

GLI ASPETTI PROSODICI DELL'ITALIANO

ATTI DELLE 4^e GIORNATE DI STUDIO DEL GRUPPO DI
FONETICA SPERIMENTALE (A.I.A.)

A cura di P.L. SALZA

COLLANA DEGLI ATTI DELL'ASSOCIAZIONE ITALIANA DI ACUSTICA

VOLUME XXI

1994



TORINO 11 - 12 Novembre 1993

VALUTAZIONE DELLA PROSODIA DI UN SINTETIZZATORE DA TESTO IN UN SISTEMA INTERATTIVO DI DIALOGO

Pier Luigi Salza, Giuseppe Di Fabrizio, Mario Oreglia

(Servizi e Tecnologie Audio CSELT - Torino)

Mauro Falcone, Ciro Sementina, Cristina Delogu

(Fondazione U. Bordini - Roma)

1. INTRODUZIONE

Il Sistema di Dialogo uomo-macchina sviluppato in CSELT è stato recentemente sperimentato nell'interrogazione dell'orario ferroviario tramite telefono. Il Sistema di Dialogo si compone di una parte relativa alla comprensione del parlato continuo [1], una parte che svolge la gestione del dialogo [2] e una che esegue la sintesi da testo. Il sistema di sintesi da testo (TTS - Text-To-Speech) per l'italiano sviluppato in CSELT, ELOQUENS[®][3], è costituito da tre moduli principali: il Modulo per il Trattamento del Testo, il Modulo per la generazione della Prosodia e il Modulo Acustico. E' stato dimostrato che nel dialogo la prosodia svolge un ruolo cruciale tanto nel convogliare l'informazione quanto nel guidare l'interazione [4], in primo luogo attraverso la varietà delle modalità intonative di frase utilizzate. Le regole prosodiche del TTS [5] cercano di correlare la durata del fonema e la curva di F_0 alla struttura fonetica, al tipo di sillaba, alla sintassi e ad altre caratteristiche linguistiche del messaggio fornite dal modulo per il trattamento del testo. Le regole di durata sono basate sul principio di sovrapposizione degli effetti e sulla possibilità di modificare la velocità di eloquio. Le regole assegnano separatamente la durata intrinseca dei fonemi e i coefficienti di variazione della durata per effetto del contesto fonetico e del contesto sintattico. Le regole di intonazione stabiliscono una sequenza di configurazioni di tono in base alla modalità di frase e alla struttura prosodico-sintattica, e convertono ogni configurazione in una sequenza di "movimenti di tono" stilizzati. Attualmente il TTS rende disponibili sette diverse modalità di frase in modo da soddisfare le esigenze di una situazione interattiva: frasi di comando (CO),

es. "Formula la tua richiesta.", frasi dichiarative semplici (DE), es. "Non ci sono collegamenti tra Milano Centrale e Lecce nell'arco di tutta la giornata."; elenchi (EL), es. "Il primo collegamento parte alle ore 15 e 19 e arriva alle ore 17 e 31; il secondo parte alle ore 7 e arriva alle 17 e 47."; interrogative "wh" (WH), es. "A che ora vuoi partire?"; interrogative "yn" (YN), es. "Ti interessa un servizio per pernottamento?"; interrogative "yn"-eco (YN+), es. "Sì o No?"; interrogative "yn"-modali (YM), es. "Vuoi cercare i collegamenti ad un cambio tra Milano Centrale e Lecce prima delle ore 7?".

L'esigenza di effettuare una valutazione soggettiva della qualità di ELOQUENS® è sorta come logica conseguenza della sua integrazione nel Sistema di Dialogo uomo-macchina, con l'obiettivo di verificare l'indice di gradimento del potenziale utente nei riguardi della voce sintetica in tale contesto, in particolare per quanto riguarda la prosodia. L'esperimento è stato realizzato nell'ambito del Progetto ESPRIT SAM-A (Speech Assessment Methodologies). Per una descrizione più particolareggiata si veda [6].

