Lorenzo Cioni

Audio e Pit, due applicazioni per l'analisi di F0

(lavoro accettato al XXIII Convegno Nazionale AIA, Bologna, Italia, 12-14 Settembre 1995)

SOMMARIO

Scopo del presente lavoro é quello di presentare due programmi che sono utilizzati presso il Laboratorio di Linguistica per l'analisi di F0. Tali programmi cooperano mediante una condivisione di file di dati e un semplice meccanismo di passaggio del controllo e mettono a disposizione dell'utente un insieme completo di comandi finalizzati all'estrazione, visualizzazione e confronto di F0. Audio consente l'acquisizione e l'ascolto di speech, l'estrazione di F0 con due algoritmi e la visualizzazione dello speech con o senza la corrispondente F0 oltre ad operazioni di misura e di segmentazione. Pit, d'altro lato, consente la visualizzazione contemporanea di più grafici di F0 e l'esecuzione su di essi di operazioni di misura, editing, allineamento e splitting.

1. INTRODUZIONE

La frequenza fondamentale F0 o pitch [1] dipende dalla velocità di vibrazione delle corde vocali secondo una relazione di proporzionalità diretta e una sua stima può essere fatta valutando la frequenza dei picchi sulla forma d'onda dello speech. Il pitch rappresenta inoltre una proprietà uditiva riferita all'ascoltatore ed é determinato da fattori quali la tensione delle corde vocali, l'intensità del flusso d'aria dai polmoni e la posizione delle corde vocali e veicola informazioni relative al sesso del parlante, (in certa misura) alla sua età e al suo stato emotivo. Oltre a ciò il pitch segna in genere il confine di unità sintattiche (infatti il pitch tende a cadere a fine frase, le sillabe finali hanno in genere valori del pitch più bassi che le non iniziali, le frasi sospese sono caratterizzate da valori crescenti del pitch e così via) sebbene in molte lingue variazioni del pitch siano associate a variazioni del significato delle parole ed esistano eccezioni a quanto affermato in precedenza e per le quali si rimanda alla letteratura specialistica.

Il pitch rappresenta, pertanto, uno dei parametri fondamentali che caratterizzano il segnale vocale e numerosi sono gli algoritmi che permettono l'estrazione del pitch da segmenti di parlato.

L'estrazione del pitch presenta tuttavia una serie di problemi legati essenzialmente [2] alle caratteristiche del segnale vocale (il segnale é quasi stazionario solo su piccoli intervalli di tempo ed é caratterizzato da transizioni continue e rapide fra segnale vocalizzato e non vocalizzato) e alle interazioni fra sorgente glottale e tratto vocale.

I metodi più utilizzati per l'estrazione sono l'autocorrelazione, il cepstrum (logaritmo della DFT) e la Codifica Predittiva Lineare o LPC. In Audio gli algoritmi utilizzati sono SIFT, che usa tecniche LPC, e un algoritmo sviluppato da Henning Reetz.

SIFT [3] (Simple Inverse Filtering Tracking) lavora nel range di frequenze compreso fra 60 e 400 HZ e si compone di una serie di passi il cui scopo principale é quello di appiattire il più possibile lo spettro del segnale di ingresso. Il segnale di partenza viene filtrato con un filtro passabasso con frequenza di taglio di 900 Hz e su di esso viene eseguita una operazione di decimazione mediante la quale si riduce la frequenza di campionamento (ad esempio da 10 a 2 kHz scartando 4 campioni su 5). Il segnale decimato é quindi analizzato con il metodo di autocorrelazione in modo da ricavare i parametri del filtro inverso mediante il quale si ricostruisce un segnale dallo spettro quasi piatto cui viene applicata una operazione di autocorrelazione e il picco più alto nell'intervallo scelto é preso come valore del pitch nel periodo. Per migliorare la risoluzione la funzione di autocorrelazione é interpolata nell'intorno del valore massimo.

