UNA VALUTAZIONE OGGETTIVA DEI METODI PIÙ DIFFUSI PER L'ESTRAZIONE AUTOMATICA DELLA FREQUENZA FONDAMENTALE

Fabio Tamburini
FICLIT – Università di Bologna
fabio.tamburini@unibo.it

1. INTRODUZIONE

Il pitch, e in particolare la frequenza fondamentale - F0 - che rappresenta la sua controparte fisica, è uno dei parametri percettivi più rilevanti della lingua parlata e uno dei fenomeni fondamentali da considerare attentamente quando si analizzano dati linguistici a livello fonetico e fonologico. L'estrazione automatica di F0 è, di conseguenza, oggetto di studio da lungo tempo e in letteratura esistono numerosissimi lavori che si pongono come obiettivo lo sviluppo di algoritmi in grado di estrarre in modo affidabile F0 dalla componente acustica degli enunciati, algoritmi che vengono comunemente identificati come PDA (*Pitch Detection Algorithm*).

Tecnicamente, l'estrazione di F0 è un problema tutt'altro che banale e la grande varietà di metodologie applicate a questo problema ne dimostra l'estrema complessità, specialmente se si considera che difficilmente è possibile predisporre un PDA che funzioni in modo ottimale per le differenti condizioni di registrazione, considerando che parametri come il tipo di parlato, il rumore, le sovrapposizioni, ecc. sono in grado di influenzare pesantemente le prestazioni di questo tipo di algoritmi.

Gli studiosi impegnati sul versante tecnologico si sono spinti alla ricerca di tecniche sempre più sofisticate per questi casi estremi, ancorché estremamente rilevanti per la costruzione di applicazioni reali, considerando risolto, o magari semplicemente abbandonando, il problema dell'estrazione di F0 per il cosiddetto "*clean speech*". Tuttavia, chiunque abbia utilizzato i più comuni programmi disponibili per l'estrazione automatica di F0 è ben cosciente che errori di *halving* o *doubling* del valore di F0, per citare solo una tipologia di problemi, sono tutt'altro che rari e che l'identificazione automatica delle zone *voiced* all'interno dell'enunciato pone ancora numerosi problemi.

D'altra parte non tentiamo nemmeno di inquadrare le varie tipologie di algoritmi presenti in letteratura nella tradizionale suddivisione tra metodi nel dominio del tempo, delle frequenze o ibridi, classificazione che, ad oggi, sembra essere di difficile identificazione vista l'estrema ibridazione e specializzazione delle varie tecnologie utilizzate.

Ogni lavoro che propone un nuovo metodo per l'estrazione automatica di F0 ha ormai da anni il dovere di eseguire una valutazione delle prestazioni ottenute in rapporto agli altri PDA, ma, di solito, queste valutazioni soffrono delle tipiche mancanze che derivano da sistemi di valutazione non ottimali: ci si limita a esaminare un insieme molto limitato di algoritmi, spesso non disponibili nella loro implementazione, tipicamente considerando corpora non distribuiti, relativi a specifiche lingue e/o che contengono particolari tipologie di lingua parlata (parlato patologico, parlato disturbato da rumore, ecc.) (Veprek & Scordilis, 2002; Wu *et alii*, 2003; Kotnik *et alii*, 2006; Jang *et alii*, 2007; Luengo *et alii*, 2007; Chu & Alwan, 2009; Bartosek, 2010; Huang & Lee, 2012; Chu & Alwan, 2012).

Sono pochi gli studi, tra i più recenti, che hanno eseguito valutazioni piuttosto complete che si basino, inoltre, su corpora scaricabili liberamente (de Cheveigné & Kawahara 2002; Camacho, 2007; Wang & Loizou 2012). Questi studi utilizzano molto spesso nella valutazione una singola metrica che misura un unico tipo di errore, non considerando o considerando parzialmente l'intero panorama di indicatori sviluppati a partire dal pionieristico lavoro di Rabiner e colleghi (1976), e quindi, a mio avviso, i risultati ottenuti sembrano essere piuttosto parziali.

