IDENTIFICAZIONE AUTOMATICA DELLA PROMINENZA FRASALE NEL LINGUAGGIO PARLATO

Fabio Tamburini
Università di Bologna
f.tamburini@cilta.unibo.it

SOMMARIO

Questo contributo presenterà uno studio che si inserisce nel filone delle indagini nell'ambito dei fenomeni prosodici e della loro identificazione con metodi automatici, proponendosi di indagare le complesse relazioni che intercorrono, a diversi livelli, tra il fenomeno percettivo della prominenza prosodica e i fenomeni di tipo acustico che possono essere desunti dalla componente fonetico/acustica degli enunciati. Con questo lavoro intendiamo mostrare come, con un'attenta misurazione dei parametri acustici che supportano il fenomeno della prominenza e identificando opportune relazioni tra essi e i fenomeni prosodici oggetto di questo studio, sia possibile costruire un sistema automatico capace di identificare le sillabe prominenti nel parlato continuo con livelli di accordo tra sistema e annotatori umani confrontabili con quelli ottenibili da esperti del settore utilizzando unicamente informazioni ricavabili dalla *waveform* dell'enunciato.

1. INTRODUZIONE

Lo studio della prosodia ha seriamente influenzato i lavori riguardanti la produzione di sistemi automatici per l'elaborazione del linguaggio parlato. Numerosi studi hanno incluso metodi per la gestione e l'utilizzo dei fatti prosodici in relazione ai vari livelli dell'analisi linguistica eseguita con metodi automatici, arrivando ad affermare che l'introduzione dell'analisi prosodica in tali sistemi risulta essere il passo più importante, allo stato dell'arte, per migliorarne le prestazioni.

Nell'ambito dei sistemi per il riconoscimento automatico del linguaggio parlato l'indagine di studiosi come (Batliner *et alii*, 2001; Hastie *et alii*, 2001; Hieronymous *et alii*, 1992; Shriberg, Stolcke, 2001) ha mostrato come tali sistemi possano migliorare sensibilmente le loro prestazioni nelle fasi di riconoscimento introducendo moduli per la disambiguazione basata su informazioni prosodiche, anche se, come sottilinea Wightman (2002), ben poche di queste metodologie sono state applicate in sistemi commerciali. Simili conclusioni sono state ottenute dagli studiosi che hanno operato indagini nell'ambito della comprensione automatica del linguaggio parlato (Beckman, Venditti, 2000; Gallwitz *et alii*, 2002; Noth *et alii*, 2000; Shriberg *et alii*, 2000).

Un ambito di indagine che ha certamente rilevanti connessioni con il trattamento delle informazioni prosodiche riguarda i sistemi per il *Text To Speech*. In questi sistemi la naturalezza della voce sintetica è un requisito fondamentale e irrinunciabile per un corretto e piacevole ascolto da parte di un ascoltatore umano e anche per una corretta comprensione del messaggio. I lavori di studiosi come (Bulyko *et alii*, 1999; Mixdorff, Jokisch, 2003; Portele, Heuft, 1997; Taylor, 2000; Wightman *et alii*, 2000) sono solo pochi esempi di studi in questo settore che trattano questo tipo di fenomeni.

Particolarmente interessante per gli studi sulla prosodia risulta essere la produzione di risorse linguistiche etichettate rispetto ai fenomeni prosodici (Beckman, Venditti, 2000; Campione, Veronis, 1998; Hirst, 2001). Queste risorse sono estremamente preziose sia per

lo sviluppo di modelli teorici della prosodia sia come fonti di informazioni statistiche per lo sviluppo dei sistemi automatici descritti in precedenza. In questo senso, strumenti automatici capaci di individuare ed etichettare questi fenomeni risultano estremamente utili.

Di estremo interesse sono i contributi di (Hirschberg, Avesani, 2000; Warren, 1996) riguardo alle complesse interazioni tra prosodia e disambiguazione sintattica e semantica. In particolare, la rassegna di Warren esamina i vari aspetti delle complesse interazioni tra la prosodia e i differenti livelli di analisi linguistica, sottolineando le fondamentali connessioni con gli aspetti cognitivi della produzione e della percezione dei fenomeni prosodici. Anche il lavoro di (Noth *et alii*, 2000) in seno al progetto VERBMOBIL, probabilmente la punta progettuale più avanzata nel settore dell'integrazione tra prosodia e sistemi di ASR/ASU, mostra come una stretta integrazione di modelli computazionali in grado di gestire informazioni prosodiche nell'ambito di compiti di disambiguazione sintattico/semantica consenta di ottenere prestazioni decisamente superiori nella comprensione automatica del linguaggio parlato.

Sistemi di *Computer Aided Language Learning* dedicati alla pronuncia possono analizzare la produzione di un discente mostrando, confrontando, correggendo e valutando le sue prestazioni in relazione alle corrispondenti produzioni di un parlante nativo. Anche in questo settore sono numerosissimi gli studi per lo sviluppo di sistemi automatici per la didattica della pronuncia che includono moduli specifici per l'insegnamento di una corretta gestione dei fenomeni prosodici (Bagshaw, 1994; Chun, 1998; Delmonte, 2000; Menzel *et alii*, 2001; Neumeyer *et alii*, 1998).

Lo studio presentato in questo contributo si inserisce nel filone delle indagini nell'ambito dei fenomeni prosodici e della loro identificazione con metodi automatici, proponendosi di sondare le complesse relazioni che intercorrono, a diversi livelli, tra il fenomeno percettivo della prominenza prosodica e i fenomeni di tipo acustico che possono essere desunti dalla componente fonetico/acustica degli enunciati. La prominenza, che possiamo preliminarmente indicare come l'enfasi posta da un locutore su specifiche parti dell'enunciato, risulta essere uno dei fenomeni prosodici maggiormente interessanti e anche uno di quelli maggiormente studiati, proprio per l'importanza che riveste nell'ambito dei processi comunicativi umani. Nonostante la grande quantità di studi nel settore, soprattutto dal punto di vista linguistico, il problema risulta ancora estremamente attuale e tutt'altro che risolto.

Sembra opportuno discutere brevemente quali sono i vincoli e le condizioni al contorno nei quali si inserisce il lavoro. Riteniamo opportuno che tale indagine utilizzi come uniche informazioni i parametri derivabili direttamente, anche se in modo estremamente articolato, dall'espressione sonora dell'enunciato, ovvero nella sua trasposizione digitale realizzata campionando opportunamente il segnale. Il modello che proporremo e l'algoritmo che implementerà il riconoscimento della prominenza non saranno basati su fonti di informazioni alternative, quali trascrizioni degli enunciati, sia ortografiche che fonetiche, risorse linguistiche etichettate dal punto di vista fonetico, fonologico, prosodico, e nemmeno risorse che contengano informazioni di tipo segmentale sugli enunciati. L'unica informazione fornita all'algoritmo di annotazione sarà la digitalizzazione dell'enunciato (*waveform* o oscillogramma).

Restringere il dominio delle informazioni fruibili in modo così netto elimina, di fatto, tutti quei modelli di analisi parametrica che utilizzano fasi di apprendimento basate su dati autentici: *hidden Markov model* (HMM), reti neurali, alberi di decisione probabilistici, classificatori basati su logiche *fuzzy*, ecc., richiedendo pesanti fasi di apprendimento, che sfruttano appieno le informazioni di vari tipi di risorse linguistiche etichettate, sono modelli

teorici esclusi a priori da questo lavoro. La richiesta di possedere risorse linguistiche accuratamente etichettate in modo manuale per poter sviluppare tali modelli ci è sembrata infatti una richiesta estremamente vincolante, essendo tali risorse abbastanza rare, estremamente costose, difficili da realizzare autonomamente e, soprattutto, avendo richiesto nell'ambito del loro sviluppo l'applicazione di modelli teorici o procedurali che potrebbero non corrispondere agli obiettivi del nostro lavoro, potrebbero fornire dati in qualche modo distorti e, una volta implementata e realizzata la fase di apprendimento, il sistema risulta vincolato alla tipologia di dati e all'intrinseca natura delle risorse utilizzate. Impostare i parametri di un modello utilizzando risorse linguistiche di una specifica lingua limita aprioristicamente ogni possibile conclusione nell'ambito della lingua utilizzata, non consentendo, prima di tutto dal punto di vista metodologico, alcuna generalizzazione inter-linguistica. Viceversa, la scelta di non utilizzare alcun tipo di informazione non derivabile dalla porzione sonora dell'enunciato può consentire lo sviluppo di modelli che prescindono completamente da dati etichettati non restringe a priori le conclusioni che si otterranno a una specifica lingua quindi, anche se il fenomeno della prominenza non può dirsi legato, da un punto di vista interlinguistico, ai parametri prosodici con relazioni costanti nella loro natura, gli algoritmi e le metodologie che proporremo potranno essere facilmente adattate ad altri contesti linguistici, senza richiedere che siano presenti risorse linguistiche etichettate nelle varie lingue trattate.

Nell'esame dei contributi scientifici sull'argomento si riscontrano notevoli problemi di uniformità terminologica, soprattutto nell'identificazione dei fenomeni che fanno parte dell'ambito di indagine. Numerosi studiosi, tra cui (Bertinetto, 1981; Taylor, 1992; Wightman, Ostendorf, 1994), sottolineano più volte come negli studi sui fenomeni prosodici ci sia una rilevante eterogeneicità tra i termini utilizzati per indicare lo stesso fenomeno, che, da studio a studio e nel tempo, tendono ad assumere differenti connotazioni e riferimenti. Il problema è ulteriormente complicato dal fatto che la quasi totalità della bibliografia del settore è scritta in lingua inglese. Le corrispondenze terminologiche, pur essendo abbastanza chiare e definite, creano in taluni casi assonanze tra termini corrispondenti a traduzioni che in ambito prosodico assumono significati differenti a seconda della lingua alla quale appartengono (es. accento *vs accent*). La scelta è stata quella di limitare al minimo indispensabile l'uso della terminologia tratta dall'ambito prosodico, mantenendo per alcuni di tali termini la notazione anglosassone, peraltro sempre evidenziati tipograficamente, evitando l'introduzione di sinonimi o equivalenti traduttivi nel corso dell'esposizione.

Questo studio si è concentrato principalmente sulla lingua inglese americana, sia per i numerosi spunti metodologici rilevabili in letteratura sia per la possibilità di confronto con altri lavori. L'esame della bibliografia del settore, specialmente in riferimento a problematiche tecnologiche, ha immediatamente evidenziato la presenza di un *corpus* costantemente citato negli studi di questo tipo, che è stato largamente utilizzato per le valutazioni di prestazioni degli algoritmi e delle tecniche presentate nel settore dell'elaborazione del linguaggio parlato. Il *corpus* TIMIT contiene 6300 enunciati di linguaggio parlato letto di inglese americano (Garofolo *et alii*, 1993) ai quali sono stati associati la trascrizione ortografica, la segmentazione allineata rispetto alle parole contenute e la segmentazione fonetica allineata.

2. LA PROMINENZA FRASALE: ASPETTI LINGUISTICI

I vari studi nel settore mostrano un panorama piuttosto vario della materia e, pur nell'evoluzione della conoscenza che si è avuta tra i primi studi negli anni '50 ai giorni

nostri, non sembra che vi siano teorie di riferimento universalmente riconosciute. Al contrario, vi sono numerosi lavori che, partendo spesso da prospettive diverse, ottengono conclusioni generali contraddittorie. Se queste contraddizioni erano evidenti negli anni '50-'70, specialmente tra scuole di pensiero diverse, anche negli studi più recenti persistono punti di vista molto diversi tra i vari studiosi della materia.

Nonostante questa apparente mancanza di riferimenti precisi a teorie linguistiche che rendano conto compiutamente del fenomeno della prominenza, soprattutto nei confronti dei parametri fonetico/acustici, pare comunque opportuno, e necessario, chiarire il quadro di riferimento linguistico di questo lavoro.

Le modalità di realizzazione fonetico/acustica della prominenza tendono a dividere le lingue umane in classi di equivalenza; il nostro interesse sarà principalmente focalizzato su lingue *stress accented*, per la complessità dei fenomeni che realizzano la prominenza e soprattutto per l'interesse che rivestono le lingue contenute in questa classe, come, per esempio, l'inglese, l'italiano, l'olandese e lo spagnolo…

A livello dei fenomeni linguistici correlati al fenomeno percettivo della prominenza, sembra esserci una sostanziale convergenza degli studi, specialmente fra quelli più moderni, nell'individuare essenzialmente due "attori" principali. I *pitch accent*, individuati partendo dall'analisi dei profili intonativi identificando i movimenti del *pitch* al loro interno, costituiscono il primo fenomeno linguistico che certamente è in grado di indurre percettivamente un'idea di prominenza nell'ascoltatore (Beckman, 1986; Bolinger, 1958; Fry, 1958; Sluijter, van Heuven, 1996a; 1996b; 1997; Streefkerk, 1996; Taylor, 2000). Il secondo, non meno importante del primo, riguarda il concetto di *stress* frasale ed è basato essenzialmente sul fenomeno dello *stress* lessicale delle unità coinvolte nell'enunciato e sull'interazione di tali unità, a livello di rinforzo o indebolimento dei loro *stress* lessicali, per arrivare alla costruzione dell'andamento ritmico dell'enunciato (Bagshaw, 1994; Beckman, 1986; Sluijter, van Heuven, 1996a; 1996b; 1997; Streefkerk, 1996).

