LIAV (LESSICO DELL' ITALIANO AUDIOVISIVO) CORPUS LESSICALE AUDIOVISIVO PER L'ANALISI, LA SINTESI E IL RICONOSCIMENTO BIMODALI DELL'ITALIANO PARLATO

E. Magno Caldognetto P.Cosi

ISTC-SPFD CNR
Istituto di Scienze e Tecnologie della Cognizione - Sezione di Padova "Fonetica e Dialettologia",
Consiglio Nazionale delle Ricerche
www: http://nts.csrf.pd.cnr.it/
e-mail: {cosi,avesani}@csrf.pd.cnr.it

1. INTRODUZIONE

Poiché nella comunicazione orale faccia-a-faccia il segnale verbale viene trasmesso contemporaneamente ad altre informazioni su più canali, in particolare ad informazioni visive, tra le tecnologie del parlato si stanno sempre più imponendo all'attenzione dei ricercatori la **sintesi** e il **riconoscimento bimodali uditivo-visivi**.

Infatti, sebbene la modalità uditiva rappresenti il canale sensoriale più importante nel processo di percezione del parlato, è stato ormai accertato, sulla base delle ricerche sullo *speech-reading* o *lip-reading* [1-3], che l'informazione estratta dai movimenti articolatori visibili (movimento delle labbra, della mandibola, della lingua e visibilità dei denti), consente di migliorarne l'intelligibilità soprattutto quando il segnale acustico risulta degradato dalla presenza di rumore [4-6] oppure quando le capacità percettive dell'ascoltatore siano state danneggiate da patologie dell'udito.

Questo successo della percezione bimodale nell'identificazione delle unità fonologiche segmentali e quindi delle uscite lessicali si basa, intuitivamente, sul *sinergismo* tra le informazioni uditive e visive relative ad uno stesso fonema, in quanto prodotti sensoriali della realizzazione fonetica di uno stesso intento fonologico e collegati tra loro dalla relazione causale esistente tra movimenti articolatori e segnale acustico risultante.

Grazie a tale sinergismo risulta raccomandabile l'applicazione sistemi di sintesi e riconoscimento bimodali in vari tipi di interazione uomo-macchina per soggetti normali e patologici.

La sintesi audiovisiva può rendere infatti più naturale, robusto e amichevole l'accesso a tutti i tipi di banche-dati, dalla lettura di notiziari all'e-commerce, dalla didattica (insegnamento della lingua materna e di lingue straniere) alla pratica clinica e logopedica per la valutazione e riabilitazione dei soggetti ipoacusici e può essere utilizzata nella videotelefonia, nelle teleconferenze, nell'industria cinematografica e televisiva (cartoni animati, videogiochi, doppiaggio, "sottotitolatura" di messaggi in LIS, Lingua Italiana dei Segni, creazione di Attori Virtuali, ecc). Anche per i sistemi di riconoscimento automatico l'approccio bimodale fa prevedere notevoli vantaggi soprattutto qualora la decodificazione del parlato debba avvenire in ambiente rumoroso, come avviene in tutte le applicazioni *reali* del riconoscimento vocale, o in condizioni che prevedono la massima sicurezza.

2. OBIETTIVI DEL LIAV (LESSICO DELL'ITALIANO AUDIOVISIVO)

La condizione fondamentale per l'implementazione di sistemi di sintesi e riconoscimento bimodali è l'individuazione delle unità fonologiche segmentali e la definizione delle loro caratteristiche acustiche e visive. Se per le informazioni uditive l'unità minima è il *fonema*, per le unità dell'informazione visiva è stato coniato il termine di *visema* con cui si individua un gruppo di consonanti che, condividendo movimenti articolatori visibili simili, trasmettono una stessa informazione fonologica (per l'italiano [7-12]). I visemi non sono però in rapporto biunivoco con

i fonemi, e da ciò deriva l'importanza di una loro corretta caratterizzazione sia in termini articolatori che percettivi per generare Facce Parlanti sempre più naturali [13-15] e per migliorare le prestazioni dei sistemi di riconoscimento tradizionali basati esclusivamente sull'informazione ottenibile attraverso il solo canale acustico [16-19].

