirizzo: http://www.itl.atr.co.jp/comp.speech/Section6/Recognition/jialong.html

5.1.11. Altri Sistemi Free Software?

Se siete a conoscenza di programmi free software che non sono inclusi in questa lista, mandatemi un'email all'indirizzo scook@gear21.com (mailto:scook@gear21.com). Se volete, potete anche dirmi come ottenere una copia del programma e qualsiasi impressione vi abbia fatto. Grazie!

5.2. Software Commerciale

5.2.1. IBM ViaVoice

IBM ha mantenuto le sue promesse di supportare Linux con la loro serie di prodotti ViaVoice per Linux. Nonostante questo, il futuro dei suoi SDK è ancora piuttosto incerta: la licenza per gli sviluppatori non è stata ancora rilasciata - maggiori informazioni in futuro.

La versione commerciale (a pagamento) di IBM ViaVoice Dictation per Linux (disponibile all'indirizzo http://www-4.ibm.com/software/speech/linux/dictation.html) funziona molto bene, ma ha esigenze di sistema più consistenti rispetto ad altri sistemi ASR meno evoluti (64MB di RAM e un processore Pentium a 233MHz). Per il prezzo di 59.95 dollari si ottiene anche un microfono Andrea NC-8 microphone. Permette anche di essere usato da più di un utente, ma non l'ho mai provato in questo modo. Se qualcuno ne avesse esperienza, me lo faccia sapere. Viene fornita documentazione (in formato PDF), un modulo per l'addestramento, il sistema per il dettato e una serie di script per l'installazione. Supporto per varie distribuzioni di Linux basate sul kernel 2.2 è disponibile nell'ultima versione.

L'SDK ASR è disponibile gratuitamente e include SMAPI, grammar API, documentazione e un certo numero di programmi di esempio. Il ViaVoice Run Time Kit fornisce un motore ASR, dati per le funzioni di dettato, altre utility. Lo stesso vale per il ViaVoice Command & Control Run Time Kit. L'SDK e i Kit richiedono almeno 128MB di RAM e un kernel 2.2 o più avanzato.

L'SDK e i Kit sono disponibili all'indirizzo: http://www-4.ibm.com/software/speech/dev/sdk_linux.html

5.2.2. Vocalis Speechware

Maggiori informazioni su Vocalis and Vocalis Speechware è disponibile all'indirizzo: http://www.vocalisspeechware.com e http://www.vocalis.com.

5.2.3. Babel Technologies

Babel Technologies fornisce un SDK per Linux chiamato Babear. È un sistema indipendente dall'utente basato su HMM e reti neurali. Dispongono inoltre di un certo numero di prodotti per effettuare text-to-speech, riconoscimento vocale degli utenti analisi dei fonemi. Maggiori informazioni sono disponibile all'indirizzo: http://www.babeltech.com.

5.2.4. SpeechWorks

Non ho trovato nulla sul loro sito web che menzionava esplicitamente Linux, ma il loro "OpenSpeech Recognizer" usa VoiceXML, che è un open standard. Maggiori informazioni sono disponibili all'indirizzo: http://www.speechworks.com.

5.2.5. Nuance

Nuance offre un prodotto per il riconoscimento del parlato, attualmente arrivato alla versione 8.0 per una varietà di piattaforme *nix. Può gestire vocabolari molto grandi e adotta un'architettura distribuita per migliorare la scalabilità e tolleranza ai guasti del sistema. Maggiori informazioni sono disponibili all'indirizzo: http://www.nuance.com.

5.2.6. Abbot/AbbotDemo

Abbot è un sistema di ASR indipendente dall'utente e in grado di gestire un vocabolario molto esteso. È stato originariamente sviluppato dal Connectionist Speech Group all'Università di Cambridge. È stato poi trasferito (commercializzato) a SoftSound. Maggiori informazioni sono disponibili all'indirizzo: http://www.softsound.com.

AbbotDemo è una demo di Abbot. È dotato di un vocabolario di circa 5000 parole e usa un algoritmo connessionista/HMM a parlato continuo. È un programma di prova di cui non è disponibile il codice sorgente.