L'algoritmo sviluppato da Reetz [4] sembra essere più accurato e promettente con le voci femminili, caratterizzate da valori più alti del pitch. É un algoritmo che lavora nel dominio del tempo, é resistente al rumore non periodico e opera in un range di frequenze compreso fra 50 e 1000 Hz. Lo si può suddividere in quattro fasi principali: riduzione dati, filtraggio logico, chaining e post-elaborazione.

Durante la fase di riduzione dei dati il segnale di speech é convertito in un insieme di picchi prominenti. Il filtraggio logico mira all'eliminazione di picchi di energia che differiscono troppo in ampiezza e distanza dai loro vicini. Tale fase si compone di una serie di passi il cui scopo é quello di eliminare sia il rumore ancora presente dopo la prima fase sia i

picchi di ampiezza inferiore al 30% dei propri vicini oppure che non soddisfano un criterio combinato di distanza vs. ampiezza.

Il chaining, infine, ricerca un tracciato ottimale del pitch sulla base dei picchi evidenziati dalle fasi precedenti mentre la fase di post-elaborazione si occupa di eliminare segmenti troppo alti o troppo bassi o troppo corti.

2. AUDIO

Audio (vedi figura 1) é un programma originariamente sviluppato da Foti dello CSELT di Torino [5] ed é essenzialmente dedicato all'estrazione e alla visualizzazione del pitch mediante gli algoritmi SIFT e Reetz descritti nell'introduzione.

La versione che viene utilizzata presso il nostro Laboratorio [6] presenta numerose modifiche ed estensioni che tuttavia non modificano la filosofia di base del programma.

Audio consente l'acquisizione di file di speech, la loro visualizzazione con o senza il grafico della F0 corrispondente, l'esecuzione di algoritmi di analisi per l'estrazione della F0, l'esecuzione di una analisi EPD, l'esecuzione di operazioni di segmentazione e di misura.

L'acquisizione avviene mediante un sistema di acquisizione della Digital Sound Corporation DSC-240 collegato all'host su cui gira Audio, un DEC Microvax, a frequenze variabili da 8 a 12 kHz, settabili da programma, utilizzando 16 bit per campione. Tale sistema gestisce anche l'ascolto, sia di un intero file sia di una sua porzione, alla frequenza memorizzata nell'header di ciascun file di speech.

I file di speech possono essere visualizzati con o senza i corrispondenti file di pitch e su di essi possono essere eseguite operazioni di zoom e di scorrimento (panning ossia scorrimento di una finestra sulla rappresentazione dei segnali). Sui grafi dello speech e del pitch possono essere fissati cursori che consentono di evidenziare su una finestra dedicata porzioni dei segnali (speech e/o pitch) e di eseguire misure frequenza÷tempo o DF0÷Dt. La visualizzazione del pitch insieme al file di speech associato consente, fra l'altro, la rilevazione di possibili errori commessi dall'algoritmo di estrazione della F0.

Gli algoritmi di analisi descritti nell'introduzione permettono di estrarre dai file di speech gli andamenti della F0 che vengono memorizzati in file di formato opportuno. Audio consente il settaggio dei valori dei parametri sulla base dei quali viene effettuata l'analisi e, ovviamente, la visualizzazione del contenuto di tali file. I parametri settabili da programma sono diversi nei due casi.

Nel caso di SIFT l'utente può impostare i valori di ampiezza del frame di analisi, del frame di autocorrelazione, del frame di pitch, il valore del fattore di pre-enfasi e l'ordine dell'analisi LPC.

Nel caso dell'algoritmo di Reetz l'utente può settare valori di frequenza (la minima e la massima frequenza ricercate dall'algoritmo nel file di speech e i valori minimo e massimo ammissibili per il pitch), l'ampiezza del frame all'interno del quale l'algoritmo calcola il valore medio del pitch, un valore di soglia per il valore medio dell'energia, la durata massima degli spike e valori delle massime variazioni ammesse in ampiezza e frequenza oltre ad altri valori di minore importanza.

L'analisi EPD mira essenzialmente ad individuare all'interno di un file di speech (una frase, in senso lato) una successione di "parole" sulla base di periodi di silenzio, cioè di valori del segnale al di sotto di una certa soglia, più lunghi di un dato intervallo di tempo. Il valore della soglia e l'ampiezza dei periodi di silenzio possono essere settati via software dall'utente di Audio.