Ci sembra quindi rilevante eseguire una valutazione completa della maggior parte dei PDA, con particolare attenzione a quelli disponibili liberamente e a quelli frequentemente utilizzati dalla comunità scientifica, misurando le prestazioni di questi sistemi con un'ampia gamma di misure quantitative.

2. I GOLD STANDARD

La valutazione si è avvalsa di due corpora considerati come *gold standard*, entrambi disponibili liberamente e largamente utilizzati in letteratura nella valutazione dei PDA:

- *Keele Pitch Database* (Plante *et alii*, 1995): è composto da 10 locutori, 5 maschi e 5 femmine, che leggono, in ambiente controllato, un piccolo brano bilanciato in lingua inglese (*'North Wind story'*). Il corpus contiene anche l'output di un laringografo, dal quale è possibile stimare con precisione il valore di F0.
- *FDA* (Bagshaw *et alii*, 1993): è un piccolo corpus contenente 5' di registrazione divisi in 100 enunciati, letti da due locutori, un maschio e una femmina, particolarmente ricchi di fricative sonore, nasali, liquide e glide, suoni particolarmente problematici da analizzare da parte dei PDA. Anche in questo caso il gold standard per i valori di F0 è stimato a partire dall'output del laringografo e la lingua di riferimento è l'inglese.

3. LE METRICHE UTILIZZATE NELLA VALUTAZIONE

Nella predisposizione dei meccanismi di valutazione è necessario, oltre alla definizione di un opportuno gold standard, introdurre idonee misure quantitative di performance che siano in grado di cogliere i differenti aspetti critici del problema in esame.

In (Rabiner *et alii*, 1976) viene stabilito di fatto un primo standard per le misure di valutazione dei PDA, standard utilizzato da molti altri dopo di lui (es. Chu & Alwan, 2009).

Se $E_{voi \rightarrow unv}$ e $E_{unv \rightarrow voi}$ rappresentano rispettivamente il numero di frame erroneamente classificati tra *voiced* e *unvoiced* e viceversa, mentre E_{f0} rappresenta il numero di frame *voiced* nei quali il valore di pitch prodotto dal PDA differisce dal gold standard per più di 16Hz, allora possiamo definire:

RabGPE (*[Rabiner] Gross Pitch Error*) $= E_{f0} / N_{voi}$
RabVDE (*[Rabiner] Voiced Detection Error*) $= (E_{voi \rightarrow unv} + E_{unv \rightarrow voi}) / N_{frame}$

dove N_{voi} è il numero di frame *voiced* nel gold standard e N_{frame} il numero di frame nei quali è suddiviso l'enunciato.

Questi indicatori, presi singolarmente o in coppia, sono stati utilizzati in un gran numero di lavori per valutare le performance di PDA. I due indicatori, tuttavia, misurano errori molto diversi tra loro ed è possibile ottimizzare il comportamento di un PDA rispetto ad un indicatore a discapito dei risultati ottenibili utilizzando l'altro. I lavori che misurano le per-

formance utilizzando un solo indicatore, di solito il RabGPE, valutano solamente una parte del problema e difficilmente forniscono un'immagine fedele del comportamento del PDA. D'altra parte considerare entrambe le misure porta a una difficile comparazione dei risultati che risultano espressi come coppie di valori.

Chu & Alwan (2009) hanno definito una versione leggermente differente di E_{f0} nel quale una stima di F0 viene considerata errata quando differisce dal gold standard per più del 20%, hanno cioè considerato una soglia dinamica anziché fissarla a 16Hz come nel caso di (Rabiner *et alii*, 1976). Chiameremo questo indicatore **GPE20**.

Per cercare di ovviare a questi problemi, Lee e Ellis (2012) hanno suggerito metriche leggermente differenti, che consentono di essere combinate in un unico indicatore:

VE (*Voiced Error*) $= (E_{f0} + E_{voi \rightarrow unv}) / N_{voi}$
UE (*Unvoiced Error*) $= E_{unv \rightarrow voi} / N_{unv}$
PTE (*Pitch Tracking Error*) = (VE+UE) / 2

dove N_{unv} è il numero di frame *unvoiced* contenuti nel gold standard.