A livello fonetico/acustico, le lingue *stress accented* utilizzano numerosi parametri che interagiscono in vari modi per veicolare la percezione della prominenza tra il parlante e l'ascoltatore. Anche se non vi è accordo totale tra i vari studiosi,un gruppo rilevante di studi, tra i quali risalta l'influente lavoro di Sluijter e van Heuven (1996a; 1996b; 1997), è concorde nell'indicare l'intensità e i movimenti nei profili di F0 come i principali correlati acustici del *pitch accent*, mentre la durata della sillaba e l'intensità in particolari bande spettrali identificano piuttosto chiaramente le sillabe *stressed* (Anastasakos *et alii*, 1995; Bagshaw, 1994; Heldner, 2001; Streefkerk, 1996).

Uno dei nodi principali da sciogliere riguarda l'interazione e le connessioni tra lo *stress* e il *pitch accent*. Sluijter e van Heuven (1996a), descrivendo questi due fenomeni linguistici come i principali fenomeni legati alla prominenza, evidenziano chiaramente questo punto. Anche altri studiosi ritengono che la connessione tra *stress* e *pitch accent* sia, se pur con sfumature differenti, quella illustrata da Sluijter e van Heuven (Beckman, 1986; Taylor, 1992), ovvero ascrivono allo *stress* lessicale il ruolo di supporto all'introduzione di *pitch accent*, motivandolo, per esempio, con considerazioni di tipo fonetico/acustico riguardanti la necessità che la durata della sillaba *pitch accented* risulti aumentata, proprio per poter ospitare temporalmente l'accento intonativo.

Queste considerazioni, pur risultando estremamente ragionevoli, almeno da un punto di vista acustico, non appaiono tuttavia essere sufficientemente generali da giustificare l'introduzione di una gerarchia stretta tra i fenomeni linguistici che supportano la prominenza, imponendo, in modo radicale, che un *pitch accent* debba forzatamente presentarsi in riferimento a una sillaba effettivamente *stressed*. L'esame di enunciati

autentici mostra come grandi movimenti nel profilo del *pitch*, che generano conseguentemente un *pitch accent* intenso, non siano di fatto associati a sillabe particolarmente lunghe, o meglio, vengono associati a sillabe che non presentano un aumento in durata e enfasi spettrale rilevante, e che quindi non possono essere effettivamente identificate come *stressed* da punto di vista acustico.

Un'altra considerazione rilevante riguarda il lavoro di alcuni studiosi come Gimson (1980): essi osservano come l'andamento ritmico della lingua inglese e in particolare la posizione delle sillabe prominenti, pur essendo abbastanza prevedibile, mostri una struttura molto più flessibile di quello che potrebbe sembrare, modificando la visione di andamento ritmico che vorrebbe una sillaba prominente per ogni parola in corrispondenza dello *stress* lessicale, almeno per quel che riguarda le parole che portano significato. Le motivazioni possono riguardare il contesto nel quale è inserita la parola, particolari scopi comunicativi, o tendenze comuni ad un gruppo di parlanti.

Queste considerazioni, anche se di tipo qualitativo e certamente non conclusive, il fatto che non vi è un totale accordo tra gli studiosi sulle connessioni tra *stress* e *pitch accent* e il forte impatto che avrebbe la scelta di gerarchizzare strettamente i due fenomeni linguistici sull'obiettivo di questo lavoro, portano a concludere che non risulta opportuno imporre questa scelta forte e utilizzare le informazioni relative allo *stress* lessicale delle parole come unica modalità di identificazione delle sillabe in grado di supportare un *pitch accent* nel riconoscimento della prominenza. D'altra parte le informazioni relative alla trascrizione degli enunciati non è disponibile e nemmeno lo è una segmentazione in unità fonetiche che arrivi ad identificarne completamente la natura, quindi, nelle ipotesi di partenza sulle quali si basa questo lavoro, non risulta possibile accedere ad informazioni relative allo stress lessicale nella lingua in esame.

Nonostante una chiara convergenza degli studiosi nel definire il fenomeno della prominenza come un fenomeno percettivo che presenta una gradazione continua (Chomsky, Halle, 1968; Fant *et alii*, 2000; Ladd *et alii*, 1994; Portele e Heuft 1997; Taylor, 2000), sono numerosi gli studi e le teorie fonologiche che propongono una classificazione di questo fenomeno, e dei fenomeni linguistici ed esso correlati, da un punto di vista discreto.

Numerosi studiosi (Campione, Veronis, 1998; Taylor, 1992; 2000; Wightman, 2002) hanno evidenziato come trattazioni discrete siano sostanzialmente inadeguate, da vari punti di vista, per affrontare fenomeni intrinsecamente continui. La difficoltà ad adattarle a lingue differenti da quelle per la quale sono state introdotte, i problemi che pongono a metodi automatici per la classificazione, il disaccordo di alcuni studiosi sulle scelte che tali metodi impongono, sono solo alcune delle osservazioni negative che tali studiosi muovono a questi metodi. Prima Campione e Veronis e poi Taylor, propongono di inserire metodologie di identificazione di questi fenomeni basate su teorie fonetico/acustiche continue come strato intermedio tra l'acustica degli enunciati e teorie fonologiche linguisticamente motivate.

Nell'ambito di questo lavoro seguiremo le indicazioni fornitaci da questi ultimi studiosi e affronteremo l'identificazione dei fenomeni in esame da un punto di vista strettamente continuo, proponendo metodi e sistemi che gestiscono le grandezze collegate a questi fenomeni evitando ogni categorizzazione, se non in estrema analisi e unicamente nell'ottica di una valutazione delle prestazioni dei sistemi automatici.

2.1 Studi sulla prominenza nelle varie lingue

La lingua inglese ha storicamente detenuto il primato di lingua maggiormente analizzata dal punto di vista prosodico, infatti molti degli studi "tradizionali" hanno avuto come oggetto la lingua sia britannica che americana. Si vedano per esempio, evitando di citare gli

studi più datati, (Beckman, 1986, 1994; Bagshaw, 1994; Sluijter, van Heuven, 1996a; 1996b; 1997; Taylor, 1992; 2000).

Negli ultimi anni, si è assistito a un fervore di studi sulla prosodia nelle altre lingue. Lingue *stress accented* come l'olandese e il tedesco hanno ricevuto l'attenzione di studiosi come (Rietveld, Kerkhoff, 2002; Sluijter, van Heuven, 1996a; 1996b; 1997; Streefkerk, 1996; 1999; Strom, 1995; van Kuijk, Boves, 1999). I risultati che presentano sono sostanzialmente in accordo con quelli discussi nelle sezioni precedenti, principalmente ricavati per la lingua inglese, mostrando come vi sia una chiara suddivisione dei due fenomeni linguistici descritti come correlati della prominenza, ovvero *stress* frasale e *pitch accent*, in special modo nei confronti dei parametri acustici che li supportano, evidenziando sostanzialmente gli stessi fenomeni fonetico/acustici applicabili alla lingua inglese.

Anche lingue *pitch accented* come lo svedese e il finlandese hanno evidenziato, negli studi di (Eriksson *et alii*, 2001; Fant *et alii*, 2000; Heldner, 1998; 2003; Suomi *et alii*, 2003) una sostanziale convergenza, nell'organizzazione dei fenomeni sia linguistici che fonetico/acustici, con le lingue *stress accented*, anche se, per quanto riguarda il finlandese, l'unità segmentale di riferimento più adeguata per analizzare il fenomeno sembra essere la mora anziché la sillaba, differenziandolo maggiormente dalle altre lingue nella identificazione dei fenomeni prosodici.

Altre lingue *pitch accented*, come per esempio il giapponese, sono state estensivamente studiate da Beckman (1986). I risultati degli esperimenti effettuati mostrano una differenziazione radicale rispetto alle lingue *stress accented*, in quanto l'uso di correlati acustici diversi dal *pitch* per segnalare fenomeni di prominenza è piuttosto raro.

Come spesso accade in linguistica, introdurre categorizzazioni tra gli oggetti di studio può essere estremamente complesso. Nel caso delle lingue studiate dal punto di vista prosodico la distinzione tra lingue *stress accented* e *pitch accented* appare tutt'altro che netta e definita. Beckman (1986) ha introdotto tale distinzione basandola sull'uso o meno di correlati acustici diversi dal *pitch* per segnalare la prominenza frasale. Di fatto, tutti gli studi descritti sembrano invece suggerire una visione molto più sfumata di tale distinzione. Sembra essere più opportuno considerare un *continuum* di possibilità tra due poli rappresentati dai prototipi proposti da Beckman per queste due categorie, ovvero la lingua inglese e la lingua giapponese rispettivamente, inserendo le altre lingue in posizioni intermedie senza disegnare nettamente il confine di separazione tra le due classi.

Una menzione particolare va senz'altro alla lingua italiana che verrà descritta, sia dal punto di vista bibliografico che dell'analisi, in una delle sezioni seguenti.

3. ANALISI FONETICO/ACUSTICA

In questo lavoro ci siamo concentrati su lingue *stress accented*, in particolare, per quel che riguarda la sperimentazione effettuata, sull'inglese americano.

Analizzando i lavori presentati nella sezione precedente emerge chiaramente come vi sia una costellazione di parametri fonetico/acustici che, attraverso le loro complesse interazioni, supportano, in ultima istanza, il fenomeno della prominenza. Appare quindi giustificato, nell'ottica di un riconoscimento automatico, analizzare i vari parametri acustici che negli studi precedenti sono stati indicati come i correlati principali del fenomeno in esame. L'analisi si dovrà occupare di discutere l'effettiva validità di tali parametri, nonché di produrre adeguate procedure per la loro determinazione in modo automatico.

Dagli studi precedenti emergono sostanzialmente tre aree di indagine, nelle quali possiamo raggruppare i parametri fonetico/acustici individuati: misure di durata temporale delle unità di riferimento (che di solito coincidono con le sillabe), misure relative ai profili

intonativi dell'enunciato e misure relative all'intensità, in senso lato, all'interno di tali unità. Le tre sezioni seguenti si occuperanno rispettivamente dell'analisi di ognuna delle aree d'indagine.

Nelle sezioni seguenti, laddove si tratteranno le metodologie proposte e si farà menzione di analisi basate su suddivisioni in frame degli enunciati, si considereranno durate dei frame di 25 msec e scostamenti temporali tra due frame successivi di 10 msec.

3.1 Identificazione delle unità di riferimento e misure di durata

Gli studi analizzati nella sezione precedente e le teorie proposte sono sostanzialmente concordi nel basare lo studio della prominenza, dal punto di vista delle unità temporali, su unità di tipo sillabico, che vengono considerate il riferimento per la misurazione di tutti i parametri fonetico/acustici correlati col fenomeno della prominenza.

Il concetto di sillaba si presenta tuttavia molto problematico dal punto di vista fonetico/acustico. Se è possibile definire univocamente la sillaba nel linguaggio scritto, dal punto di vista del linguaggio parlato lo scenario cambia radicalmente; la definizione di sillaba infatti pertiene a livelli linguistici più alti di quello considerato in questo lavoro, risultando un concetto derivabile dalle teorie fonologiche di una determinata lingua. La trasposizione di tale unità segmentale a livello fonetico comporta numerosi problemi e risulta difficilmente definibile con una chiarezza confrontabile alla definizione che riceve negli altri livelli.

Su queste difficoltà di identificazione dei confini sillabici nel dominio fonetico risulta esserci un sufficiente accordo tra gli studiosi (Kopecek, 1999; Noetzel, 1991; Pfitzinger *et alii*, 1996; Wu *et alii*, 1997), che, come nel caso di (Goslin *et alii*, 1999), sottolineano come tali problemi rendano il processo di segmentazione in sillabe degli enunciati una operazione estremamente complessa anche per annotatori umani. Quest'ultimo studio, dopo aver verificato sperimentalmente queste difficoltà sottoponendo annotatori umani a verifiche incrociate, propone l'utilizzazione di unità sillabiche differenti dalla sillaba.

D'altra parte numerosi studi sull'influenza della prominenza sulle varie componenti sillabiche (Greenberg, et al 2003; Jenkin, Scordilis, 1996; Silipo, Greenberg 1999; van Bergem, 1993; van Kuijk, Boves, 1999) hanno mostrato come le principali modificazioni siano a carico del nucleo. Vari risultati sperimentali hanno correlato in maniera affidabile la presenza di prominenza nelle sillabe con un allungamento della durata della vocale che ne costituisce il nucleo, mostrando inoltre come solo il nucleo sillabico appaia subire queste modificazioni in presenza di fenomeni di prominenza.

Per un'ulteriore verifica di queste asserzioni abbiamo misurato la lunghezza di sillabe prominenti e non di un certo numero di enunciati tratti dal TIMIT corpus, utilizzando le segmentazioni manuali, e le abbiamo confrontate con le lunghezze dei nuclei sillabici nei medesimi enunciati rispetto al loro grado di prominenza. Il grado di separazione dei due insiemi di misure ottenuti conferma sostanzialmente i risultati descritti dagli altri studiosi, mostrando come misure di durata effettuate sui nuclei sillabici siano altrettanto valide, rispetto alla prominenza delle sillabe, di quelle effettuate considerando l'intera estensione temporale delle sillabe.