2. METODO

Sulla base di una rassegna delle diverse metodologie disponibili per la valutazione della qualità dei sistemi TTS [7], è stato prescelto il metodo soggettivo basato sul Punteggio Medio di Opinione (MOS - Mean Opinion Scores), raccomandato dal CCITT [8]. Il MOS utilizza sette scale di opinione, contenenti opportune definizioni descrittive (aggettivi) graduate in cinque valori, che vanno dalla valutazione migliore (indice 5) a quella peggiore (indice 1). Ciascuna scala di opinione è associata ad una delle seguenti sette caratteristiche della voce: Impressione Globale, Sforzo di ascolto (comprensione a livello del messaggio), Articolazione (comprensione a livello di fonema), Pronuncia (adeguatezza della prosodia), Gradevolezza della Voce, problemi di Comprensione (a livello di parola), Velocità di eloquio (adeguatezza). I giudizi vengono espressi dagli ascoltatori su appositi questionari. Per esempio, la scala di opinione dell'Impressione Globale utilizza i seguenti aggettivi: (5) Ottimo (4) Buono (3) Sufficiente (2) Scarso (1) Insufficiente. La progettazione della valutazione MOS qui illustrata ha comportato tre novità sostanziali rispetto al MOS standard: 1) ai soggetti è stato chiesto di compilare i questionari dopo aver effettuato l'interazione con il sistema di dialogo, durante una apposita sessione di test nella quale sono state fatte loro riascoltare le risposte vocali del sintetizzatore (misura successiva); 2) i messaggi del sintetizzatore sono stati presentati agli ascoltatori all'interno di un breve turno di dialogo (valutazione dipendente dal contesto); 3) relativamente alla prosodia, il parametro Pronuncia, piuttosto generico, è stato sostituito da tre parametri specifici, vale a dire Intonazione, Naturalità e Pause, dando luogo ad un totale di 9 parametri o caratteristiche della voce. Le scale di opinione connesse a questi parametri sono state distribuite in due diversi tipi di questionario, detti tipo A e tipo B, i quali comprendono,

rispettivamente, i seguenti parametri: Tipo A: *Impressione Globale, Sforzo di ascolto, Articolazione, Intonazione, Naturalità, Pause*; Tipo B: *gradevolezza della Voce, problemi di Comprensione, Velocità di eloquio, Intonazione, Naturalità, Pause*.

Uno dei principali obiettivi dell'esperimento è stato il confronto tra le risposte di ascoltatori "interattivi" e "non-interattivi", da un lato, e tra quelle di ascoltatori "naive" ed "esperti", dall'altro. L'esecuzione degli esperimenti ha comportato diverse fasi.

15 soggetti "naive" hanno interagito individualmente con il Sistema di Dialogo. Il loro compito è consistito nell'ottenere informazioni (orari, tariffe, stazioni di arrivo e/o partenza, etc.) sugli orari ferroviari per determinate tratte, mediante un apparecchio telefonico localizzato in una stanza isolata. I soggetti erano equamente distribuiti tra maschi e femmine, l'età variava dai 19 ai 63 anni e nessuno di loro aveva utilizzato servizi telefonici computerizzati prima di allora. Ciascun soggetto ha effettuato un numero di dialoghi variabile da 8 a più di 20, a seconda della diversa capacità di interazione e di dipendenza di vari fattori relativi al Sistema di Dialogo.

L'operatore, seguendo ciascuna interazione, ha selezionato 20 turni di dialogo per ognuno dei 15 soggetti. Si considera turno di dialogo l'insieme di domande-risposte nel seguente ordine: "soggetto - sistema di sintesi - soggetto". La selezione ha incluso pertanto 300 enunciati prodotti dal sintetizzatore, dai quali è stato ricavato un ulteriore sottoinsieme di 100 enunciati. Per ciascuna interazione è stata condotta una scelta equilibrata di modalità di frase.