Tuttavia, tentare di interpretare i risultati ottenuti da un PDA alla luce della misura PTE risulta piuttosto complesso: non è infatti immediato identificare dal risultato ottenuto la fonte di errore maggiormente rilevante.

Alla luce di quanto detto finora, ci sembra opportuno introdurre una nuova misura di performance che sia in grado di catturare in modo semplice le prestazioni di un PDA in un unico indicatore, chiaro e che consideri equamente rilevanti tutte le tipologie di errore possibili. Definiamo quindi, in modo simile alla definizione di *Word Error Rate* utilizzata nel riconoscimento automatico della lingua parlata (*Automatic Speech Recognition*), il *Pitch Error Rate* come:

$$\mathbf{PER} = (E_{f0} + E_{voi \rightarrow unv} + E_{unv \rightarrow voi}) / N_{frame}$$

Questa misura somma, senza privilegiare o ridurre il contributo di alcuna componente, tutte le tipologie d'errore possibili nell'elaborazione di un frame, consentendo, a nostro avviso, un'interpretazione più semplice dei risultati ottenuti.

4. GLI ALGORITMI CONSIDERATI

La tabella 1 elenca gli algoritmi compresi nella valutazione e l'implementazione considerata. Nella scelta, oltre a includere i programmi maggiormente utilizzati dagli studiosi della nostra comunità scientifica, si è scelto di privilegiare quelli disponibili gratuitamente in formato *open-source*.

Per la valutazione sono stati utilizzati i parametri standard, o di *default*, per ogni algoritmo considerato, imponendo unicamente a tutti gli algoritmi uno *shift* tra i frame di 0.01 sec. per poter confrontare i risultati coi dati contenuti nei *gold standard* senza applicare operazioni di interpolazione o ricampionamento che avrebbero immancabilmente introdotto un certo grado di arbitrarietà.

Algoritmo	Implementazione	Rif. Bibliografico
FXANAL	SFS v4.8/win (http://www.phon.ucl.ac.uk/resource/sfs/)	(Secrest & Doddington, 1983)
ESRPD	Edimburgh Speech Tools (pda) (http://www.cstr.ed.ac.uk/projects/speech_tools/)	(Bagshaw *et alii*, 1993; Medan *et alii*, 1991)
PRAAT	Praat v5.3.35 ("*To Pitch (ac)*", l'algoritmo risultato migliore tra tutti i disponibili, con o senza l'applicazione dell'opzione "*Kill octave jump*")	(Boersma, 1993)
RAPT	ESPS get_f0, Snack/Wavesurfer, SFS v4.8/win, e altri…	(Talkin, 1995)
SHR/ Papyrus	Implementazione all'interno del progetto Papyrus (http://www.ict-papyrus.eu)	(Sun, 2000)
YIN	http://audition.ens.fr/adc/sw/yin.zip	(de Cheveigné & Kawahara, 2002)
WU	http://www.cse.ohio-state.edu/pnl/shareware/wu-tsap03/	(Wu *et alii*, 2003)
SWIPE'	SPTK, v3.5 (http://sp-tk.sourceforge.net/)	(Camacho, 2007)
YAAPT	http://ws2.binghamton.edu/zahorian/yaapt.htm	(Zahorian & Hu, 2008)
PEFAC	VoiceBox per Matlab (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html)	(Gonzalez & Brookes, 2011)
SAcC	http://labrosa.ee.columbia.edu/projects/SAcC/	(Lee & Ellis, 2012)

Tabella 1: Gli algoritmi utilizzati nella valutazione e le rispettive implementazioni (link verificati il 23/5/2013).

5. RISULTATI

La Tabella 2 mostra i valori di performance ottenuti dai vari algoritmi rispetto a tutte le misure considerate (per entrambi i gold standard utilizzati nella valutazione), ordinati rispetto alla nuova metrica proposta (**PER**).

Le migliori prestazioni si evidenziano per gli algoritmi RAPT, PRAAT, SWIPE', YAAPT e SAcC, che risultano essere più stabili e performanti sia rispetto al PER che agli altri indicatori. Mentre per gli altri algoritmi più performanti non si rilevano differenze rispetto ai due gold standard utilizzati, PRAAT mostra performance maggiori nella classificazione degli enunciati contenuti nel corpus FDA rispetto al gold standard KEELE; probabilmente, considerando che la differenza principale tra i due corpora riguarda la lunghezza degli enunciati, è possibile concludere che le prestazioni di PRAAT sono in qualche modo influenzate negativamente dalla quantità di dati da elaborare.