3.2 Identificazione dei nuclei sillabici

Affrontando il problema dell'identificazione delle sillabe nel linguaggio parlato, gran parte degli studiosi si è concentrata sull'introduzione di opportune misure, o tratti fonetico/acustici, che potessero essere utili nell'individuazione dei nuclei sillabici, o, in modo equivalente, che potessero manifestare la presenza di una sillaba. In questo senso molti degli sforzi compiuti sono stati diretti nell'identificazione di quelli che sono stati

definiti da Stevens (1992) "*landmark*", o marcatori di specifici eventi acustici nell'enunciato. Sulla linea tracciata da Stevens si inseriscono i lavori di Howitt (2000), Liu (1996) e Bitar, Espy-Wilson (1996) tesi a sviluppare metodi automatici per l'identificazione di marcatori di specifici eventi.

Questo modo di procedere "per marcatori" è stato spesso utilizzato negli studi sull'identificazione dei nuclei sillabici, in particolare nell'ottica di individuare opportune posizioni nell'enunciato ove sia possibile rilevare le caratteristiche fonetico/acustiche tipiche dei nuclei. L'identificazione della sillaba, in letteratura, si è principalmente concentrata sull'individuazione del "centro di gravità" sillabico piuttosto che nell'identificazione dei confini della sillaba stessa, che abbiamo visto essere estremamente problematici, se non addirittura indefinibili dal punto di vista acustico. Questo modo di procedere è stato trasposto inalterato nei lavori volti all'identificazione del nucleo sillabico: anche in questo caso gli studi più rilevanti hanno fornito metodi che si concentrano sull'identificazione di una posizione nell'enunciato ove i tratti acustici suggeriscono un'elevata possibilità di essere all'interno di un nucleo sillabico, ma ben pochi di essi hanno affrontato il problema di individuare i confini di tali nuclei.

Nello studio in esame è tuttavia necessaria una identificazione delle dimensioni dei vari nuclei sillabici, al fine di una corretta misura della durata degli stessi, e quindi dei loro confini nei confronti delle unità che li circondano. Il lavoro verrà quindi diviso in due parti: la prima volta all'identificazione delle posizioni dei nuclei sillabici attraverso una suddivisione dell'enunciato in segmenti consecutivi ognuno contenente un nucleo, la seconda volta all'esplorazione dell'enunciato all'interno di un singolo segmento per l'individuazione di possibili confini che delimitino il nucleo rispetto alle unità circostanti.

Isolamento dei nuclei

Tra i metodi che non prevedono fasi di apprendimento, spicca il citatissimo studio effettuato da Mermelstein (1975). L'algoritmo originale proposto da Mermelstein utilizza come parametro fondamentale per l'identificazione dei nuclei sillabici l'energia del segnale in una specifica banda di frequenze (500-4000 Hz) successivamente mediata nel tempo tra 5 frame consecutivi. Per quanto riguarda l'identificazione dei picchi nel profilo energetico, che Mermelstein propone come indicatori di un nucleo sillabico, utilizza un algoritmo ricorsivo basato su *convex hull*. Dato un segmento di enunciato, e la relativa funzione energetica, si definisce la *convex hull* come la funzione non decrescente che ha la minima differenza con la funzione energetica prima del massimo globale nel segmento considerato, e la funzione non crescente che ha la minima differenza col profilo energetico dopo il massimo globale.

Vi sono altri metodi in letteratura utilizzati per l'identificazione di regioni contenenti nuclei sillabici. Howitt (2000) sottolinea come un massimo nell'ampiezza associata alla prima formante sia un indicatore affidabile della presenza di una vocale nell'enunciato (*vowel landmark*), e elabora un approccio alternativo alla determinazione di questi punti di massimo, riconoscendo che, se si concentra l'analisi sull'energia calcolata nella banda spettrale da 300 a 900 Hz, si ottengono risultati molto simili, evitando completamente l'uso di algoritmi per l'estrazione automatica delle formanti, riconosciuti da molti studiosi come piuttosto problematici (si veda ad esempio Vallabha, Tuller, 2002). Le informazioni ottenute analizzando il profilo energetico nella banda indicata vengono fornite a una rete neurale in grado, dopo un'adeguata fase di apprendimento su dati etichettati manualmente, di individuare punti nell'enunciato (*landmark*) ove è più probabile la presenza di una vocale e punti ove quasi certamente non si è in presenza di una vocale.

Un approccio radicalmente diverso al problema coinvolge gli studi volti alla risoluzione del generico problema della segmentazione temporale degli enunciati; questi studi riguardano essenzialmente l'identificazione di regioni spettralmente "quasi stazionarie", all'interno delle quali le caratteristiche spettrali dell'enunciato restano approssimativamente costanti. In teoria regioni di questo tipo si riferirebbero biunivocamente alle unità segmentali che compongono l'enunciato, ma in pratica i fenomeni di coarticolazione e di riduzione o soppressione dei suoni rendono le segmentazioni basate su parametri acustici non coincidenti con quelle basate su criteri fonologici.

Esiste una vastissima bibliografia sull'argomento e numerosi di questi metodi sono stati effettivamente implementati nel corso di questo studio per valutarne le prestazioni su casi reali. Uno dei metodi che si è dimostrato più efficace è stato proposto da Andre-Obrecht (1988): l'enunciato viene considerato come una sequenza di zone quasi-stazionarie e ogni zona viene modellizzata da un modello autoregressivo gaussiano che corrisponde a un modello LPC (*Linear Predictive Coding*). Due modelli autoregressivi *M0* e *M1* vengono definiti e, sulla base dell'introduzione di un'oportuna metrica basata su entropia mutua condizionata (W_n), vengono valutate due ipotesi distinte:

- *H0* - non esiste alcuna discontinuità all'interno del modello *M0*, ovvero W_n non varia;
- *H1* - esiste un punto di discontinuità r ($r < n$) nel quale W_n ha una deriva negativa superiore a una determinata soglia.

Questo metodo si è dimostrato estremamente efficace nell'individuazione dei punti di transizione tra una unità segmentale e la successiva (92.84% di corrette classificazioni rispetto ai dati ricavati dal TIMIT corpus con una tolleranza di 20 msec), generando però una quantità decisamente eccessiva di falsi allarmi ovvero inserimenti di transizioni inesistenti (68.8% in più rispetto a quelle reali).

Come hanno sottolineato alcuni studiosi nei loro lavori (Glass, Zue, 1988), appare estremamente difficile produrre algoritmi in grado di identificare stati quasi-stazionari negli enunciati confrontabili con quelli prodotti dalle segmentazioni manuali. Sembra quindi ragionevole, alla luce anche delle alte percentuali di falsi allarmi introdotti dai vari metodi, considerare segmentazioni a livello più alto. Glass e Zue hanno mostrato come la corretta segmentazione dell'enunciato sia ottenibile componendo opportunamente alcuni dei segmenti ottenuti in prima istanza. Glass e Zue presentano un metodo che di fatto ottiene buoni risultati, ma utilizza modelli di Markov, e quindi algoritmi che necessitano di fasi di apprendimento.

Un'analisi approfondita dei risultati ottenuti utilizzando il metodo proposto da Andre-Obrecht mostra che il 79% dei nuclei correttamente identificati da una sequenza di intervalli stazionari, a causa dell'elevata stabilità spettrale delle vocali, viene suddiviso in massimo due segmenti contigui (osservazione confermata a livello qualitativo anche dagli esperimenti dell'autrice stessa), che, considerati globalmente, hanno un'alta corrispondenza coi dati delle segmentazioni forniti dal TIMIT corpus.

Il metodo che proponiamo per l'identificazione dei nuclei risulta essere un ibrido dei tre presentati precedentemente. L'idea è quella di restringere la scelta sulla determinazione dei confini dei nuclei alle transizioni proposte dall'algoritmo di Andre-Obrecht, che abbiamo visto sperimentalmente essere in elevato accordo con le segmentazioni di riferimento, considerando però che l'elevato tasso di falsi allarmi ci costringerà a determinare raggruppamenti tra i segmenti individuati da tali transizioni.

Questa prima fase si avvale dell'algoritmo proposto da Mermelstein, con alcune modifiche sostanziali che ne migliorano le prestazioni: (a) in primo luogo la banda di frequenza utilizzata è quella proposta da Howitt nel suo studio (300-900 Hz) che si è

dimostrata più efficace, sia nello studio di Howitt che negli esperimenti condotti all'interno di questa ricerca; (b) in secondo luogo le informazioni ottenute dalla segmentazione effettuata col metodo di Andre-Obrecht, restringendo le possibilità di individuazione dei confini della sillaba, evitano tutte quelle problematiche legate alla variabilità del profilo energetico che diversamente dovrebbero essere catturate con soglie, casi speciali, ecc., procedimenti sempre molto delicati e spesso fallimentari; (c) infine il profilo energetico considerato viene annullato ove un analizzatore di zone sonore basato sull'autocorrelazione non segnala la presenza di questo tipo di suoni. La figura 1 mostra graficamente come l'algoritmo di Mermelstein e le informazioni dovute alla segmentazione, ottenuta utilizzando l'algoritmo di Andre-Obrecht, vengono utilizzate per individuare i confini sillabici entro i quali risiedono i nuclei per un enunciato tratto dal TIMIT corpus. La soglia utilizzata per decidere se suddividere un intervallo o meno viene determinata dinamicamente in funzione dell'altezza del massimo globale all'interno dell'intervallo stesso.

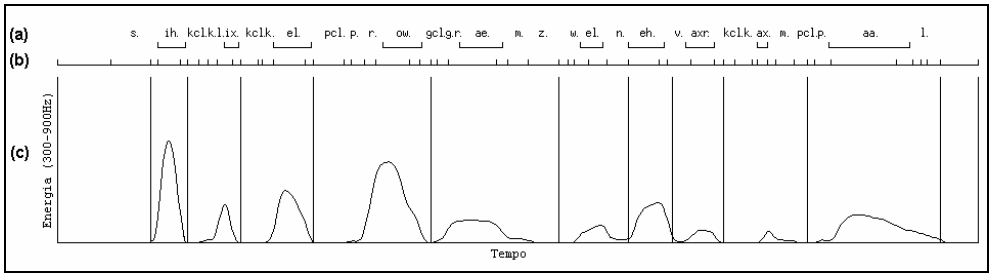

Figura 1. Applicazione del metodo proposto per l'identificazione dei confini entro i quali risiedono i nuclei sillabici alla frase "*Cyclical programs will never compile*". (a) Posizione dei nuclei sillabici tratta dalla segmentazione di riferimento; (b) Intervalli spettralmente quasi-stazionari; (c) profilo energetico e suddivisione dell'enunciato prodotta dal metodo proposto.

Identificazione dei confini dei nuclei

La procedura descritta nella sezione precedente ha prodotto una partizione dell'enunciato, rispetto all'asse temporale, composta da segmenti contigui ciascuno dei quali contiene un nucleo sillabico e quindi un massimo nel profilo energetico dell'enunciato calcolato nella banda di frequenza 300-900 Hz.

Il metodo proposto si concentra ora su uno dei segmenti prodotti dalla parte precedente e deve produrre una partizione di tale segmento il più possibile corrispondente alla definizione del nucleo sillabico estratta dalla segmentazione fornita col *corpus* di riferimento. La suddivisione in sezioni spettralmente quasi-stazionarie effettuata precedentemente può essere estremamente utile anche in questo caso, per tutte le proprietà già considerate precedentemente. Anche le informazioni fornite dal profilo energetico nella banda 300-900 Hz possono essere utilizzate per la determinazione dei confini. Abbiamo considerato che la parte più energetica della sillaba corrisponda al suo nucleo, quindi la forma del massimo ad essa associato ne suggerirà l'estensione temporale. Si consideri la figura 2: il grafico (a) mostra un ipotetico profilo energetico all'interno di uno dei segmenti prodotti nella fase di identificazione dei nuclei e (b) la suddivisione in zone spettralmente quasi-stazionarie. Il grafico (c) mostra un profilo energetico calcolato considerando come intervallo di integrazione gli intervalli quasi-stazionari. Come si può vedere, due dei sei intervalli contengono la maggior parte dell'energia di questo nucleo e, molto

probabilmente, i confini dell'unione di questi due intervalli identificano correttamente i confini del nucleo in esame.

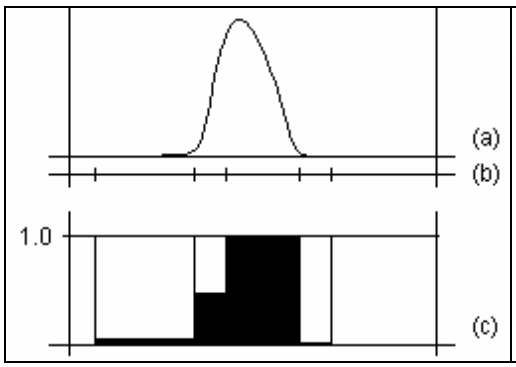

Figura 2: Identificazione dei confini dei nuclei sillabici. (a) Profilo energetico; (b) suddivisione in intervalli spettralmente quasi-stazionari; (c) profilo energetico considerando la suddivisione in intervalli.