L'operatore ha quindi preparato un "file di ascolto" per ogni soggetto in cui ognuno dei 20 turni di dialogo selezionati era seguito dall'enunciato prodotto dal sistema di sintesi ripetuto due volte. Durante una successiva sessione, a ciascuno dei soggetti che hanno utilizzato il sistema di Dialogo, che chiameremo **Ascoltatori Interattivi**, è stato fatto riascoltare in cuffia il relativo "file di ascolto" ed è stato chiesto di valutare gli enunciati prodotti dal sintetizzatore sugli appositi questionari MOS, dopo aver letto un foglio di istruzioni.

Lo stesso materiale è stato presentato per la valutazione MOS, in condizioni del tutto simili, a 15 ascoltatori "naive" che non hanno partecipato alla sessione di dialogo e che chiameremo **Ascoltatori Semplici**, sia uomini che donne di età variabile dai 25 ai 55 anni.

Infine, il sottoinsieme di 100 enunciati è stato presentato per la valutazione MOS a 5 ascoltatori "esperti", che chiameremo **Ascoltatori di Controllo**, i quali, al pari dei precedenti, non hanno partecipato alla sessione di dialogo.

In breve, l'impianto complessivo dell'esperimento è il seguente:

- 15 Ascoltatori Interattivi x 20 frasi x 6 giudizi MOS (tipo A o B) = 1800 giudizi MOS
- 15 Ascoltatori Semplici x 20 frasi x 6 giudizi MOS (tipo A o B) = 1800 giudizi MOS

Controllo x 20 frasi x 6 giudizi (tipo A o B) = 600 giudizi MOS.

ATI
 o stati raccolti 4200 giudizi MOS.
 gruppo di ascoltatori sono stati calcolati il valore medio MOS
 Standard per ogni parametro e per ogni tipo di frase. I valori
 sono riportati in forma grafica nelle Figure 1, 2 e 3, rispettiva-
 gli Ascoltatori Interattivi, gli Ascoltatori Semplici e gli
 di Controllo.

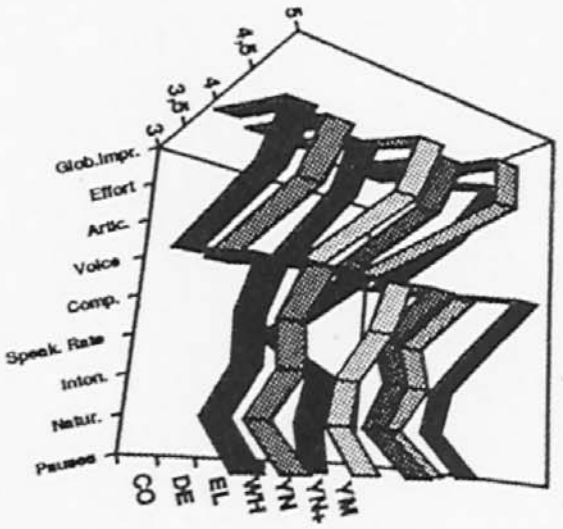


Fig. 1 - MOS mediato tra tutti gli Ascoltatori Interattivi: MOS medio globale = 4.33.
 Osservando queste figure, si può affermare che: 1) la tendenza delle
 vi è una buona
 caratteristiche della voce;
 alla dimensione della frase. I due
 lungo la dimensione tipo di frase. I due
 Figure 4 e 5.

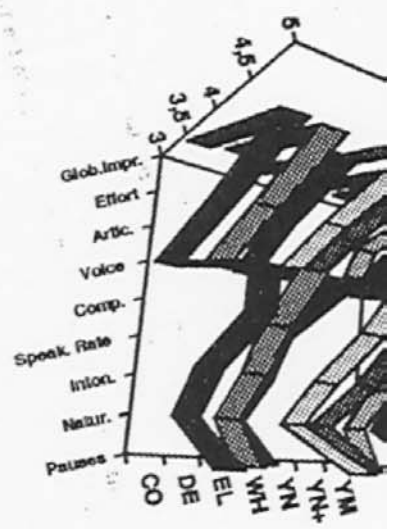


Fig. 2 - MOS mediato tra tutti gli Ascoltatori Semplici: MOS medio globale = 4.
 MOS medio globale = 4.