5.2.7. Entropic

Le buone persone di Entropic sono state comprate da Micro$oft... I loro prodotti e il loro servizio di supporto sono scomparsi. Il supporto per HTK e ESPS/waves+ non è più disponibile e il loro futuro è nelle mani di M$. Il loro vecchio sito web è accessibile all'indirizzo http://www.entropic.com e fornisce ulteriori informazioni.

K.K. Chin mi ha avvisato che gli sviluppatori originali di HTK (il Speech Vision and Robotic Group a Cambridge) forniscono ancora supporto per HTK. Esiste anche una versione "free" disponibile all'indirizzo: http://htk.eng.cam.ac.uk. Nota che Microsoft possiede ancora il copyright per il codice di HTK...

5.2.8. Altri Prodotti Commerciali

Ci sono voci di altri sistemi ASR commerciali presto disponibili (incluso L&H). Ho parlato con un paio di rappresentanti di L&H a Comdex 2000 (Las Vegas) ma non mi hanno fornito nessuna informazione su eventuali versioni per Linux o nemmeno se programmano di rilasciare alcun prodotto per Linux. Se avete ulteriori informazioni, fatemelo sapere scrivendo a scook@gear21.com (mailto:scook@gear21.com).

6. Capire il Riconoscimento del Parlato

6.1. Come Funzionano i Riconoscitori

I sistemi di riconoscimento possono essere suddivisi in due classi. Sistemi basati sul riconoscimento di pattern confrontano il parlato con dei pattern noti/appresi per determinare delle corrispondenze. Sistemi basati sulla fonetica acustica sfruttano conoscenze sul corpo umano (emissione della voce e ascolto) per confrontare feature del parlato (proprietà fonetiche come il suono delle vocali). La maggior parte dei sistemi moderni utilizza l'approccio di riconoscimento di pattern perché questo si adatta molto bene alle tecniche computazionali esistenti e tende a presentare migliori valori di accuratezza.

L'attività della maggior parte dei riconoscitori si può suddividere nei seguenti passi:

1. Registrazione del suono e Riconoscimento degli enunciati
2. Pre-filtraggio (pre-amplificazione, normalizzazione spostamento di banda, ecc.)
3. Suddivisione in Frame/Finestre (suddivisione dei dati in un formato utilizzabile)
4. Filtraggio (ulteriore filtraggio di ciascuna finestra/frame/banda di frequenze)
5. Confronto e Matching (riconoscimento degli enunciati)
6. Azione (Esegue la funzione associata col pattern riconosciuto)

Anche se ciascun passo sembra facile, in realtà richiede l'esecuzione di una moltitudine di tecniche diverse e talvolta completamente opposte.

(1) Registrazione dell'audio/enunciati: può essere fatta in un certo numero di modi. Il punto di inizio di un enunciato può essere determinato confrontando livelli audio dell'ambiente (l'energia acustica in alcuni casi) con il campione appena registrato. Il punto terminale dell'enunciato è più difficile da determinare in quanto l'utente tende ad inserire nel parlato degli "artefatti", come respiri, rumore di denti, echi.

(2) Pre-Filtraggio: lo si può fare in molti modi diversi, a seconda di altre caratteristiche del sistema di riconoscimento. I metodi più comuni sono il metodo "Banco di Filtri" che usa una serie di filtri audio per preparare il campione audio, e la Codifica Lineare Predittiva che usa una funziona di predizione per calcolare le differenze (errori). Sono anche utilizzate diverse forme di analisi spettrale.

(3) La suddivisione in Frame/Finestre separa il campione audio in parti di dimensioni specifiche. Spesso questa operazione viene effettuata direttamente nei passi 2 o 4. In questo passo, inoltre, si preparano le parti estreme del campione per l'analisi: si rimuovono gli edge click, ecc.

(4) Un ulteriore passo di Filtraggio non è sempre presente e, quando lo è, è utilizzato per effettuare gli ultimi aggiustamenti del campione prima delle fasi di confronto e matching. Spesso si fanno operazioni di allineamento temporale e normalizzazione.