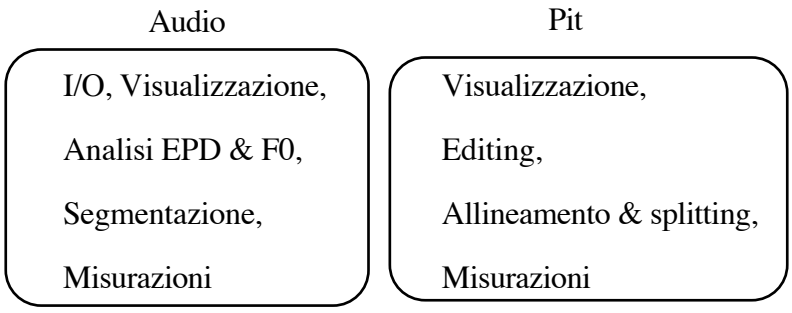

Figura 1. I domini delle applicazioni

La segmentazione, infine, opera solo su file di speech e permette di suddividere un file sia sulla base delle parole individuate dall'analisi EPD sia sulla base di intervalli di tempo settati dall'utente. I singoli segmenti vengono a loro volta memorizzati in file di speech e possono essere sottoposti ad analisi.

3. PIT

Pit é un programma esplicitamente progettato per l'esecuzione di operazioni di confronto fra più grafici della F0 contenuti in file indicati come F1 e Fn in figura 2 e ottenuti applicando uno degli algoritmi disponibili in Audio (ad esempio) ai corrispondenti file di speech.

Pit permette la visualizzazione su di una stessa finestra di un insieme di grafici della F0 e l'esecuzione su di questi di svariate operazioni mediante finestre ausiliarie e/o pulsanti. La visualizzazione può essere effettuata in modalità tempo (msec) vs. frequenza (F0 in scala logaritmica) oppure durata percentuale vs. frequenza: la seconda modalità consente un agevole confronto di file di speech di diversa durata. Il programma consente un agevole passaggio da una all'altra modalità di rappresentazione.

Le suddette operazioni possono coinvolgere tutti i grafici visualizzati (allineamento) oppure un singolo grafico alla volta (editing, splitting, misurazioni).

L'allineamento prevede che su ciascun grafico della F0 sia possibile posizionare un marker che costituisce il riferimento comune rispetto al quale le diverse traiettorie vengono ritracciate. Mediante questa semplice operazione, che viene eseguita su ogni singola forma d'onda indipendentemente dalle altre, é possibile ad esempio confrontare fra di loro gli andamenti del pitch di alcune parole considerate chiave inserite in contesti diversi.

Le operazioni di editing interattivo classiche (*cut*, *clear*, *save*, *undo*) danno all'utente la possibilità di eliminare dal grafico di una F0 valori spuri e/o indesiderati in modo da eseguire il confronto solo sulle porzioni significative dei grafici delle F0.

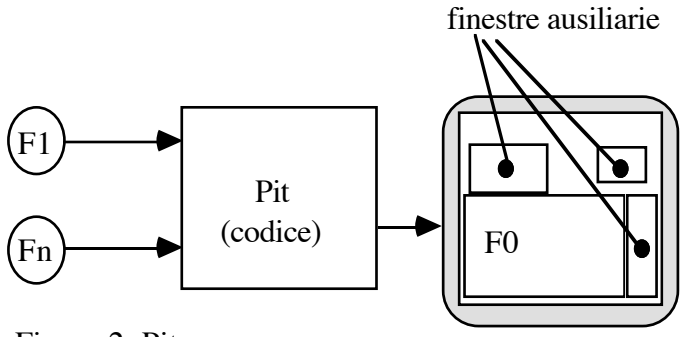

Figura 2. Pit.

Con il termine splitting si definisce la possibilità, dato un grafico della F0, di fissare su di esso un marker e di allineare le due porzioni così definite sull'origine (o allo 0%).