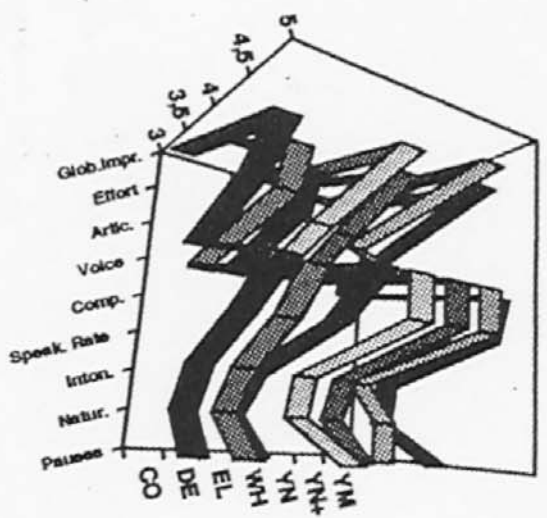


Fig. 3 - MOS mediato tra tutti gli Ascoltatori di Controllo: MOS medio globale = 3.
 MOS medio globale = 3

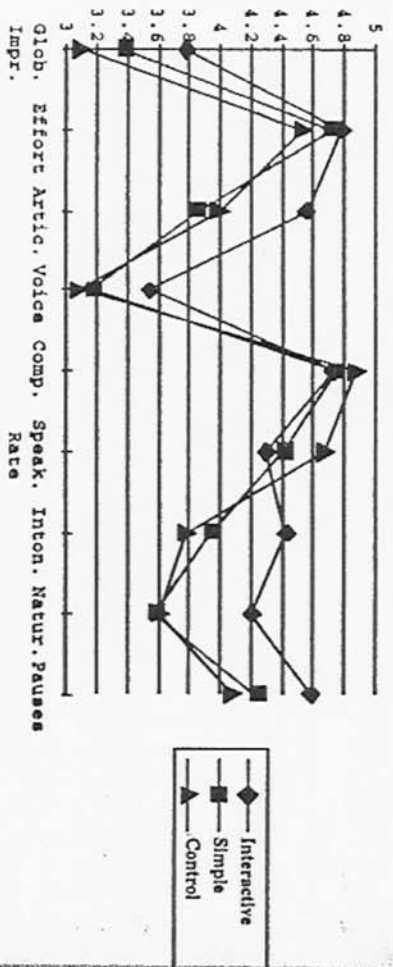


Fig. 4 - Punteggio medio dei giudizi dei tre gruppi di ascoltatori, mediato tra tutti i tipi di frase, per le diverse caratteristiche della voce.

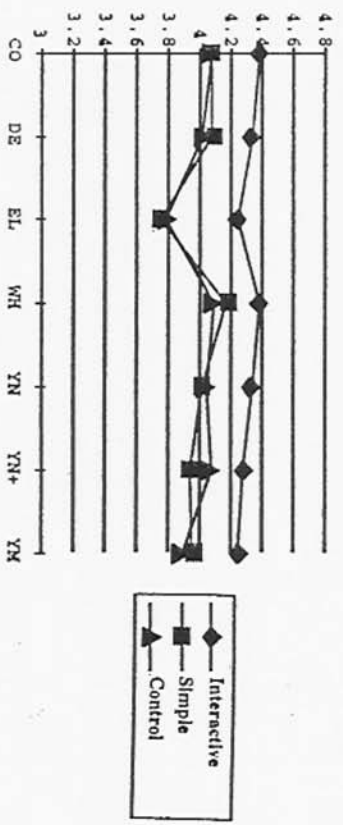


Fig. 5 - Punteggio medio dei giudizi dei tre gruppi di ascoltatori, mediato tra tutte le caratteristiche della voce, per i diversi tipi di frase.