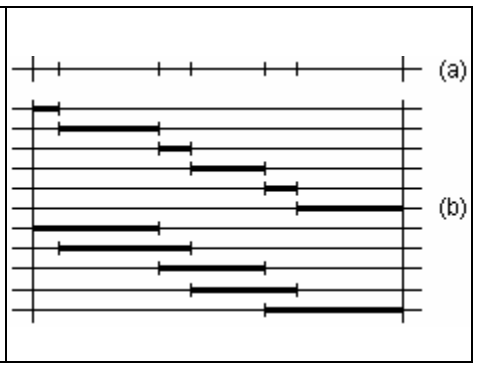

Figura 3: Creazione delle partizioni. (a) intervalli quasi-stazionari; (b) possibili partizioni: i segmenti evidenziati individuano la partizione P^+, gli altri la partizione P^-.

L'algoritmo sviluppato appositamente, all'interno di questo studio, per l'identificazione dei confini dei nuclei procede generando tutte le possibili partizioni binarie degli intervalli quasi-stazionari tali che la partizione che viene logicamente associata al nucleo sia composta al massimo da due intervalli. Il numero di tali partizioni è estremamente contenuto, sia perché il numero di intervalli coinvolti (N_i) è tipicamente minore di 10, sia a causa delle proprietà che abbiamo imposto alla partizione, che riducono il numero totale delle possibilità a $2N_i$-*1*. La figura 3 mostra le possibili partizioni per N_i=*6*. Ogni partizione *P* suddivide l'insieme degli intervalli in due parti, la prima (P^+) che contiene gli intervalli che identificano il nucleo, mentre la seconda (P^-) gli intervalli esterni al nucleo. Per ogni intervallo è possibile calcolare l'area del rettangolo individuato dalla sua estensione temporale e dalla energia contenuta in tale intervallo, che chiameremo $Area^+$ (indicata in nero nella figura 2c) e l'area differenza tra l' $Area^+$ e il valore *1.0*, tetto della normalizzazione sul massimo delle energie degli intervalli, che indicheremo con $Area^-$ (marcata in bianco nella figura 2c).

Se consideriamo una funzione di *Score* per una partizione *P* definita come

$$Score(P) = \sum_{\text{int}_j \in P^+} Area^+_{\text{int}_j} + \sum_{\text{int}_k \in P^-} Area^-_{\text{int}_k}$$

è possibile calcolare la migliore partizione tra quelle generate come

$$P_{best} = \arg\max_P (Score(P)) = \arg\max_P \left(\sum_{\text{int}_j \in P^+} Area^+_{\text{int}_j} + \sum_{\text{int}_k \in P^-} Area^-_{\text{int}_k} \right).$$

Durata dei nuclei sillabici

Una volta identificati i confini dei nuclei sillabici, il calcolo della loro durata temporale, obiettivo di tutto il processo di segmentazione, risulta essere immediato.

Resta tuttavia da chiarire l'influenza della velocità di elocuzione (*rate-of-speech*). Sembra opportuno considerare procedimenti di normalizzazione di questi valori, al fine di

poter effettuare un confronto tra enunciati diversi realizzati da differenti locutori. Tra i numerosi studi relativi alla normalizzazione dei parametri di durata temporale alcuni metodi si basano su misure statistiche all'interno del singolo enunciato (Neumeyer *et alii*, 1996; Gadde, 2000); tra i metodi proposti da questi ultimi studi, quello di normalizzare la durata di ogni nucleo contenuto nell'enunciato rispetto alla durata media dei nuclei nell'enunciato stesso è sembrato un metodo utile al fine di ottenere una normalizzazione adeguata delle durate senza peraltro richiedere l'utilizzo di risorse linguistiche annotate o complessi algoritmi.

4. PROFILI INTONATIVI E *PITCH*

La determinazione della frequenza fondamentale di un enunciato risulta essere un problema estremamente complesso. Nel corso degli ultimi 30 anni si è assistito a un proliferare di studi nel settore che hanno determinato un panorama estremamente vario e articolato di metodi automatici per la determinazione di F0. Potremmo fornire numerosi riferimenti bibliografici a lavori in questo campo, ma preferiamo rimandare agli studi condotti da Bagshaw (1994) e da de Cheveigné, Kawahara (2002) per un confronto degli spunti metodologici e degli principali algoritmi.

Nell'ambito del nostro studio abbiamo utilizzato l'algoritmo di *pitch tracking* che viene considerato tra gli studiosi del settore come il più performante (RAPT - *Robust Algorithm for Pitch Tracking* - Talkin, 1995), tanto da essere comunemente ritenuto il metodo di riferimento per la valutazione delle prestazioni di nuovi algoritmi per l'estrazione del *pitch*.

Tuttavia, nell'ambito del nostro studio, è parso opportuno introdurre un'ulteriore fase di post-elaborazione al fine di rimuovere principalmente errori di *halving* e *doubling* che risulterebbero molto penalizzanti nella determinazione dei *pitch accent*. Per la rimozione di questi errori è stato utilizzato il metodo presentato da Bagshaw (1994).

5. MISURE DI INTENSITÀ

All'interno di questo lavoro l'energia di un frammento di enunciato viene calcolata utilizzando l'energia *root mean square* (RMS) all'interno dei segmenti corrispondenti ai nuclei sillabici individuati col processo di segmentazione.

Come abbiamo più volte sottolineato nelle sezioni precedenti, i correlati acustici che possono fornire informazioni sul livello di prominenza sembrano essere, durata, frequenza fondamentale e intensità. All'interno di questi parametri la durata e le configurazioni assunte dal profilo del *pitch* sono visti come i più rilevanti e robusti. Allo stesso tempo però vi sono evidenze sperimentali che attribuiscono all'intensità un ruolo non secondario nell'identificazione delle sillabe prominenti, in particolare per quanto riguarda le sue elaborazioni legate principalmente a misure effettuate in specifiche bande spettrali, che sono state definite in letteratura come *spectral balance*, *spectral emphasis* o *spectral tilt* (Fant *et alii*, 2000; Heldner, 2001; 2003; Sluijter, van Heuven, 1996a; 1996b; 1997). In particolare, i lavori di Sluijter e van Heuven analizzano i contributi energetici di 4 bande spettrali contigue (0-500-1000-2000-4000 Hz), mostrando sperimentalmente che il contributo energetico della prima banda spettrale (0-500 Hz) non risulta essere in alcun modo correlato con la percezione di prominenza, mentre, al contrario, il contributo energetico delle altre 3 bande risulta correlato positivamente con la prominenza all'interno della sillaba.

All'interno di questo studio sono stati realizzati alcuni esperimenti a verifica di tali risultati, utilizzando bande di frequenza leggermente diverse, estremamente simili a quelle

utilizzate in (Strom, 1995) all'interno del progetto VERBMOBIL, in particolare 0-300 Hz, 300-2200 Hz e 2200-4000 Hz. I contributi energetici di queste tre bande spettrali sono stati messi in relazione alla prominenza delle sillabe corrispondenti. Dalle prove emerge piuttosto chiaramente che la banda centrale di frequenza (300-2200 Hz), quella che contiene le principali formanti associate alle vocali presenti nel nucleo sillabico, mostra come le sillabe prominenti tendano ad avere contributi energetici superiori a quelle non prominenti. Ai fini di questo studio utilizzeremo quindi l'energia RMS misurata nella banda 300-2200 Hz come misura di enfasi spettrale.

Come nel caso precedente, relativo alla durata, sembra opportuno, per evitare la variabilità di condizioni sia ambientali che comunicative, normalizzare le misure energetiche rispetto a parametri interni ad ogni enunciato: normalizzare queste misure utilizzando la media delle energie dei nuclei sillabici dell'enunciato risulta essere un adeguata procedura per rimuovere quelle informazioni non rilevanti ai fini dell'individuazione automatica della prominenza.

6. IDENTIFICAZIONE DEI FENOMENI

In base alle argomentazioni delineate precedentemente e all'esame degli studi condotti nel settore, abbiamo proposto un modello del fenomeno percettivo della prominenza basato essenzialmente sull'occorrenza di due specifici fenomeni linguistici, entrambi sufficienti ad individuare una sillaba prominente, ma nessuno strettamente necessario: lo *stress* della sillaba, fenomeno essenzialmente a carico del nucleo sillabico, o la presenza di un *pitch accent* all'interno della sillaba analizzata. A loro volta i fenomeni linguistici indicati sono stati correlati, in letteratura, con fenomeni acustici direttamente ricavabili dai parametri fisici dell'enunciato. Si delinea quindi una gerarchia nella quale possiamo inserire i fenomeni in gioco, siano essi percettivi, linguistici o acustici (si veda la tabella 1).

Fen. percettivi	Prominenza			
Fen. linguistici	*Stress*		*Pitch accent*	
Fen. acustici	durata	enfasi spettrale	mov. di F0	intens. globale

Tabella 1: Gerarchia dei fenomeni coinvolti in questo studio. Ogni livello si basa sui livelli sottostanti per l'identificazione dei corrispondenti fenomeni.

L'identificazione di uno specifico fenomeno si fonda quindi sulla individuazione e sulla combinazione dei livelli sottostanti attraverso opportune manipolazioni.
Ci concentreremo quindi sull'analisi delle relazioni che intercorrono tra i differenti parametri nella determinazione della prominenza, in particolare per quanto riguarda i fenomeni linguistici.

6.1 Stress

I correlati acustici che identificano lo *stress* frasale in modo più affidabile sembrano essere la durata dei nuclei sillabici e picchi energetici in specifiche bande spettrali; in particolare la banda da 300 a 2200 Hz sembra essere quella che meglio si correla al fenomeno della prominenza indotto appunto dallo *stress*. I lavori di numerosi studiosi, tra cui (Sluijter, van Heuven, 1996a; 1996b; 1997), sono concordi nell'indicare che un aumento dei valori di queste due grandezze è da associare direttamente alla percezione dello *stress* frasale e quindi alla prominenza.

L'andamento generale delle distribuzioni dei due insiemi di dati (sillabe prominenti *vs* sillabe non prominenti) delineatosi negli esperimenti condotti all'interno di questo studio

sembra confermare le considerazioni effettuate. Valori elevati di energia RMS (300-2200 Hz) e elevate estensioni temporali dei nuclei sillabici risultano essere indicatori sufficientemente affidabili della presenza di *stress* nella sillaba in esame.

Possiamo quindi desumere, almeno come indicazione qualitativa, che, dovendo ridurre la determinazione dello *stress* ad un unico parametro, il prodotto della durata per l'energia (300-2200 Hz) può essere considerato una misura sufficientemente affidabile, che attribuisce valori "grandi" alle sillabe prominenti e valori "piccoli" alle sillabe non prominenti.

6.2 Pitch Accent

Nell'ambito di questo lavoro è stato estensivamente sperimentato il modello presentato da Taylor (Tilt) (1992; 2000). Nel modello Tilt ogni evento viene identificato con un movimento crescente, decrescente o crescente-decrescente e descritto in modo completo misurando le escursioni in frequenza e le durate temporali di ogni segmento e definendo, tra altri, i parametri seguenti:

$$A_{event} = |A_{rise}| + |A_{fall}| \qquad D_{event} = D_{rise} + D_{fall}$$

dove A_{rise} e A_{fall} sono rispettivamente le escursioni in frequenza del segmento crescente e decrescente relativo a un evento mentre D_{rise} e D_{fall} sono le rispettive durate temporali.

Il profilo del *pitch*, è stato suddiviso in *frame* lunghi 25 msec e i dati all'interno del *frame* sono stati interpolati utilizzando un metodo *Least Median Squares* per ottenerne una regressione lineare affidabile e non affetta dal problema dei valori erratici. Ogni frame è stato poi classificato in base al gradiente della retta interpolante come crescente (*Rise*-"R"), decrescente (*Fall*-"F") o costante (*Connection*-"C") utilizzando come soglia tra le varie classi il valore ±100 Hz/sec come suggerito in (Taylor, 1993). Tutti i frame consecutivi classificati allo stesso modo sono quindi stati compattati in un unico intervallo temporale, contenente un movimento uniforme del profilo intonativo in un determinato intervallo di tempo.

Il modello *RFC* del profilo intonativo ottenuto risulta essere un punto di partenza ideale per l'identificazione degli eventi che potenzialmente possono definire un *pitch accent*: secondo (Taylor, 2000) un evento intonativo candidato per identificare un *pitch accent* è composto da una sezione crescente seguita da una sezione decrescente del profilo di F0, di conseguenza l'individuazione delle sequenze di intervalli marcate con "RF" nel profilo intonativo consente di isolare gli eventi che rappresentano i *pitch accent* all'interno dell'enunciato.

Le teorie autosegmentali metriche tendono a definire numerosi profili di F0 in grado di produrre un *pitch accent*, ma, come osservato in (Taylor, 2000), le differenze tra i vari profili sembrano essere dovute a differenze di allineamento del *pitch accent* rispetto al nucleo sillabico piuttosto che a reali differenze nei profili. Altri studiosi (Ladd, Shepman, 2003; Rietveld, Kerkhoff, 2002) hanno notato come vi siano delle sovrapposizioni tra i differenti profili introdotti originalmente da Pierrehumber e ripresi poi nel modello Tobi, riducendo, di fatto, il numero di profili che sono in grado di definire un *pitch accent*. D'altra parte studi statistici dello stesso Taylor (2000) hanno mostrato come nel 79% dei casi un *pitch accent* venga realizzato con un profilo H* e nel 15% dei casi con profilo L+H*. Quindi appare plausibile considerare i picchi nel profilo di F0 come gli eventi intonativi maggiormente rilevanti ai fini dell'identificazione dei *pitch accent*, spostando eventualmente il problema al loro allineamento coi nuclei sillabici. Nell'ambito di questo studio, considerando anche l'importanza che alcuni studiosi hanno attribuito ai segmenti

crescenti, includeremo nell'analisi anche gli eventi degeneri formati unicamente da una sezione crescente.