Nella valutazione dei risultati ottenuti ci sembra opportuno studiare le tipologie d'errore che i 5 algoritmi più performanti compiono nel processo di identificazione di F0; la Tabella 3 mostra come si distribuisce l'errore totale (PER) rispetto alle tre tipologie di errore che compongono la sua definizione (E_{f0}, $E_{voi \rightarrow unv}$, $E_{unv \rightarrow voi}$). Abbiamo tentato di evidenziare visivamente l'entità dell'errore introducendo nella tabella una sfumatura di grigi, col significato più scuro – errore più rilevante.

FDA corpus

PDA	PER	GPE20	RabGPE	RabVDE	PTE	VE	UE
PRAAT	0.05992	0.00817	0.02703	0.05070	0.05742	0.07480	0.04004
RAPT	0.07128	0.01642	0.03958	0.05723	0.06523	0.06824	0.06222
SWIPE'	0.07517	0.00241	0.02158	0.06614	0.07765	0.12483	0.03047
YAAPT	0.07929	0.02102	0.05184	0.06153	0.06841	0.05890	0.07792
PRAAT+koj	0.08401	0.07095	0.08961	0.05070	0.08645	0.13286	0.04004
SAcC	0.08541	0.00626	0.02573	0.07723	0.09327	0.14046	0.04609
WU	0.10327	0.01087	0.03518	0.09117	0.08843	0.05324	0.12363
YIN	0.11228	0.01674	0.03715	0.10019	0.10990	0.13009	0.08972
FXANAL	0.11657	0.02881	0.05672	0.09791	0.11675	0.14586	0.08765
ESRPD	0.11801	0.06126	0.07140	0.09760	0.14673	0.28213	0.01132
PEFAC	0.14273	0.05188	0.09448	0.10896	0.12072	0.10295	0.13849
SHR/Papyrus	0.18300	0.02958	0.06000	0.16368	0.16964	0.14322	0.19605

KEELE corpus

PDA	PER	GPE20	RabGPE	RabVDE	PTE	VE	UE
RAPT	0.07441	0.01792	0.03283	0.05866	0.06811	0.07117	0.06505
SWIPE'	0.08097	0.00290	0.00885	0.10864	0.11057	0.19941	0.02173
YAAPT	0.08139	0.01948	0.03307	0.06548	0.07462	0.06828	0.08096
PRAAT	0.09297	0.01317	0.02129	0.08390	0.08990	0.12828	0.05151
SAcC	0.09836	0.01377	0.02067	0.08981	0.09538	0.14503	0.04574
YIN	0.12689	0.01431	0.02271	0.11710	0.12501	0.14651	0.10352
WU	0.12801	0.02791	0.03598	0.11190	0.12572	0.11393	0.13751
FXANAL	0.14714	0.04124	0.05870	0.12097	0.13907	0.13365	0.14450
ESRPD	0.18417	0.04690	0.05545	0.16225	0.17789	0.34164	0.01413
PRAAT+koj	0.20324	0.25643	0.26486	0.08894	0.20158	0.34324	0.05991
SHR/Papyrus	0.25158	0.05248	0.06974	0.21995	0.24428	0.13729	0.35128
PEFAC	0.29194	0.10602	0.18897	0.21376	0.25881	0.28114	0.23647

Tabella 2: Le performance ottenute dagli algoritmi considerati (per entrambi i gold standard utilizzati nella valutazione) ordinate rispetto alla nuova metrica proposta (PER).

Dalla Tabella 3 si evidenziano comportamenti piuttosto differenti tra i migliori PDA: gli errori commessi dai vari algoritmi sembrano infatti distribuirsi tra le tre tipologie di errore in modo non uniforme e con configurazioni differenti tra i PDA.