In particolare, dalla Figura 4 si osserva che l'Impressione Globale e la Gradevolezza della Voce, tra loro evidentemente correlate, hanno ricevuto da tutti e tre i gruppi di ascoltatori un MOS più basso rispetto alle rimanenti caratteristiche della voce. Il MOS degli Ascoltatori Semplici e degli Ascoltatori Interattivi, pur essendo più alto di mezzo punto, conserva tuttavia una "forma" molto simile. Dalla Figura 5 emerge che i punteggi dei vari tipi di frase, operando una media tra tutte le caratteristiche della voce, non differiscono tra loro in modo significativo, ad eccezione della modalità EL (frasi elenco). I giudizi degli ascoltatori esperti non differiscono sostanzialmente da

quelli degli ascoltatori naive, mentre esiste una sostanziale diversità di giudizio tra gli ascoltatori che hanno effettuato il dialogo e quelli che non vi hanno partecipato. Sembra quindi che la differenziazione più significativa fra i soggetti sia quella tra gli ascoltatori Interattivi e quelli Non-Interattivi.

4. RISULTATI CONCERNENTI I PARAMETRI PROSDICI

Sono stati confrontati i punteggi dei tre gruppi di ascoltatori (Interattivi, Semplici e di Controllo) calcolati separatamente per i sette tipi di frasi per ognuno dei tre parametri prosodici. I risultati sono riportati nelle Figure 6, 7 e 8, rispettivamente per l'Intonazione, la Naturalità e le Pause. Queste figure confermano che le valutazioni MOS fornite dalle due classi di ascoltatori Non-Interattivi (cioè: Ascoltatori Semplici e Ascoltatori di Controllo) sono molto simili e, in media, più basse rispetto alle valutazioni MOS date dai soggetti Interattivi. La distanza tra i punteggi degli ascoltatori Interattivi e Non-Interattivi per i parametri prosodici è in media di 0.6. Tra le diverse modalità, le frasi EL (Elenco) e, in misura minore, quelle YN+ (interrogative "yn"-eco) hanno ricevuto un punteggio basso dagli ascoltatori Non-Interattivi per tutti e tre i parametri prosodici.

La modalità EL ha ricevuto un punteggio basso per la Naturalità e le Pause anche dagli ascoltatori Interattivi. Gli ascoltatori di Controllo hanno fornito punteggi bassi anche alla modalità YM (interrogative "yn"-modali).

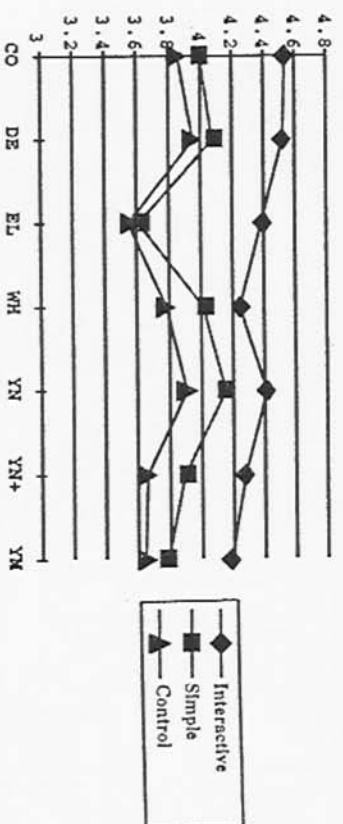


Fig. 6 - Punteggi del parametro prosodico Intonazione dati rispettivamente dagli Ascoltatori Interattivi, dagli Ascoltatori Semplici e dagli Ascoltatori di Controllo ai diversi tipi di frase.