Numerosi studi sulle correlazioni tra la posizione dei *pitch accent* e le unità segmentali (Taylor, 2000; van Santen, Möbius, 1997; Wichmann *et alii*, 2000; Xu, 2002) hanno mostrato come la posizione dei *pitch accent* sia influenzata fortemente da fattori quali l'effettiva composizione delle unità segmentali e vincoli articolatori che pongono precisi limiti temporali e di estensione ai movimenti della frequenza fondamentale. Il risultato di tali fenomeni genera piccoli spostamenti temporali dei *pitch accent* rispetto alle unità segmentali che, pur non annullando completamente le connessioni tra sillabe e eventi intonativi, rendono ardua la definizione delle corrispondenze necessarie per una corretta analisi dei fenomeni in esame.

Nonostante la complessità dell'analisi vi sono numerosi studi che indicano utili regolarità nel comportamento di questi fenomeni. In (Taylor, 2000) viene sottolineato come la posizione dei picchi nel profilo intonativo sia stata correlata con le differenti definizioni dei *pitch accent* nei modelli autosegmentali metrici. I lavori di altri studiosi confermano sostanzialmente le considerazioni di Taylor sulla lingua inglese (van Santen, 2001; Wichmann *et alii*, 2000) e anche in altre lingue come ad esempio l'olandese (Rietveld, Kerkhoff, 2002). Sulla base di questi risultati sull'allineamento tra eventi e unità segmentali sembra emergere un ruolo rilevante dei profili crescenti all'interno dei *pitch accent* come punto di riferimento temporale maggiormente affidabile per definire queste corrispondenze. Esistono infatti studi specifici che mostrano come i segmenti crescenti contenuti nei *pitch accent* siano da associarsi temporalmente con la sillaba alla quale si riferiscono (Ladd *et alii*, 1999; Ladd, Shepman, 2003; van Santen, Möbius, 1997; van Santen, 2002). Sembra quindi ragionevole utilizzare i segmenti crescenti degli eventi intonativi come punto di riferimento privilegiato per determinare la corrispondenza, in termini di sovrapposizione, coi nuclei sillabici che compongono le nostre unità segmentali.

I lavori di Sluijter e van Heuven, ai quali ci siamo più volte riferiti, indicano l'ampiezza dei movimenti nel profilo di F0 e l'energia globale, calcolata su tutta la banda spettrale o almeno su una banda molto ampia (nel nostro caso 50-5000 Hz), come i due parametri acustici in grado di individuare sillabe che possiedono un alto livello di prominenza, attraverso il meccanismo dei *pitch accent* (anche se non utilizzano la nostra stessa terminologia). Seguendo il modello Tilt, nell'implementazione appena presentata, e in particolare considerando il parametro A_{event} che misura proprio l'ampiezza dell'escursione in frequenza dell'evento intonativo come somma delle ampiezze della sezione crescente e decrescente in valore assoluto, è possibile valutare l'impatto dei due parametri acustici sull'identificazione della prominenza in un modo del tutto simile allo studio condotto per lo *stress* nella sezione precedente. In questo caso però, data la natura e la forma degli eventi intonativi e soprattutto la complessità dell'identificazione affidabile del profilo con metodi automatici, sembra opportuno considerare il parametro A_{event} in relazione alla durata (D_{event}) dell'evento, in particolare moltiplicando i due valori e calcolando così un parametro al quale potremmo approssimativamente dare l'interpretazione di "area" dell'evento intonativo. Misurando questa sorta di area anziché l'ampiezza dell'evento si riduce di molto l'impatto di errori nel calcolo del profilo di F0, errori che si manifestano, solitamente, come bruschi e isolati cambiamenti di altezza del *pitch*.

E' possibile studiare l'andamento dei parametri acustici che dovrebbero supportare la presenza di un *pitch accent*. In modo simile alle considerazioni fatte relativamente allo *stress*, dagli esperimenti si evince che una maggior energia globale nel nucleo e una maggiore ampiezza del parametro relativo ai movimenti di F0 correlano positivamente con

la prominenza confermando gli studi presi a riferimento per questo lavoro. Possiamo quindi concludere che valori elevati di energia in concomitanza con movimenti rilevanti nel profilo di F0 sono buoni indicatori della presenza di un *pitch accent* e quindi della prominenza della sillaba.

7. IDENTIFICAZIONE AUTOMATICA DELLA PROMINENZA

Nella sezione introduttiva abbiamo posto e motivato gli obiettivi di questo lavoro e il contesto e i vincoli nei quali intendevamo inserirlo; uno di questi vincoli riguarda l'eliminazione della possibilità di utilizzare metodi "supervisionati", ossia che richiedono lunghe fasi di apprendimento su dati autentici etichettati manualmente. Tuttavia, considerando il panorama scientifico ricco di metodologie che sfruttano fasi di apprendimento per ottenere prestazioni migliori, ci è sembrato opportuno sperimentare alcuni di tali modelli (in particolare reti neurali MLP e *Support Vector Machine*) anche per valutare nel modo più corretto, confrontandolo con i metodi supervisionati più performanti, le prestazioni del metodo non-supervisionato che proporremo alla fine del capitolo. L'obiettivo di sviluppare metodologie non dipendenti da fasi di apprendimento posto nel capitolo introduttivo resta quindi valido per quanto riguarda le tecniche e gli algoritmi proposti dall'autore.

7.1 La nostra proposta: un metodo non supervisionato

Gran parte della discussione sviluppata in letteratura a proposito della definizione del fenomeno della prominenza è volta a sottolineare come esso abbia una grossa componente di confronto sull'asse sintagmatico, che porta ad adottare la definizione

> *Prominence is the property by which linguistic units are perceived as standing out from their environment.*

introdotta da Terken (1991). Il concetto stesso di prominenza appare quindi intimamente connesso con una analisi del contesto della sillaba in esame: definire come prominente o meno una sillaba esaminando unicamente i valori dei parametri all'interno di essa appare quindi estremamente riduttivo. Al contrario, una corretta determinazione della prominenza di una determinata unità segmentale dovrebbe essere basata sull'esame dei parametri fonetico/acustici della sillaba stessa e su un loro confronto coi parametri derivati delle sillabe che la circondano.

In questa sezione intendiamo presentare un metodo per l'identificazione automatica della prominenza di tipo non supervisionato che tenga conto opportunamente delle informazioni derivate dal contesto della sillaba in esame. Si definisce matematicamente un'opportuna "funzione di prominenza", legata ai parametri acustici dei nuclei sillabici identificati nella fase di segmentazione, in modo che produca valori su un asse continuo. In accordo con (Taylor, 2000) e in base alle considerazioni effettuate precedentemente, riteniamo che un'applicazione per il riconoscimento automatico della prominenza debba fornire valori e valutazioni su una scala continua piuttosto che una scelta forzatamente dicotomica; definiremo quindi il comportamento del metodo che proporremo su un codominio continuo di valori e forzeremo la scelta tra i due estremi del *continuum* solo per esigenze di valutazione delle prestazioni e di confronto coi metodi precedenti.

Le considerazioni qualitative introdotte nelle sezioni precedenti si possono trasformare in legami quantitativi definendo una funzione di prominenza del tipo:

$$\mathrm{Prom}^{i} = \max\left\{ en^{i}_{300-2200} \cdot dur^{i} \;,\; en^{i}_{glob} \cdot (A^{i}_{event} \cdot D^{i}_{event}) \right\} \qquad (1)$$

dove $en_{300\text{-}2200}$ è l'energia RMS nella banda 300-2200 Hz, *dur* è la durata temporale del nucleo, en_{glob} è l'energia globale nel nucleo e A_{event} , D_{event} sono i parametri del modello

TILT come sono stati definiti in precedenza (si noti che nel caso non sia presente alcun evento intonativo all'interno della sillaba questi due parametri valgono 0), riferiti al generico nucleo *i*. La struttura della funzione *Prom*, sebbene sembri scelta arbitrariamente, riflette in realtà le relazioni trai parametri che abbiamo stabilito all'interno di questo lavoro e, in particolare, la funzione *max* esprime la disgiunzione logica che definisce la prominenza in funzione dei parametri linguistici di *stress* e *pitch accent*, in un modo del tutto simile alla definizione adottata nell'ambito della logica *fuzzy* relativamente all'unione di due insiemi. Onde evitare interferenze dovute ai differenti intervalli di valori dei due argomenti della funzione *max*, questi, prima del calcolo della funzione, sono stati entrambi normalizzati rispetto al corrispondente valore massimo all'interno dell'enunciato.

Sulla base della definizione della funzione *Prom*, e considerando l'attribuzione della prominenza come un parametro legato al contesto, l'identificazione delle sillabe prominenti corrisponde alla ricerca dei massimi relativi della funzione *Prom*: in quest'ottica il valore della funzione di prominenza per ogni nucleo viene confrontato coi corrispondenti valori dei nuclei adiacenti e, se rappresenta un massimo, il nucleo sillabico corrispondente (e di conseguenza anche la sillaba ad esso associata) viene etichettato come prominente (D'Anna *et alii*, 2001).

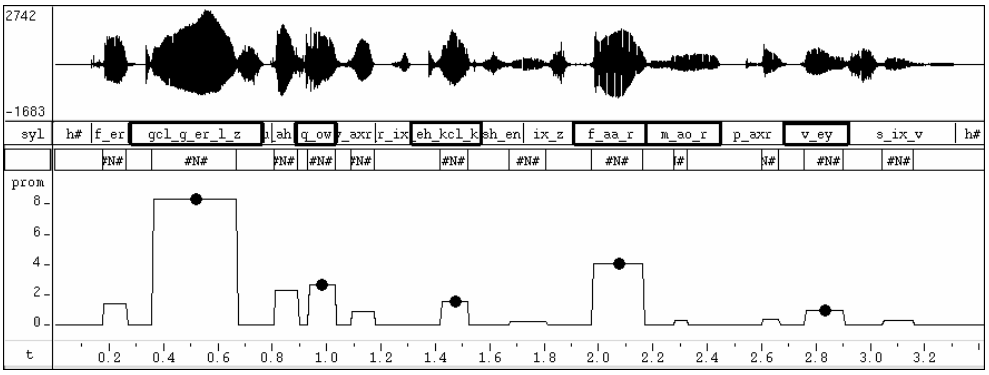

Figura 4: Profilo della funzione di prominenza per l'enunciato "*For girls the overprotection is far more pervasive*". Dall'alto: la forma d'onda originale, la segmentazione in sillabe (mostrata unicamente come riferimento), i nuclei sillabici identificati dal sistema automatico (segnalati dall'etichetta *#N#*), e infine il valore della funzione di prominenza per ogni nucleo identificato dalla procedura di segmentazione. I nuclei prominenti identificati dal sistema automatico sono marcati con un cerchio nero sul profilo della funzione, mentre le sillabe prominenti, classificate manualmente, sono indicate da un rettangolo nero.

Nella lingua inglese americana, come del resto nella lingua inglese britannica, può tuttavia accadere che due sillabe che risultano consecutive nell'enunciato siano percettivamente entrambe prominenti; è il caso di due parole successive monosillabiche entrambe prominenti. Le due sillabe presenteranno probabilmente diversi valori, o "livelli" della funzione di prominenza, ma in una prospettiva di classificazione dicotomica i livelli di prominenza vengono mascherati dall'attribuzione all'una o all'altra classe. L'algoritmo di ricerca dei massimi è stato modificato in modo da ignorare la sillaba immediatamente vicina nel caso che la differenza tra i due valori della funzione *Prom* siano inferiori al 15%; in tal caso la sillaba considerata nel metodo di ricerca del massimo locale non è quella immediatamente adiacente, ma la successiva. Inoltre nuclei sillabici che presentano un

valore della funzione di prominenza maggiore del 70% del valore del massimo globale nell'enunciato vengono etichettati anch'essi come prominenti.

La figura 4 mostra un grafico della funzione di prominenza per la frase "*For girls the overprotection is far more pervasive*" tratta dal *corpus* TIMIT.

Il metodo proposto, che d'ora in poi chiameremo *Blind Continuous Prominence Function* - BCPF, è stato sottoposto alla stessa valutazione di prestazioni dei metodi supervisionati considerati utilizzando però un maggior numero di dati grazie al fatto che, essendo un metodo non supervisionato, non necessita di alcuna fase di apprendimento.

7.2 Gli esperimenti effettuati

Tutti i metodi considerati sono stati valutati utilizzando dati estratti dal *corpus* TIMIT. In particolare sono stati formati due insiemi distinti di enunciati da utilizzarsi nelle fasi di apprendimento e di test di tutti i modelli supervisionati, insiemi che sono stati uniti per formare un terzo gruppo di test utilizzato per il metodo non supervisionato che presenteremo al termine del capitolo. La tabella 2 riassume le dimensioni e le composizioni dei vari insiemi utilizzati nelle sperimentazioni.

Insieme	**Num. Enunciati**	**Num. locutori**	**Num. sillabe**
TrainingSET	191	25	2853
TestSET	193	26	2855
GlobalSET	384	51	5708

Tabella 2: Insiemi di dati utilizzati nella sperimentazione dei vari metodi.