Potrebbe quindi essere utile considerare la possibilità di combinare i contributi dei differenti algoritmi che, essendo diversi tra loro, potrebbero innescare un circolo virtuoso nel quale il comportamento di un PDA potrebbe correggere gli errori fatti dagli altri e viceversa, producendo, globalmente, una diminuzione dell'errore totale (PER).

Una possibilità per combinare i valori di vari algoritmi è quella di considerare come stima del valore del pitch in uno specifico frame la mediana dei valori calcolati da un numero dispari di algoritmi diversi (nei nostri esperimenti 3 o 5).

La Tabella 4 mostra alcune delle combinazioni più performanti tra i 5 migliori PDA.

FDA corpus

	PER	E_{f0}	$E_{voi \rightarrow unv}$	$E_{unv \rightarrow voi}$
PRAAT	0.05992	0.00922	0.02687	0.02383
RAPT	0.07128	0.01406	0.02091	0.03632
SWIPE'	0.07517	0.01139	0.01691	0.04686
YAAPT	0.07840	0.01767	0.01597	0.04476
SAcC	0.08541	0.00818	0.05072	0.02651

KEELE corpus

	PER	E_{f0}	$E_{voi \rightarrow unv}$	$E_{unv \rightarrow voi}$
RAPT	0.07441	0.01575	0.02689	0.03177
YAAPT	0.07991	0.01479	0.02653	0.03859
SWIPE'	0.08097	0.00748	0.04218	0.03131
PRAAT	0.09297	0.00907	0.05855	0.02534
SAcC	0.09836	0.00855	0.06714	0.02268

Tabella 3: Entità degli errori compiuti dai migliori 5 PDA rispetto alle tre tipologie di errore che compongono la definizione del PER.

FDA corpus

	PER	E_{f0}	$E_{voi \rightarrow unv}$	$E_{unv \rightarrow voi}$
SAcC-SWIPE1-PRAAT	0.05356	0.00860	0.02337	0.02159
RAPT-SWIPE1-YAAPT-SAcC-PRAAT	0.05534	0.01031	0.01926	0.02576
RAPT-SWIPE1-PRAAT	0.05848	0.01033	0.01914	0.02902
RAPT-SWIPE1-YAAPT	0.06017	0.01212	0.01407	0.03398

KEELE corpus

	PER	E_{f0}	$E_{voi \rightarrow unv}$	$E_{unv \rightarrow voi}$
RAPT-SWIPE1-YAAPT-SAcC- PRAAT	0.05914	0.00848	0.03025	0.02040
RAPT-SWIPE1-YAAPT	0.06109	0.00980	0.02442	0.02687
RAPT-SWIPE1- PRAAT	0.06120	0.00651	0.03326	0.02143
SAcC-SWIPE1- PRAAT	0.06370	0.00522	0.04183	0.01665

Tabella 4: Prestazioni ottenute combinando, considerando il valore mediano del pitch, 3 o 5 tra i PDA più performanti.

Dalla Tabella 4 emerge piuttosto chiaramente come la combinazione di più algoritmi diversi con il metodo delle mediane riduca sensibilmente il PER e tutti gli altri tipi di errore, e, in particolare, come riduca in modo rilevante E_{f0}, ovvero l'errore di stima dei valori di F0 su frame considerati voiced, i cosiddetti errori di *halving* e *doubling*.

CONCLUSIONI

Questo lavoro ha presentato una valutazione oggettiva di un gran numero di algoritmi per l'estrazione automatica del valore della frequenza fondamentale nella lingua parlata utilizzando un cospicuo insieme di metriche differenti.

Dopo aver analizzato nel dettaglio le prestazioni dei PDA considerati, abbiamo proposto e verificato come la combinazione dei risultati ottenuti da vari algoritmi possa migliorare le performance totali consentendo di ridurre l'errore di stima di F0 sfruttando le migliori possibilità di ogni algoritmo.

BIBLIOGRAFIA

Bagshaw P. C., Hiller S. M. & Jack M. A. (1993), Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching, in Proceedings of Eurospeech '93, Berlin, 1003–1006.

Bartosek, J. (2010), Pitch Detection Algorithm Evaluation Framework, in Proceedings of 20th Czech-German Workshop on Speech Processing, Prague, 118–123.