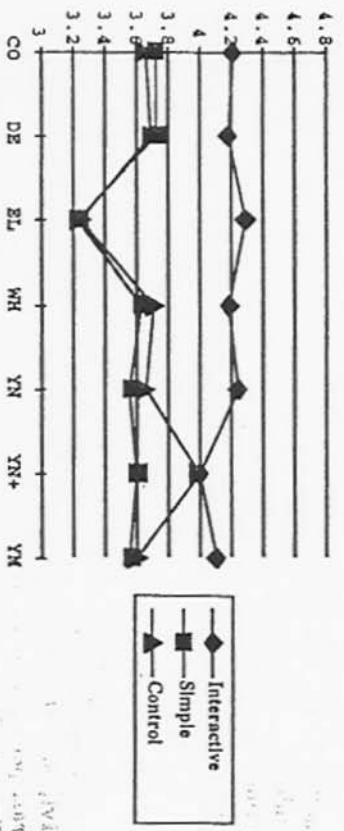


Fig. 7 - Punteggi del parametro prosodico *Naturalzza* dati rispettivamente dagli Ascoltatori Interattivi, dagli Ascoltatori Semplici e dagli Ascoltatori di Controllo ai diversi tipi di frase.

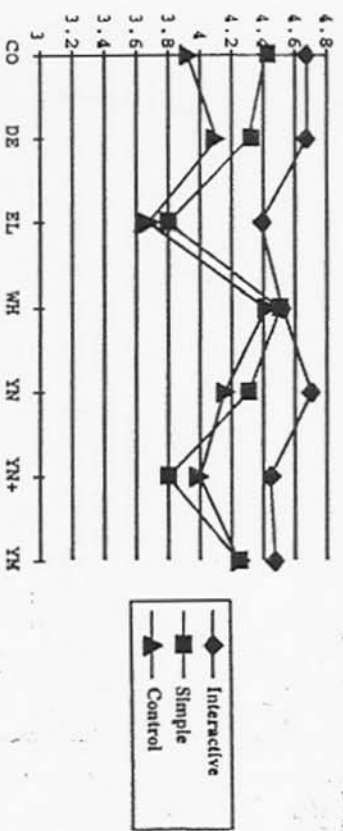


Fig. 8 - Punteggi del parametro prosodico *Pause* dati rispettivamente dagli Ascoltatori Interattivi, dagli Ascoltatori Semplici e dagli Ascoltatori di Controllo ai diversi tipi di frase.

5. VALIDAZIONE STATISTICA DEL MOS

E' stata effettuata un'analisi della Varianza (ANOVA) per dati non bilanciati [9] dei MOS medi di ogni gruppo di ascolto. Il calcolo è stato eseguito tenendo conto del "tipo di frase" quale variabile separata. L'analisi mette in evidenza che la variabile *ascoltatore* è sempre significativa (i valori di F sono compresi tra 1.77 e 22.77 mentre i corrispondenti valori di p variano tra .0001 e .1547) mentre la variabile *tipo di frase* è significativa solo in alcuni casi, prevalentemente nei dati degli Ascoltatori Semplici (i valori di F sono compresi tra 0.24 e 4.34 mentre i corrispondenti valori di p variano tra .0003 e

.9609). Le distribuzioni cumulative dei giudizi MOS evidenziano comportamenti differenti: gli Ascoltatori Interattivi hanno concentrato il 55% dei loro giudizi sul punteggio 5 e circa il 28% sul 4, mentre le percentuali dei punteggi più bassi sono drasticamente minori. Sia gli Ascoltatori Semplici sia quelli di Controllo hanno distribuito i loro giudizi in modo equivalente (35%) sui punteggi 5 e 4, mentre le percentuali dei punteggi più bassi diminuiscono in modo lineare. Dall'analisi ANOVA a una via è emerso che in tutti e tre i confronti possibili, Interattivi verso Semplici, Interattivi verso Controllo, Semplici verso Controllo, F è bassa, compresa tra 0.0108 e 1.7187, e p è compresa tra .2094 e .9183. I giudizi sono consistenti per tutti e tre i gruppi. Il valore più basso di F, vicino a zero, si ha nel test Semplice verso Controllo, dove i dati risultano altamente correlati.