In questo modo, ad esempio, é possibile effettuare un confronto dell'andamento del pitch a inizio e a fine frase oppure, utilizzando le operazioni di editing, su due porzioni qualunque del file. Tale operazione può essere eseguita ricorsivamente fino ad allineare tutte le "parole" contenute in un file di pitch sull'origine (in Audio infatti l'analisi EPD permette di conoscere gli istanti di tempo in cui tali parole iniziano e finiscono).

Le operazioni di misura sono di due tipi: puntuali e globali. É infatti possibile eseguire misure di coppie tempo÷F0 (operazioni puntuali) ed eseguire su insiemi di tali coppie operazioni di media, ricerca del minimo e del massimo valore della F0 e calcolo del valore del "gradiente" (operazioni globali). Il gradiente viene definito come il rapporto fra il valore assoluto di un $\Delta F0$ ($= F0_1 - F0_2$) e di un Δt ($= t_1 - t_2$) ottenuti da due misurazioni successive.

4. COOPERAZIONE FRA PROGRAMMI

Audio e Pit sebbene possano essere usati come moduli autonomi sono stati progettati in modo che Pit sia visto come un comando di Audio così da realizzare una pipeline acquisizione-analisi (Audio) e visualizzazione-confronti-misure (Pit) schematizzata in figura 3.

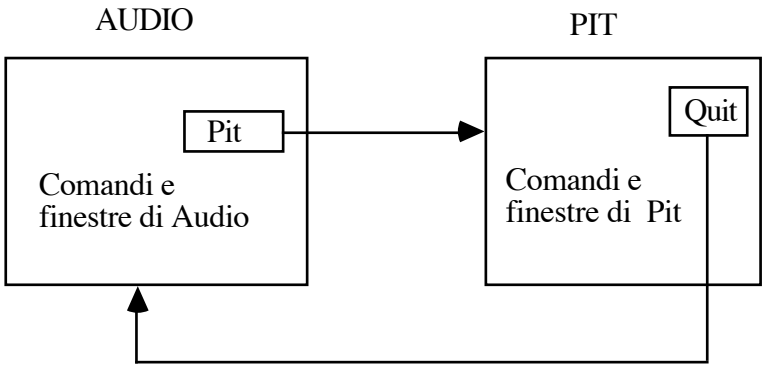

Figura 3. Passaggio del controllo fra Audio e Pit e viceversa

La disponibilità dei suddetti programmi rappresenta un modo efficace per eseguire l'analisi della F0 di file di speech: Audio consente di eseguire operazioni di I/O e di analisi mentre Pit permette di eseguire il confronto fra file di pitch, ottenuti anche da altri programmi purché aventi formato compatibile.

5. UNA APPLICAZIONE: L'ANALISI DELL'INTONAZIONE NELL'ITALIANO REGIONALE

I due programmi brevemente illustrati nel presente articolo sono stati utilizzati per la descrizione di alcune caratteristiche dell'intonazione dell'italiano regionale. Il lavoro cui ci si riferisce é stato svolto da Endo Reiko [7] con il supporto tecnico dell'autore e si é basato sull'analisi dell'andamento della F0 nelle frasi pronunciate da locutori maschili nativi di varie località italiane. Le frasi componevano un breve testo di senso compiuto ed erano di tipo sia interrogativo sia dichiarativo.

Scopo del lavoro era quello di individuare andamenti significativi in funzione dell'area geografica di provenienza dei parlanti. Nel presente articolo ci si limita ad illustrare l'utilizzo di Audio e Pit nell'ambito del lavoro in questione.

Punto di partenza é stata la registrazione su nastro, nelle località di origine, delle produzioni dei singoli locutori, tre per ciascuno: delle tre é stata scelta quella che sembrava la più spontanea.

Le registrazioni sono state quindi trasferite in file di speech sul Microvax ed analizzate con Audio utilizzando SIFT, dato che i parlanti erano tutti di sesso maschile.

Una volta ottenuti i file di pitch é stato possibile visualizzarli insieme al file di speech corrispondente mediante Audio oppure visualizzare solamente l'andamento di un insieme di file di pitch utilizzando Pit.

In tal modo é stato possibile effettuare il confronto fra andamenti intonativi di una stessa frase nello stesso contesto, frase pronunciata da parlanti diversi.