Esiste un enorme numero di tecniche per il passo (5): Confronto e Matching. La maggior parte è basata sul confronto della Finestra corrente con dei campioni noti. Ci sono, poi, metodo basati su Hidden Markov Models (HMM), analisi della frequenza, analisi differenziale, tecniche di algebra lineare, distorsione spettrale, distorsione temporale. Tutti questi metodi sono usati per generare un valore di probabilità e accuratezza del match.

(6) Le Azioni sono qualsiasi cosa che lo sviluppatore voglia fare. *GRIN*

6.2. Introduzione all'Audio Digitale

L'audio è un fenomeno inerentemente analogico. La registrazione di un campione digitale richiede di convertire il segnale analogico che proviene dal microfono in un segnale digitale attraverso un convertitore A/D presente sulla scheda audio. Quando un microfono è attivo, le onde sonore fanno vibrare l'elemento magnetico presente nel microfono, generando, così, una corrente elettrica che raggiunge la scheda audio (si può pensare ad uno speaker che funziona al contrario). Il compito del convertitore A/D è, sostanzialmente, quello di registrare il valore della tensione elettrica a intervalli di tempo specifici.

Ci sono due fattori importanti da considerare durante questo processo. Il primo è la "frequenza di campionamento", ovvero quanto spesso vengono registrati i valori della tensione. Il secondo fattore è il "numero di bit per campione", ovvero quanto accurato è il valore registrato. Un terzo elemento da considerare è il numero di canali (mono o stereo), ma per la maggior parte delle applicazioni di ASR mono è sufficiente. Nella maggior parte dei casi si usano dei valori predefiniti per questi parametri e l'utente non dovrebbe modificarli a meno che non sia richiesto nella documentazione del programma. Gli sviluppatori, invece, dovrebbero sperimentare diversi valori per determinare quale garantisce un funzionamento ottimale dei loro algoritmi.

Qual è un buon valore della "frequenza di campionamento" per ASR? Considerato che il parlato richiede relativamente poca banda (compresa tra le frequenze 100Hz e 8kHz), 8000 campionamenti al secondo (8kHz) è un valore sufficiente per la maggior parte delle applicazioni. Alcuni, tuttavia, preferiscono effettuare 16000 campionamenti al secondo (16kHz) perché con questo valore è possibile ottenere informazioni più accurate sui suoni ad alta frequenza. Se si dispone di sufficiente potenza di elaborazione, è opportuno impostare la frequenza di campionamento a 16kHz. Per la maggior parte delle applicazioni di ASR, utilizzare frequenze di campionamento maggiori di circa 22kHz è uno spreco.

Qual è, invece, un buon valore per il "numero di bit per campione"? 8 bit permettono di esprimere valori compresi tra 0 e 255, ovvero di registrare 256 posizioni diverse dell'elemento magnetico del microfono. 16 bit permettono di distinguere 65536 possibili posizioni. Similmente a quanto detto per la frequenza di campionamento, se si dispone di sufficiente potenza di elaborazione, si dovrebbe impostare il numero di bit per campione a 16 bit. Come termine di paragone, un Compact Disc è codificato usando 16 bit per campione e la frequenza di campionamento è circa 44kHz.

Il formato di codifica usato dovrebbe essere semplice - lineare con segno o senza segno. Usare un algoritmo U-Law/A-Law o qualche altro schema di compressione non è generalmente utile, dal momento che comporta un costo computazionale e non fornisce grandi vantaggi.

7. Pubblicazioni

Se ci sono delle pubblicazioni che non sono elencate in questa lista e che pensate dovrebbero esserlo, fatemelo sapere mandandomi un messaggio all'indirizzo scook@gear21.com (mailto:scook@gear21.com).