E' opportuno sottolineare che gli insiemi *TrainingSET* e *TestSET* sono disgiunti da ogni punto di vista, ossia contengono enunciati diversi prodotti da differenti locutori appartenenti a entrambi i sessi. Tutti questi enunciati sono stati sillabati e classificati manualmente rispetto alla prominenza percepita e verranno quindi utilizzati come base di confronto per la classificazione proposta dai metodi automatici.

Gli esperimenti basati su reti neurali MLP (utilizzando il pacchetto software *LNKnet*) e *Support Vector Machines* (utilizzando SVM^{light}) sono stati eseguiti fornendo in ingresso vettori composti dai dati di tre sillabe, la sillaba in esame e le due adiacenti, per un totale di 12 valori reali. Per quanto riguarda le reti MLP abbiamo definito una struttura dei livelli composta da 12 neuroni in ingresso, 24 neuroni per ognuno dei due livelli nascosti e 2 neuroni in uscita.

La Tabella 3 mostra i risultati ottenuti considerando tutti gli esperimenti effettuati.

Metodo	**Tipo**	**Accuracy**	**Recall**	**Precision**
Reti neurali MLP	Superv.	81.60%	70.72%	72.51%
Support vector machine	Superv.	81.60%	66.56%	74.61%
BCPF	Non Superv.	80.80%	66.30%	72.46%

Tabella 3: Tavola riassuntiva dei risultati ottenuti considerando gli esperimenti effettuati.

7.3. Discussione dei risultati ottenuti

Nel capitolo introduttivo abbiamo più volte sottolineato come l'etichettatura di estese risorse linguistiche relative al linguaggio parlato sia un compito estremamente complesso che richiede risorse umane considerevoli. Proprio da queste considerazioni è nata l'idea di questo studio, quindi sembra ragionevole, volendo costruire un sistema capace di annotare

risorse linguistiche automaticamente, confrontare le prestazioni nell'etichettatura automatica della prominenza frasale con l'equivalente processo effettuato da annotatori umani.

Alcuni studi hanno analizzato in considerevole dettaglio le abilità di annotazione della prominenza, e di altri fenomeni prosodici, relativamente ad esperti umani. In particolare, un dato di estremo interesse per questo studio riguarda l'accordo tra gruppi di annotatori umani ai quali viene affidata l'annotazione del medesimo insieme di enunciati (Buhmann *et alii*, 2002; Eriksson *et alii*, 2002; Jenkin and Scordilis 1996; Pickering *et alii*, 1996).

Vi è una generale convergenza in letteratura nell'identificare un livello di accordo, tra i diversi annotatori chiamati a giudicare il livello di prominenza dei medesimi enunciati, attorno all'80-90% dei casi esaminati. L'intervallo di variazione è piuttosto ampio, ma, come sottolineato più volte nel corso dello studio, la valutazione di questi processi e il confronto di tali valutazioni risultano essere estremamente complesse. Le variabili in gioco sono molteplici e spesso i differenti studi utilizzano criteri di valutazione estremamente diversi che rendono i loro risultati difficilmente confrontabili.

L'annotatore automatico basato sulla definizione di un'opportuna funzione di prominenza e sull'identificazione dei massimi locali significativi all'interno del suo profilo (BCPF) è in accordo con l'annotatore umano che ha etichettato i campioni utilizzati nel corso di questo studio nell'80.80% dei casi, assegnando il corretto valore di prominenza alle sillabe degli enunciati esaminati. Il metodo proposto non utilizza alcuna informazione aggiuntiva (trascrizioni fonetiche e/o ortografiche dell'enunciato, informazioni sintattiche, ecc.), ma unicamente parametri acustici estratti direttamente dal campione e non richiede alcuna fase di apprendimento basata su dati etichettati manualmente. I risultati ottenuti dall'annotatore automatico, riguardo all'accordo coi dati annotati manualmente, sono in linea con l'accordo tra annotatori umani, quindi il metodo proposto sembra essere una valida alternativa al processo di annotazione manuale della prominenza per la costruzione di risorse linguistiche per la ricerca e la didattica, o quantomeno un buon punto di partenza per annotare sistematicamente grandi moli di dati che saranno sottoposte, in una fase successiva, a verifica manuale da parte di esperti.

Il confronto del metodo non supervisionato BCPF con i metodi che richiedono fasi di apprendimento, reti MLP e SVM, non ha mostrato una sostanziale differenza tra le prestazioni ottenute, che nel caso dei metodi supervisionati, superano quelle del metodo proposto per meno di una unità percentuale, non risultando quindi globalmente significative. In questo senso, non sembra quindi particolarmente utile, ai fini delle prestazioni nella fase di classificazione, utilizzare complessi metodi che richiedono lunghe fasi di apprendimento, preferendo, viceversa, metodi diretti basati sulla definizione di funzioni opportune che potremmo classificare come "basati su regole".

Gli studi precedenti nel settore dell'identificazione automatica della prominenza hanno spesso seguito approcci radicalmente diversi.

Bagshaw (1994) ha presentato un sistema automatico per il riconoscimento della prominenza per la didattica delle lingue, in particolare per la lingua inglese. Nel suo lavoro, come negli altri con obiettivi simili, la trascrizione dell'enunciato è data come una risorsa acquisita, grazie all'intrinseca operazione di verifica controllata che i sistemi *CALL* solitamente implementano. Utilizzando la trascrizione come supporto informativo aggiuntivo, il problema della segmentazione viene risolto con metodi di *automatic speech recognition* (ASR) basati su modelli HMM e le stesse operazioni di identificazione possono sfruttare l'identità dei foni a fini statistici. Il sistema presentato da Bagshaw ottiene un

accordo coi dati etichettati manualmente del 61.6% che è molto più basso del sistema presentato in questo studio.

Il sistema presentato da Jenkin e Scordilis (1996) misura automaticamente 6 parametri all'interno dell'enunciato, basati su caratteristiche di F0 e su misure di intensità e durata, impiegandoli come parametri di base su cui costruire numerosi esperimenti utilizzando differenti modelli teorici; tali metodi vengono applicati a enunciati estratti dal *corpus* TIMIT. Il primo, e più performante, sistema che presentano è basato su reti neurali e classifica correttamente la prominenza delle sillabe nell'84% dei casi. Il secondo è basato su modelli probabilistici e catene di Markov e ottiene performance comprese tra il 77 e l'80% di corrette classificazioni, mentre il terzo sistema, essenzialmente basato su regole, classifica correttamente la prominenza delle sillabe nel 70-75% dei casi esaminati. Tutti i tre sistemi automatici presentati da Jenkin e Scordilis sono basati su modelli che richiedono fasi di apprendimento su dati etichettati, e solo nel caso del primo sistema, raggiungono prestazioni leggermente superiori al metodo presentato nel nostro studio; inoltre nel loro lavoro il problema della segmentazione e dell'identificazione dei nuclei sillabici, e i relativi errori che verrebbero introdotti, viene completamente evitato sfruttando la trascrizione fonetica allineata all'enunciato che viene fornita nel *corpus* TIMIT.

Considerazioni simili possono essere effettuate sui risultati ottenuti da Wightman e Ostendorf (1994) col loro sistema di annotazione automatica applicato alla lingua inglese americana. Essi utilizzano metodi di ASR basati su modelli HMM per identificare le unità segmentali e isolare i nuclei sillabici che saranno la base per l'estrazione di un certo numero di parametri acustici, come nei casi precedenti. Il vettore di parametri acustici viene utilizzato come base per un modello che implementa alberi di decisione simili a HMM discrete per ottenere un'etichettatura della prominenza e di altri parametri prosodici. Questo sistema classifica correttamente la prominenza delle sillabe esaminate nell'85-86% dei casi. Come nel caso precedente, le prestazioni sono leggermente superiori al sistema presentato, ma sono state ottenute attraverso l'utilizzazione di metodi pesantemente supervisionati che richiedono lunghe fasi di apprendimento e risorse linguistiche rilevanti (si pensi alla complessità, alla pesantezza e alla delicatezza del modulo che implementa il riconoscimento automatico del linguaggio parlato).

Vereecken *et alii*, (1998) presentano un sistema simile al precedente per quanto riguarda la fase di identificazione delle unità segmentali nell'enunciato, basato quindi su metodi di ASR, al quale aggiungono informazioni linguistiche ricavate dalla trascrizione dell'enunciato come posizioni degli *stress* all'interno degli elementi lessicali, *part-of-speech tag*, frequenze delle parole, trascrizioni fonetiche, ecc., oltre che i parametri acustici più volte discussi. Tutte le informazioni ottenute vengono fornite a una rete MLP per il processo di classificazione. Il sistema è stato applicato a numerose lingue, ottenendo risultati di accordo con etichettatori umani molto vari. Il lavoro risulta estremamente interessante da un punto di vista interlinguistico, ma, concentrandosi sull'inglese americano, non presenta un sistema capace di fornire prestazioni superiori a quelli dei sistemi esaminati in questo studio.

Silipo e Greenberg (1999) presentano uno studio molto articolato basato anch'esso sull'inglese americano. Propongono 4 differenti sistemi di classificazione basati sull'estrazione automatica di 4 parametri acustici: durata, intensità, variazione e valor medio del *pitch*. Tutti questi parametri sono calcolati in base alla segmentazione fonetica che viene fornita come informazione aggiuntiva e data per acquisita; anche in questo caso il processo di segmentazione non è presente. Nessuno dei sistemi descritti, alla luce anche del fatto che il problema della segmentazione viene evitato utilizzando la trascrizione fonetica

degli enunciati, sembra avere un impatto rilevante sulle valutazioni di questa sezione. Viceversa il quarto sistema che presentano segue un approccio simile a quello utilizzato nel nostro studio. Silipo e Greenberg hanno sperimentato la costruzione di funzioni di prominenza combinando i quattro parametri acustici in vario modo, utilizzando però come operatore di combinazione unicamente l'operazione di moltiplicazione. I parametri e le soglie utilizzati nella definizione della funzione di prominenza sono fissati con un'opportuna fase di apprendimento. I migliori risultati ottenuti con questo metodo riguardano una funzione di prominenza basata sul prodotto della durata dei nuclei e dell'intensità, risultati che, pur trascurando completamente il contributo delle curve intonative e quindi dei *pitch accent*, sono migliori di quelli ottenuti dagli altri metodi che gli stessi autori presentano, anche se ancora inferiori a tutti gli altri sistemi discussi finora, compreso quello descritto in questo lavoro (BCPF).

Batliner *et alii*, (1997) hanno presentato un lavoro sulla lingua tedesca all'interno del progetto VERBMOBIL. 276 parametri acustici vengono definiti in parte estraendoli dalle caratteristiche acustiche degli enunciati e in parte utilizzando la trascrizione degli stessi e metodi di ASR. Una rete MLP viene addestrata, utilizzando questi parametri, a riconoscere la presenza di prominenza all'interno delle parole (non delle sillabe) ottenendo un risultato di corrette classificazioni (prominente/non prominente) nell'82.4% dei casi esaminati. Gli autori sottolineano come le prestazioni più alte siano state ottenute utilizzando il maggior numero di parametri a disposizione.

Il confronto del metodo presentato in questo studio con quelli sviluppati nei lavori descritti evidenzia sostanzialmente alcuni punti:

- i metodi automatici che presentano le migliori performance si collocano attorno all'80-85% di corrette classificazioni, in linea con il livello di accordo ottenuto confrontando diversi annotatori umani;
- i metodi in assoluto più performanti utilizzano complesse fasi di apprendimento e richiedono il pesante utilizzo di risorse linguistiche e di informazioni non contenute all'interno dell'enunciato, quali la trascrizione testuale, la trascrizione fonetica allineata, modelli acustici dei foni necessari ai metodi basati su ASR, *part-of-speech tag*, ecc.;
- le prestazioni dei sistemi automatici presentati negli altri studi sono estremamente vicine a quelle dal metodo presentato in questo lavoro (BCPF), basato sulla definizione di una appropriata funzione di prominenza che non richiede alcuna informazione aggiuntiva al campione sonoro dell'enunciato, non necessita di alcuna fase di apprendimento e soprattutto esegue il processo di analisi in tutte le sue fasi, segmentazione e classificazione.

8. ALCUNI ESPERIMENTI SULL'ITALIANO

Alla luce dei risultati ottenuti sulla lingua inglese americana, invero piuttosto incoraggianti, potrebbe essere di un certo interesse proiettare il lavoro effettuato in una prospettiva interlinguistica tentando di trasporre le metodologie sviluppate in un differente sistema linguistico.

A questo scopo è stata avviata un'indagine sull'identificazione automatica della prominenza nella lingua italiana: ci sembra quindi opportuno presentare i risultati ottenuti, seppur estremamente preliminari, applicando le metodologie descritte a dati tratti da un piccolo *corpus* di lingua italiana.

8.1. Quadro bibliografico

In modo simile all'inglese, l'olandese, lo spagnolo, ..., l'italiano è ascrivibile all'insieme delle lingue definite come *stress accented*, nel senso che la prominenza metrica o frasale

viene indicata per mezzo di una combinazione dei contributi derivati dalla frequenza fondamentale, dall'intensità e dalla durata delle unità segmentali che compongono il nucleo sillabico. Le considerazioni e i risultati che abbiamo ottenuto per la lingua inglese sono, in linea di principio, compatibili con la visione che si ha della lingua italiana nei confronti di questi fenomeni. Va segnalato che, a differenza di quanto si rileva per la lingua inglese, gli studi mirati alla determinazione delle correlazioni tra parametri acustici e prominenza applicati alla lingua italiana sono in numero nettamente inferiore. Di conseguenza il panorama risulta essere meno definito e, al momento e a conoscenza dell'autore, non pare esserci una teoria di riferimento completa dalla quale attingere spunti e linee guida.