6. CONCLUSIONI

I risultati più importanti di tale esperimento possono essere così riassunti:
 - il metodo MOS ha mostrato dati consistenti tra i tre gruppi di ascoltatori;
 - la distinzione tra soggetti "esperti" e "naïve" non è rilevante mentre, piuttosto, le opinioni degli ascoltatori sembrano influenzate positivamente dalla loro partecipazione ai dialoghi;
 - un'analisi più approfondita dei punteggi relativi ai parametri prosodici rivela che le frasi EL (Elenco) e, in misura minore, le frasi YN+ (interrogative "yn"-eco) e YM (interrogative "yn"-modali) vengono pronunciate dal sintetizzatore in modo non del tutto soddisfacente;
 - per quanto riguarda i valori medi assoluti dei punteggi MOS, il sistema di sintesi CSELT ELOQUENS® ha ottenuto punteggi più alti nelle situazioni di dialogo, persino dai soggetti Non-Interattivi, rispetto a quanto è avvenuto in semplici test di ascolto [10]. Un confronto tra i punteggi medi MOS ottenuti dal TTS in situazioni di dialogo e di non-dialogo è riportato nella seguente tabella.

DIALOGO		NON-DIALOGO
Ascoltatori Interattivi	Ascoltatori Non-Interattivi (Semplici - di Controllo)	Ascoltatori Non-Interattivi
4.33	4 - 3.97	3.66

Poiché il sistema non è stato sostanzialmente migliorato nel periodo di tempo intercorrente tra i due esperimenti, il confronto suggerisce che l'accettabilità della sintesi migliora in una applicazione interattiva grazie, probabilmente, al maggior coinvolgimento dell'utente.

NOTA - L'esperimento, parzialmente sovvenzionato dalla CEE nell'ambito del Progetto ESPRIT SAM-A n. 6819, si è avvalso degli utili suggerimenti di Giuseppe Castagneri e della preziosa assistenza di Sheyla Miltello.

BIBLIOGRAFIA

- [1] Clementino D. Fisore L. (1993), "A man-machine dialogue system for speech access to train timetable information", *Proc. EUROSPEECH '93*, Berlin, Sept. 1993, Vol. 3, pp. 1863-1866.
- [2] Garbino E., Danielli M. (1993), "Managing Dialogue in a Continuous Speech Understanding System", *Proc. EUROSPEECH '93*, Berlin, Sept. 1993, Vol. 3, pp. 1661-1664.
- [3] Balasri M., Lazzarato S., Salza P.L., Sandri S. (1993), "The CSELT system for Italian text-to-speech synthesis", *Proc. EUROSPEECH '93*, Berlin, September 1993, Vol. 3, pp. 2091-2094.
- [4] Gelykens R., Swerts M. (1993), "Local and global prosodic cues to discourse organisation in dialogues", *Proc. ESCA Workshop on Prosody*, Lund, Sept. 1993, pp. 108-111.
- [5] Quazza S., Salza P.L., Spini A. (1993), "Prosodic Control in a Text-to-Speech System for Italian", *Proc. ESCA Workshop on Prosody*, Lund, Sept. 1993, pp. 78-81.
- [6] ESPRIT (1993), "Development of a context dependent methodology for Text-To-Speech Synthesis evaluation in interactive Dialogue systems", One Year Deliverable, *ESPRIT Project 6819 SAM-A*, Nov. 1993.
- [7] Pola L.C. (1992), "Quality Assessment of Text-to-Speech Synthesis by Rule", in: Furui S., Sondhi, M. M. (eds.), *Advances in Speech Signal Processing*, Marcel Dekker Inc., New York-Basel-Hong Kong, pp. 387-416.
- [8] CCITT (1993), Draft Recommendation P.8S of Study Group XII, January 1993, "Subjective performance assessment of the quality of speech of voice output devices", Special Rapporteur for Question 6/XII.
- [9] Welkowitz, R. Ewen and Cohen J. (1982), *Introductory Statistics for the Behavioral Sciences*, Harcourt Brace Jovanovich, San Diego.
- [10] Salza P.L., Foi E., Nebbia L. and Oraglia M. (1993), "MOS and Pair Comparison Combined Techniques for Quality Evaluation of Text-to-Speech Systems", in attesa di pubblicazione su *Acta Acustica*.

ESPRIT
W. and R. Jovanovich