Le diverse produzioni sono, tuttavia, caratterizzate da durate diverse (ad esempio la frase "volevo fare la spesa al mercato" ha un andamento della F0 che ha una durata di 1700 msec se letta da un parlante leccese e di 1300 msec se letta da un maceratese). In questo caso la possibilità di visualizzare i grafici normalizzati rispetto al più lungo permette un agevole confronto fra i diversi andamenti intonativi. Qualora l'informazione di durata fosse significativa il programma consente la visualizzazione su una scala tempo÷frequenza e lo shifting fra le due modalità.

Il confronto fra gli andamenti della F0 su frasi diverse può essere utile, ad esempio, per analizzare le differenze esistenti fra una domanda eco, una interrogativa parziale (una WH question) e una interrogativa globale (una yes-no-question). Il programma permette di evidenziare facilmente che in generale gli andamenti sono, rispettivamente, piatto, con picco iniziale e andamento mosso nella parte finale e picco iniziale e discesa sulla parte finale.

La possibilità offerta da Pit di allineare i grafici sulla base di marker fissati indipendentemente fra loro consente inoltre di confrontare gli andamenti intonativi di singole parole inserite in contesti diversi mentre le operazioni di splitting, combinate con quelle di editing, consentono, ad esempio, di confrontare l'andamento della F0 su due porzioni qualunque di un grafico (ad esempio "volevo" e "mercato"). Lo splitting, allineando tutte le

parole sull'origine, consente di evidenziare su quale porzione del segnale si hanno maggiori variazioni della F0.

6. CONCLUSIONI E SVILUPPI FUTURI

Audio e Pit costituiscono una coppia di programmi che consentono un'analisi agevole della F0 nonostante utilizzino un hardware decisamente obsoleto. I loro pregi maggiori stanno essenzialmente nella semplicità d'uso, nella buona accuratezza e nella facilità con cui le applicazioni si scambiano dati e con cui l'utente può passare dall'una all'altra.

Sviluppi futuri comprendono una migrazione di tali applicazioni su una workstation DEC Alpha 3000/300 LX, una migliore integrazione delle applicazioni stesse e un ampliamento delle funzioni disponibili. Scopo delle funzioni aggiunte é quello di consentire confronti più approfonditi fra grafici di F0 con o senza la presenza dei corrispondenti grafici dello speech.

L'analisi di un gruppo di grafici attualmente é essenzialmente di tipo qualitativo mentre sarebbe utile poter effettuare misure di quanto i diversi grafici si discostano fra loro, dato un istante temporale o una durata percentuale.

RINGRAZIAMENTI

L'autore desidera rinnovare esplicitamente la sua gratitudine a E. Foti dello CSELT e ringraziare i suoi colleghi ed amici del Laboratorio di Linguistica. Un grazie particolare all'amico Primo Coltelli, dell'Istituto CNUCE del CNR, per i suoi stimoli continui al dubbio, e a Reiko Endo, in passato borsista del Laboratorio, al cui lavoro si ispira il paragrafo 5.

BIBLIOGRAFIA

[1] Ladefoged P., *A Course in Phonetics*, Harcourt Brace Janovich, New York, 1975

[2] Malcagni M., *Introduzione all'elaborazione digitale dei segnali*, Gruppo Editoriale Jackson, Milano.

[3] Rabin I. R., Schafer R. W., *Digital processing of signals*, Prentice Hall, New Jersey, 1978.

[4] Reetz H., *A fast expert algorithm for pitch extraction* , Proceedings of the European Conference on Speech Technology, Paris, 1989.

[5] Foti E., *Codice sorgente di Audio*, CSELT, Torino.

[6] Cioni L., *Audio & Pit: co-operating applications for analysis and visualization of F0*, articolo accettato al 4th International Symposium on Social Communication, Santiago de Cuba, Cuba, 25-27 Gennaio 1995.

[7] Endo R., *Alcune caratteristiche dell'intonazione dell'italiano regionale*, documento non pubblicato, Scuola Normale Superiore, Pisa, 1993.