Boersma P. (1993), Accurate short-term analysis of the fundamental and the harmonics-to-noise ratio of a sampled sound, in Proceedings of the Institute of Phonetic Sciences, University of Amsterdam, 17, 97–110.

Camacho A., (2007), SWIPE: A sawtooth waveform inspired pitch estimator for speech and music, PhD Thesis, University of Florida.

Chu, W. & Alwan A. (2009), Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend, in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP2009, 3969–3972.

Chu, W. & Alwan A. (2012), SAFE: A Statistical Approach to F0 Estimation Under Clean and Noisy Conditions, IEEE Transactions on Audio, Speech, and Language Processing, 20, 933–944.

de Cheveigné A. & Kawahara H. (2002), YIN, a fundamental frequency estimator for speech and music, Journal of the Acoustical Society of America, 111, 1917–30.

Gonzalez S. & Brookes, M. (2011), A pitch estimation filter robust to high levels of noise (PEFAC), in Proceeedings of 19th European Signal Processing Conference - EUSIPCO 2011, Barcelona, 451–455.

Huang, F. & Lee, T. (2012), Robust Pitch Estimation Using l1-regularized Maximum Likelihood Estimation, in Proceedings of 13th Annual Conference of the International Speech Communication Association – Interspeech 2012, Portland (OR).

Jang, S.J., Choi, S.H., Kim, H.M., Choi, H.S. & Yoon Y.R. (2007), Evaluation of performance of several established pitch detection algorithms in pathological voices, In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society - EMBC, Lyon, 620–623.

Kotnik, B., Höge, H. & Kacic, Z. (2006), Evaluation of Pitch Detection Algorithms in Adverse Conditions, in Proceedings of Speech Prosody 2006, Dresden, PS2–8–83.

Lee, B.S. & Ellis, D. (2012), Noise Robust Pitch Tracking by Subband Autocorrelation Classification, in Proceedings of 13th Annual Conference of the International Speech Communication Association – Interspeech 2012, Portland (OR).

Luengo, I., Saratxaga, I., Navas, E., Hernaez, I., Sanchez, J. & Sainz I. (2007), Evaluation of Pitch Detection Algorithm under Real Conditions, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2007, Honolulu, Hawaii, 4, 1057–1060.

Medan, Y., Yair, E. & Chazan D. (1991). Super resolution pitch determination of speech signals, IEEE Transactions on Signal Processing, 39, 40–48.

Plante, F., Ainsworth, W.A. & Meyer, G. (1995), A Pitch Extraction Reference Database, in Procedings of Eurospeech'95, Madrid, 837–840.

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. & McGonegal C.A. (1976), A Comparative Performance Study of Several Pitch Detection Algoritms, IEEE Transaction on Acoustics, Speech and Signal Processing, 24, 399–418.

Secrest, B. & Doddington, G. (1983), An integrated pitch tracking algorithm for speech systems, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP83, 1352–1355.

Sun X. (2000), "A Pitch Determination Algorithm Based on Subharmonic-to-Harmonic Ratio", in Proceedings of the 6th International Conference of Spoken Language Processing - ICSLP2000, Beijing, 676–679.

Talkin D. (1995), A robust algorithm for pitch tracking (RAPT), in Speech Coding and Synthesis (W. B. Kleijn & K. K. Paliwal, editors), New York: Elsevier, 495–518.

Veprek, P. & Scordilis, M.S. (2002), Analysis, enhancement and evaluation of five pitch determination techniques, Speech Communication, 37, 249–270.

Wang, D. & Loizou, P.C. (2012), Pitch Estimation Based on Long Frame Harmonic Model and Short Frame Average Correlation Coefficient, in Proceedings of 13th Annual Conference of the International Speech Communication Association – Interspeech 2012, Portland (OR).

Wu, M., Wang, D.L. & Brown, G.J. (2003), A multipitch tracking algorithm for noisy speech, IEEE Transactions on Speech and Audio Processing, 11, 229–241.

Zahorian, S.A., Hu, H. (2008), A Spectral/temporal method for Robust Fundamental Frequency Tracking, Journal of the Acoustical Society of America, 123, 4559–4571.