7.1. Libri

- "Fundamentals of Speech Recognition". L. Rabiner & B. Juang. 1993. ISBN: 0130151572.
- "How to Build a Speech Recognition Application". B. Balentine, D. Morgan, and W. Meisel. 1999. ISBN: 0967127815.
- "Speech Recognition : Theory and C++ Implementation". C. Becchetti and L.P. Ricotti. 1999. ISBN: 0471977306.
- "Applied Speech Technology". A. Syrdal, R. Bennett, S. Greenspan. 1994. ISBN: 0849394562.
- "Speech Recognition : The Complete Practical Reference Guide". P. Foster, T. Schalk. 1993. ISBN: 0936648392.
- "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition". D. Jurafsky, J. Martin. 2000. ISBN: 0130950696.
- "Discrete-Time Processing of Speech Signals (IEEE Press Classic Reissue)". J. Deller, J. Hansen, J. Proakis. 1999. ISBN: 0780353862.
- "Statistical Methods for Speech Recognition (Language, Speech, and Communication)". F. Jelinek. 1999. ISBN: 0262100665.
- "Digital Processing of Speech Signals" L. Rabiner, R. Schafer. 1978. ISBN: 0132136031
- "Foundations of Statistical Natural Language Processing". C. Manning, H. Schutze. 1999. ISBN: 0262133601.
- "Designing Effective Speech Interfaces". S. Weinschenk, D. T. Barker. 2000. ISBN: 0471375454.

Per una VASTA bibliografia consultabile online, si consulti il sito dell'Institut Fur Phonetik: httpwww.informatik.uni-frankfurt.de.de/~ifb/bib_engl.html

7.2. Internet

news:comp.speech

Newsgroup dedicato a computer e linguaggio parlato.

- US: http://www.speech.cs.cmu.edu/comp.speech/
- UK: http://svr-www.eng.cam.ac.uk/comp.speech/
- Aus: http://www.speech.su.oz.au/comp.speech/

news:comp.speech.users

Newsgroup dedicato a utenti di software per il parlato.

- http://www.speechtechnology.com/users/comp.speech.users.html

news:comp.speech.research

Newsgroup dedicato alla ricerca su software e hardware per il linguaggio parlato.

news:comp.dsp

Newsgroup dedicato all'elaborazione digitale dei segnali.

news:alt.sci.physics.acoustics

Newsgroup dedicato alla fisica del suono.

DDLinux Email List

Mailing list dedicata al riconoscimento del parlato su Linux.

- Homepage: http://leb.net/ddlinux/
- Archivi: http://leb.net/pipermail/ddlinux/

Archivio di programmi per Linux per il linguaggio parlato

http://sunsite.uio.no/pub/linux/sound/apps/speech/

Link di Russ Wilcox sul Riconoscimento del Parlato

(eccellente) http://www.tiac.net/users/rwilcox/speech.html

Bibliografia Online

Bibliografia Online di Fonetica e Tecnologie del Parlato http://www.informatik.uni-frankfurt.de/~ifb/bib_engl.html

Homepage dello Spoken Language Systems al MIT

http://www.sls.lcs.mit.edu/sls/

Oregon Graduate Institute

Centro Spoken Language Understanding presso l'Oregon Graduate Institute. Un'eccellente risorsa per sviluppatori e ricercatori http://cslu.cse.ogi.edu/

SDK IBM ViaVoice per Linux

http://www-4.ibm.com/software/speech/dev/sdk_linux.html

Mississippi State

Homepage del Mississippi State Institute for Signal and Information Processing con una gran quantità di informazioni utili per sviluppatori http://www.isip.msstate.edu/projects/speech/

Speech Technology

Software e accessori per ASR. http://www.speechtechnology.com

Speech Control

Speech Controlled Computer Systems. Microfoni, cuffie e prodotti wireless per ASR. http://www.speechcontrol.com

Microphones.com

Microfoni e accessori per ASR. http://www.microphones.com

21st Century Eloquence

"Specialisti di Riconoscimento del Parlato." http://voicerecognition.com

Computing Out Loud

Soprattutto per utenti Windows, ma contiene informazione di buona qualità. http://www.out-loud.com

Say I Can.com

"La Fonte di Informazione sul Riconoscimento del Parlato." http://www.sayican.com