Due studi sembrano essere considerati come maggiormente rilevanti: Bertinetto (1980; 1981) ha analizzato l'impatto dei vari parametri acustici a livello lessicale predisponendo un numero rilevante di esperimenti percettivi che, basandosi sull'opposizione "papà"/ "papa", gli hanno permesso di effettuare interessanti osservazioni sulla realizzazione della prominenza all'interno di parole isolate. Bertinetto conclude che il ruolo della durata nell'ambito di questi fenomeni è decisamente superiore rispetto a quello dell'intensità o della frequenza fondamentale. Sebbene estremamente interessanti, le conclusioni di Bertinetto non ci consentono di trarre informazioni definitive per questo studio, che si pone in una prospettiva più ampia di quella lessicale, tentando di identificare i parametri che governano la realizzazione fonetico/acustica della prominenza a livello di enunciato. In questo senso il lavoro di D'Imperio (2000), analizzando i contributi dei vari parametri acustici nella percezione della prominenza negli enunciati, sebbene utilizzi un enunciato estremamente ridotto nelle sue componenti sintattiche ("Mario esce"), giunge in prima approssimazione alle medesime conclusioni di Bertinetto. Nella lingua italiana la durata delle unità segmentali sembra assumere un ruolo predominante nell'individuazione della prominenza, anche se i contributi degli altri parametri acustici e dei fenomeni linguistici che essi supportano (*stress* e *pitch accent*) non vengono certamente considerati nulli. A questo proposito ci pare opportuno segnalare i lavori di Avesani (1995) e D'Imperio (2002) per quanto riguarda l'analisi dei contributi alla prominenza dovuti all'intonazione, in particolare nell'indagine dei profili intonativi che generano i *pitch accent*.

Per quanto riguarda questo studio, ci atterremo sostanzialmente alle discussioni effettuate nei capitoli precedenti relativamente alla lingua inglese per definire gli schemi metodologici che andremo ad applicare, salvo poi riconsiderare questi aspetti nella sezione che esaminerà i risultati ottenuti sulla lingua italiana.

8.2. I dati

Per mantenere il livello di analisi confrontabile col lavoro svolto per la lingua inglese è sembrato necessario riferirsi a risorse contenenti campioni di parlato letto.

I *corpora* disponibili all'autore non si prestano particolarmente al tipo di ricerca che abbiamo effettuato. Il primo, per ragioni linguistiche, in quanto costituito da frasi costruite artificialmente che spesso non hanno senso compiuto e il secondo per le modalità di stratificazione diatopica che rendono la risorsa estremamente eterogenea e il contenuto di tipo dialogico che la rende poco adatta ai nostri scopi, almeno in questa fase preliminare dell'analisi della lingua italiana.

Si è quindi pensato di creare un piccolo *corpus* sperimentale che assolvesse appieno le necessità di questo studio, in particolare non differisse troppo nelle tipologie degli enunciati dal *corpus* TIMIT che abbiamo utilizzato nello studio sulla lingua inglese americana.

Sfruttando le possibilità offerte dal sito RAI sono stati registrati alcuni "giornali radio" (GR1 e GR2): le registrazioni sono state suddivise in enunciati, per un totale di 29 ripartiti tra 4 locutori femmine e 5 maschi, sono stati segmentati manualmente, sillabati e annotati

col livello di prominenza percepito, ottenendo un totale di 859 sillabe. Chiameremo questo *corpus* "RADIO".

Le dimensioni del *corpus* sono estremamente limitate, ma, come vedremo nella sezione successiva, consentiranno ugualmente un'analisi della prominenza nella lingua italiana.

8.3. I risultati ottenuti

Ci preme sottolineare che, date le dimensioni ridotte del campione utilizzato e l'elaborazione relativamente recente dei risultati, ogni conclusione è da considerarsi provvisoria e suscettibile di ulteriori approfondimenti.

Ripercorrendo l'analisi effettuata per la lingua inglese americana, il primo passo riguarda un'analisi qualitativa del comportamento delle distribuzioni di sillabe prominenti e non prominenti rispetto ai quattro parametri acustici considerati in questo studio. Le distribuzioni dei parametri acustici all'interno dei nuclei sillabici per i due fenomeni linguistici analizzati sembrano essere simili a quelle ottenute per la lingua inglese, mostrando sostanzialmente una generale convergenza nell'uso dei parametri acustici per supportare la prominenza nelle due lingue esaminate. Le relazioni tra le due coppie di parametri acustici nel sostenere i relativi fenomeni linguistici, secondo la gerarchia proposta nella tabella 1, sembrano permanere immutate, suggerendo, anche in questo caso, una funzione di prominenza simile a quella indicata nella sezione 7.1 per la lingua inglese americana.

La tabella 4 mostra i risultati ottenuti utilizzando la funzione e le metodologie descritte precedentemente (BCPF) per identificare la prominenza negli enunciati del *corpus* RADIO.

Lingua italiana				
Metodo	**Tipo**	**Accuracy**	**Recall**	**Precision**
BCPF	Non Superv.	80.32%	70.00%	58.10%

Tabella 4: Risultati della sperimentazione del metodo proposto (BCPF), applicato al *corpus* di lingua italiana.

I risultati ottenuti per la lingua italiana confermano essenzialmente quelli ottenuti nell'analisi della lingua inglese americana attraverso il *corpus* TIMIT. Le percentuali di corrette classificazioni, superando l'80% dei casi esaminati, ci consentono di confermare le considerazioni e le valutazioni effettuate nella sezione 7.3 a proposito dei risultati ottenuti dal classificatore per la lingua inglese.

9. CONCLUSIONI E PROSPETTIVE PER IL FUTURO

L'obiettivo principale di questo lavoro è stato la costruzione di un sistema di identificazione automatica della prominenza frasale nel linguaggio parlato continuo. I vincoli posti in fase di definizione del lavoro hanno riguardato principalmente l'esclusione dell'uso di informazioni aggiuntive, quali trascrizioni fonetiche o ortografiche dell'enunciato, informazioni sintattiche o segmentazioni dell'enunciato in unità allineate temporalmente. Inoltre, per varie ragioni esposte nella sezione introduttiva, si è scelto di evitare metodi e schemi teorici che richiedessero l'uso di algoritmi legati a fasi di apprendimento basate su dati autentici etichettati manualmente.

Sulla base dell'analisi bibliografica nel settore sia linguistico sia modellistico/tecnologico sono stati identificati alcuni parametri fonetico/acustici in grado di supportare il fenomeno della prominenza e sono stati definiti opportuni algoritmi per

misurare i valori di tali parametri a partire dalla componente acustica degli enunciati. In particolare è stato analizzato estensivamente il problema dell'identificazione automatica dei nuclei sillabici che risultano essere le unità segmentali rilevanti ove effettuare le misure dei parametri acustici necessari all'identificazione della prominenza. Per poter estrarre queste unità segmentali è stato sviluppato un algoritmo in grado di determinarne posizione e confini all'interno dell'enunciato in modo sufficientemente affidabile da consentire le elaborazioni richieste.

Alcuni modelli teorici di grande potenza e attualità (reti neurali e *support-vector machine*) sono stati estensivamente sperimentati per classificare le unità segmentali rispetto ai rispettivi livelli di prominenza utilizzando i parametri fonetico/acustici individuati. Un modello radicalmente differente, che non richiede alcuna fase di apprendimento su dati autentici etichettati, è stato proposto dall'autore; esso consiste nella definizione di un'opportuna funzione di prominenza a valori continui in grado di valutare il livello di prominenza di un nucleo sillabico basandosi su quattro parametri acustici ricavati utilizzando unicamente le informazioni contenute nella *waveform* dell'enunciato. E' stato presentato anche un metodo che, basandosi sulla funzione di prominenza, identifica quali sillabe siano effettivamente prominenti, valutando, per ognuna di esse, il contesto nella quale è inserita.

Le metotologie proposte per affrontare il problema, sia dal punto di vista dell'identificazione dei nuclei sillabici, sia della determinazione della prominenza consentono di identificare le sillabe prominenti con un accordo con annotatori umani confrontabile con quello ottenibile tra esperti del settore. Le percentuali di corretta classificazione sono molto incoraggianti e, lungi dal dichiarare che il lavoro sia completo, riteniamo che gli obiettivi posti all'inizio del lavoro siano stati compiutamente raggiunti.

Di estremo interesse sembra essere la prospettiva interlinguistica che lo studio può assumere. I risultati ottenuti sulla lingua italiana, ancorché preliminari, possono far supporre un certo livello di regolarità tra le lingue *stress accented*. Lo studio di estensioni del metodo BCPF in questa prospettiva potrebbe portare a interessanti conclusioni sulla natura della prominenza come fenomeno universale, almeno per questo gruppo di lingue. Questo risulta particolarmente significativo in quanto, come abbiamo più volte sottolineato nei capitoli bibliografici, la prominenza, nella sua manifestazione acustica, assume differenti connotazioni a seconda della lingua in esame e, benché alcune lingue sembrino mostrare tratti comuni, non risulta finora possibile definire una matrice costante tra di esse.

Una possibile prospettiva di questa indagine potrebbe essere quella di estendere la funzione di prominenza nel modo seguente

$$\mathrm{Prom}^{\mathrm{i}} = \max\left\{(en^{i}_{300-2200})^{\alpha}\cdot(dur^{i})^{\beta}\ ,\ (en^{i}_{glob})^{\gamma}\cdot(A^{i}_{event}\cdot D^{i}_{event})^{\delta}\right\} \qquad (2)$$

ossia introducendo quattro esponenti in grado di "pesare" i contributi dei parametri acustici che abbiamo utilizzato nell'identificazione della prominenza, che raccoglieremo in un vettore $\boldsymbol{w}_p$. In questo modo, alla lingua inglese americana e alla lingua italiana si potrebbero far corrispondere gli assegnamenti dei pesi seguenti

$$\boldsymbol{w}_p = (\alpha = 1.0,\ \beta = 1.0,\ \gamma = 1.0,\ \delta = 1.0)$$

che, sulla base degli esperimenti effettuati, generano una funzione di prominenza in grado di produrre buoni risultati, classificando correttamente più dell'80% delle sillabe esaminate.

In questa prospettiva, l'equazione 2 rappresenterebbe una componente universale nella definizione della prominenza nelle lingue *stress accented*, mentre il vettore dei pesi $\boldsymbol{w}_p$ permetterebbe di adattare il comportamento della funzione *Prom* alla specifica lingua,

modificando i rapporti tra i vari parametri acustici fino a rimuovere i contributi di alcuni di essi, nell'ipotesi che i quattro parametri identificati siano tutti e soli quelli universalmente coinvolti nel supportare il fenomeno in esame.

Il quadro che si va delineando, ancorché contenente ipotesi non ancora soddisfacentemente validate, aprirebbe prospettive affascinanti nella definizione di questo tipo di fenomeni, pertanto risulta, nell'opinione dell'autore, estremamente utile e interessante approfondire l'analisi in questa direzione.

Un'ulteriore dimensione dell'analisi che sarebbe senz'altro interessante approfondire è la trasposizione dei risultati ottenuti sul linguaggio parlato spontaneo. Tutti gli esperimenti effettuati in questo studio hanno coinvolto enunciati di linguaggio parlato letto: le condizioni espressive e comunicative del parlato spontaneo sono relativamente differenti da quelle nelle quali sono stati registrati i campioni utilizzati, quindi sembra opportuna una verifica delle conclusioni ottenute su *corpora* di linguaggio parlato spontaneo, al fine di assicurarsi della estendibilità dei risultati ottenuti nello studio anche in questo dominio d'analisi.

Abbiamo più volte sottolineato come la prominenza sia un fenomeno essenzialmente percettivo, quindi, sempre dal punto di vista tecnologico, sembrerebbe opportuno approfondire i rapporti tra questo fenomeno e modelli uditivo/percettivi umani (Seneff 1988). Introdurre opportuni correttivi nelle misurazioni dei parametri acustici nella direzione della percezione umana potrebbe migliorare notevolmente le prestazioni globali del sistema automatico di classificazione della prominenza accentuale.

BIBLIOGRAFIA

Anastasakos, A., Schwartz, R. & Shu, H. (1995), Duration modeling in large vocabulary speech recognition, in *Proc. of ICASSP '95*, 628-631.

Andre-Obrecht, R. (1988), A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals, *IEEE Transactions on Acoustics, Speech and Signal processing*, 29-40.

Avesani, C. (1995), ToBIt: un sistema di trascrizione per l'intonazione italiana, in *Proc. of Atti delle 5e Giornate di Studio del Gruppo di Fonetica Sperimentale*, Povo, Italy, 85-98.

Bagshaw, P.C. (1994), *Automatic prosodic analysis for computer-aided pronunciation teaching*, PhD thesis, University of Edinburgh.

Batliner, A., Kießling, A., Kompe, R., Niemann, H. & Nöth, E. (1997), Can we tell apart intonation from prosody (if we look accents and boundaries)?, in *Proc. of ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, 39-42.

Batliner, A., Möbius, B., Möler, G., Schweitzer, A. & Nöth, E. (2001), Prosodic models, automatic speech understanding and speech syntesis: toward the common ground, in *Proc. of Eurospeech 2001*, Aalborg, Denmark, 2285-2288.

Beckman, M.E. (1986), *Stress and Non-stress Accent*, Dordrecht, Holland: Foris Publications.

Beckman, M.E. & Venditti, J.J. (2000), Tagging prosody and discourse structure in elicited spontaneous speech, in *Proc. of Science and Technology Agency Priority Program Symposium on Spontaneous Speech: Corpus and Processing Technology*, Tokyo, 87-98.

Bertinetto, P.M. (1980), The perception of stress by Italian speakers, *Journal of Phonetics*, 385-395.

Bertinetto, P.M. (1981), *Strutture prosodiche dell'italiano*, Firenze: Accademia della Crusca.

Bitar, N. & Espy-Wilson, C.Y. (1996), Knowledge-based parameters for HMM speech recognition, in *Proc. of ICASSP '96*, Atlanta, 29-32.

Bolinger, D. (1958), A theory of pitch-accent in English, *Word*, 109-149.

Buhmann, J., Caspers, J., van Heuven, V.J., Hoekstra, H., Martens, J.P. & Swerts, M. (2002), Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus, in *Proc. of LREC 2002*, Las Palmas, Spain, 779-785.

Bulyko, I., Ostendorf, M. & Price, P. (1999), On the Relative Importance of Different Prosodic Factors in Improving Speech Synthesis, in *Proc. of ICPhS '99*, San Francisco, 81-84.

Campione, E. & Veronis, J. (1998), A multilingual prosodic database, in *Proc. of ICSLP '98*, Sydney, 3163-3166.

Chomsky, N. & Halle, M. (1968), *The Sound Pattern of English*, Harper and Row.

Chun, D.M. (1998), Signal Analysis Software for Teaching Discourse Intonation, *Language Learning and Technology*, 61-77.

D'Anna, L., Petrillo M., Zovato E. (2001), Elaborazioni automatiche dei parametri prosodici, nuovi sviluppi di APA, in *Proc. of Giornate di studio del Gruppo di Fonetica Sperimentale*, Macerata.

de Cheveigné, A. & Kawahara, H. (2002), YIN, a fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, 111, 1917-1930.

Delmonte, R. (2000), SLIM prosodic automatic tools for self-learning instruction, *Speech Communication*, 145-166.

D'Imperio, M. (2000), Acoustical-perceptual correlates of sentence prominence in Italian, in, *The Ohio State University Working Papers in Linguistics*, Columbus OH: The Ohio State University, 59-79.

D'Imperio, M. (2002), Italian intonation: An overview and some questions, *Probus*, 37-69.

Eriksson A., Thunberg G.C., Traunmüller H. (2001), Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing, in *Proc. of Eurospeech 2001*, Aalborg, Denmark, 399-402.

Eriksson A., Grabe E., Traunmüller H. (2002), Perception of Syllable Prominence by Listener with and without Competence in the Tested Language, in *Proc. of Speech Prosody 2002*, Aix-en-Provence.

Fant G., Kruckenberg A., Liljencrants J. (2000), Acoustic-phonetic Analisys of Prominence in Swedish, in Botinis A. (Ed.), *Intonation*, Kluwer, 55-86.

Fry D.B. (1958), Experiments in the perception of stress, *Language and Speech*, 126-152.

Gadde, Venkata Ramana Rao (2000), Modeling word duration for better speech recognition, in *Proc. of Speech Transcription Workshop*, Maryland.

Gallwitz F., Niemann H., Nöth E., Warnke V. (2002), Integrated recognition of words and prosodic phrase boundaries, *Speech Communication*, 36, 81-95.

Garofolo J.S., Lamel L.F., Fisher W.M., Fiscus J.G., Pallett D.S., Dahlgren N.L. (1993), *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM,* NIST.

Gimson A.C. (1980), *An introduction to the pronunciation of English,* London: Edward Arnold.

Glass J. & Zue V. (1988), Multi-level acoustic segmentation of continuous speech, in *Proc. of ICASSP '88*, New York, 429-432.

Goslin J., Content A., Frauenfelder U.H. (1999), Syllable segmentation: are humans consistent?, in *Proc. of Eurospeech '99*, Budapest.

Greenberg S., Carvey H., Hitchcock L., Chang S. (2003), The Phonetic Patterning of Spontaneous American English Discourse, in *Proc. of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo.

Hastie H.W., Poesio M., Isard S. (2001), Automatically predicting dialog structure using prosodic features, *Speech Communication*, 63-79.

Heldner M. (1998), Is an F0-rise a necessary or a sufficient cue to perceived focus in Swedish?, in Werner, S. (Ed.), *Nordic Prosody: Proceedings of the VIIth Conference*, Frankfurt am Main: Peter Lang, 109-125.

Heldner M. (2001), Spectral Emphasis as an Additional Source of Information in Accent Detection, in *Proc. of Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, 57-60.

Heldner M. (2003), On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish, *Journal of Phonetics*, 39-62.

Hieronymus J.L., McKelvie D., McInnes F.R. (1992), Use of acoustic sentence-level and lexical stress in HMM speech recognition, in *Proc. of ICASSP '92*, San Francisco, 225-229.

Hirshberg J., Avesani C. (2000), Prosodic disambiguation in English and Italian, in Botinis A. (Ed.), *Intonation*, Kluwer, 87-95.

Hirst D. (2001), Automatic analysis of prosody for multilingual speech corpora, in Keller E., Bailly G., Terken J., Huckvale M. (Ed.), *Improvements in Speech Synthesis*, Chichester, UK: Wiley.

Howitt A.W. (2000), Vowel Landmark Detection, in *Proc. of ICSLP 2000*, Beijing, 628-631.

Jenkin K.L., Scordilis, M.S. (1996), Development and Comparison of Three Syllable Stress Classifiers, in *Proc. of ICSLP '96*, Philadelphia, 733-736.

Kopecek, I. (1999). Speech recognition and syllable segments, in *Proc. of Workshop on Text, Speech and Dialogue - TSD '99*, LNAI 1692, 203-208.

Ladd D.R., Verhoeven J., Jacobs K. (1994), Influence of adjacent pitch accents on each other perceived prominence: two contradictory effects, *Journal of Phonetics*, 87-99.

Ladd D.R., Faulkner D., Faulkner, H., Schepman A. (1999), Constant 'segmental anchoring' of F0 movements under changes in speech rate, *Journal of the Acoustical Society of America*, 1543-1554.

Ladd D.R., Shepman A. (2003), 'Sagging transitions' between high pitch accents in English: experimental evidence, *Journal of Phonetics*, 81-112.

Liu S. (1996), Landmark detection for distinctive feature-based speech recognition, *Journal of the Acoustical Society of America*, 3417-3430.

Menzel W., Herron D., Morton R., Pezzotta D., Bonaventura P., Howarth P. (2001), Interactive pronunciation training, *ReCALL*, 13, 67-78.

Mermelstein P. (1975), Automatic segmentation of speech into syllabic units, *Journal of the Acoustical Society of America*, 880-883.

Mixdorff H., Jokisch O. (2003), Evaluating the Quality of an Integrated Model of German Prosody, *International Journal of Speech Technology*, 45-55.

Neumeyer L., Franco H., Weintraub M., Price P. (1996), Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech, in *Proc. of ICSLP '96*, Philadelphia, 1457-1460.

Neumeyer L., Franco H., Abrash L.J., Ronen O., Bratt H., Bing J., Digalakis V., Rypa M. (1998), WebGrader: A Multilingual Pronunciation Practice Tool, in *Proc. of ESCA - STiLL '98*, Marholmen, Sweden, 61-64.

Noetzel A. (1991), Robust syllable segmentation of continuous speech using neural networks, in *Proc. of IEEE Electro International Conference Record*, New York, 580-585.

Nöth, E., Batliner A., Kießling A., Kompe R., Niemann H. (2000), VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System, *IEEE Transactions on Speech and Audio Processing*, 519-532.

Pfitzinger H., Burger S., Hid S. (1996), Syllable detection in read and spontaneous speech, in *Proc. of ICSLP '96*, Philadelphia, 1261-1264.

Pickering B., Williams B., Knowles G. (1996), Analysis of transcriber differences in SEC, in Knowles G., Wichmann, A. & Alderson, P. (Ed.), *Working with speech*, London: Longman, 61-86.

Portele T., Heuft B. (1997), Towards a prominence-based syntesis system, *Speech Communication*, 61-72.

Rietveld T., Kerkhoff J. (2002), The Temporal Alignment of L*H Accents, in *Proc. of Speech Prosody 2002*, Aix-en-Provence.

Seneff S. (1988), A joint Syncrony/Mean-Rate Model of Auditory Speech Processing, *Journal of Phonetics*, 55-76.

Shriberg E., Stolcke A., Hakkani-Tür D., Tür G. (2000), Prosody-based automatic segmentation of speech into sentences and topics, *Speech Communication*, 32, 127-154.

Shriberg E., Stolcke A. (2001), Prosody modeling for automatic speech recognition and understanding, in *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, 13-16.

Silipo R., Greenberg S. (1999), Automatic transcription of prosodic stress for spontaneous English discourse, in *Proc. of ICPhS '99*, San Francisco, 2351-2354.

Sluijter A., van Heuven V., Pacilly J. (1997), Spectral balance as a cue in the perception of linguistic stress, *Journal of the Acoustical Society of America*, 503-513.

Sluijter A., van Heuven V. (1996a), Spectral balance as an acoustic correlate of linguistic stress, *Journal of the Acoustical Society of America*, 2471-2485.

Sluijter A., van Heuven V. (1996b), Acoustic correlates of linguistic stress and accent in Dutch and American English, in *Proc. of ICSLP '96*, Philadephia, 630-633.

Stevens K.N., Manuel S.Y., Shattuck-Hufnagel S., Liu S. (1992), Implementation of a model for lexical access based on features, in *Proc. of ICSLP '92*, Banff, 499-502.

Streefkerk B.M. (1996), Prominent accent and pitch movements, *Inst. of Phon. Sciences Proceedings, University of Amsterdam*, 111-119.

Streefkerk B M., Pols L.C.W., ten Bosch L.F.M. (1999), Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's, in *Proc. of Eurospeech '99*, Budapest, 551-554.

Strom V. (1995), Detection of Accents, Phrase Boundaries and Sentence Modality in German with Prosodic Features, in *Proc. of Eurospeech '95*, Madrid, 2039-2041.

Suomi K., Toivanen J., Ylitalo R. (2003), Durational and tonal correlates of accent in Finnish, *Journal of Phonetics*, 113-138.

Talkin D. (1995), A robust algorithm for pitch tracking (RAPT), in Kleijn W.B., Paliwal K.K. (Ed.), *Speech coding and synthesis*, New York: Elsevier, 495-518.

Taylor P.A. (1992), *A phonetic model of English intonation*, PhD thesis, University of Edimburgh.

Taylor P.A. (1993), Automatic Recognition of Intonation from F0 Contours using the Rise/Fall/Connection Model, in *Proc. of Eurospeech '93*, Berlin.

Taylor P.A. (2000), Analysis and Synthesis of Intonation using the Tilt Model, *Journal of the Acoustical Society of America*, 1697-1714.

Terken J. (1991), Fundamental frequency and perceived prominence, *Journal of the Acoustical Society of America*, 1768-1776.

Vallabha G.K., Tuller B. (2002), Systematic errors in the formant analysis of steady-state vowels, *Speech Communication*, 141-160.

van Bergem D. (1993), Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels, *Speech Communication*, 1-23.

van Kuijk D., Boves L. (1999), Acoustic characteristic of lexical stress in continuous telephone speech, *Speech Communication*, 95-111.

van Santen J. (2002), Quantitative Modelling of Pitch Accent Alignment, in *Proc. of Speech Prosody 2002*, Aix-en-Provence.

van Santen J., Möbius B. (1997), Modeling pitch accent curves, in *Proc. of ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens.

Vereecken H., Martens J., Grover C., Fackrell J., Van Coile B. (1998), Automatic prosodic labeling of 6 languages, in *Proc. of ICSLP '98*, Sydney, 1399-1402.

Warren P. (1996), Prosody and Parsing: an introduction, *Language and Cognitive Processes*, 1-16.

Wichmann A., House J., Rietveld T. (2000), Discourse Constraints on F0 Peak Timing in English, in Botinis A. (Ed.), *Intonation*, Kluwer, 163-182.

Wightman C.W., Syrdal A.K., Stemmer G., Conkie A. (2000), Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis, in *Proc. of ICSLP 2000*, Beijing, 71-74.

Wightman C.W. (2002), ToBI or not ToBI?, in *Proc. of Speech Prosody 2002*, Aix-en-Provence.

Wightman C.W., Ostendorf M. (1994), Automatic Labeling of Prosodic Patterns, *IEEE Transactions on Speech and Audio Processing*, 469-481.

Wu S, Shire M.L., Greenberg S., Morgan N. (1997), Integrating syllable boundary information into speech recognition, in *Proc. of ICASSP '97*, Munich, 987-990.

Xu Y. (2002), Articulatory Constraints and Tonal Alignment, in *Proc. of Speech Prosody 2002*, Aix-en-Provence.