Per lo sviluppo di queste tecnologie e per un efficace confronto scientifico degli algoritmi ad esse associati (tecniche di codifica di immagine quali MPEG [15], algoritmi e sistemi di estrazione automatica dei movimenti articolatori direttamente dall'immagine visiva), è necessario disporre di corpora, di notevoli dimensioni in termini di contenuto e di quantità, di segnale vocale di cui si sia acquisito in sincronia il corrispondente segnale video (per una revisione cfr. Chibelushi et. al. [20]).

Visto il notevole sforzo scientifico ed economico necessario per la realizzazione di simili corpora, l'unione di più laboratori favorisce sicuramente un migliore sfruttamento delle risorse che altrimenti risultano parcellizzate, duplicate e spesso disorganizzate.

E' evidente inoltre, com'è avvenuto in nazioni tecnologicamente avanzate in questo campo, che alla disponibilità di questi corpora presso i vari laboratori ed enti di ricerca consegue sempre un notevole sviluppo di tutte le metodologie associate a queste problematiche di ricerca.

Per questo riteniamo necessario proporre anche per l'italiano l'organizzazione e l'acquisizione di un corpus simile, il LIAV-Lessico dell'Italiano AudioVisivo, con lo scopo specifico di implementare sistemi di sintesi e riconoscimento bimodali uditivo-visivi del parlato.

La specificità linguistica deve essere prima documentata e poi applicata, tanto a livello fonologico quanto lessicale, soprattutto per quanto riguarda le informazioni inviate lungo il canale visivo. Infatti, anche se confrontando i risultati delle ricerche sull'informazione fonologica trasmessa dai movimenti labiali nelle diverse lingue, si possono individuare dei parallelismi dovuti alla migliore visibilità dei *loci* articolatori anteriori rispetto a quelli posteriori, i risultati variano invece da lingua a lingua poiché è diverso l'inventario fonologico (per numero di vocali e consonanti, per la diversa utilizzazione dello spazio articolatorio, per le diverse regole fonotattiche) cui corrisponde un diverso inventario visemico.

Solo dopo l'acquisizione di tali conoscenze si potranno implementare regole di trascrizione grafema-fonema-visema e regole di coalescenza in uno stesso visema di vari fonemi, indispensabili per l'individuazione delle possibili competizioni lessicali da applicarsi tanto nella realizzazione di programmi di sintesi bimodale quanto nei sistemi di riconoscimento automatico bimodale.

Il risultato finale di un tale progetto avrà sicuramente ricadute scientifiche in vari settori multidisciplinari anche indipendenti fra loro quali quelli della fonetica, della fonologia e della psicologia della percezione oltre a quelli dell'informatica, dell'elaborazione del segnale vocale e delle immagini, della computer graphics, dell'interazione uomo-macchina, ecc.

3. DESCRIZIONE SINTETICA DEL CORPUS AUDIO-VIDEO E DELLE MODALITÀ DI ACQUISIZIONE

Il sistema da noi ipotizzato per la raccolta del corpus AV dovrebbe essere basato su una workstation in grado di acquisire mediante una telecamera di alta qualità fino a 25 frame (*interlacciati*) per secondo di immagini video RGB, ad una risoluzione di 768x576 pixel (PAL standard). In aggiunta, utilizzando 2 microfoni, caratterizzati da una differente qualità in termini di rapporto segnale/disturbo (*S/N*), saranno acquisiti due canali audio a 16 bit campionati a 16 kHz. Quindi per ogni singolo item (frasi o parole acquisite) verranno memorizzati un file video, 2 file audio e 2 file contenenti alcune informazioni di sincronizzazione audio e video.

I soggetti siederanno davanti alla video-camera, ai microfoni e ad un terminale video dove appariranno gli stimoli da produrre.

Particolare cura dovrà essere riposta nel determinare correttamente l'illuminazione dei soggetti per poter sfruttare appieno le tecniche automatiche di elaborazione d'immagine.

Per un sottoinsieme del corpus (set di parole isolate) verrà acquisito un altro canale in cui saranno memorizzati gli andamenti dei movimenti articolatori rilevati da markers opportunamente posizionati sulle labbra dei soggetti e registrati mediante un sistema optoelettronico (ELITE [55], [56]), già utilizzato presso la Sezione di Fonetica e Dialettologia di Padova dell'ISTC per lo studio analitico dei movimenti labiali nella produzione delle consonanti e delle vocali dell'italiano in sequenze VCV[9-12]. Alcuni soggetti produrranno inoltre un sottoinsieme degli stimoli dopo un opportuno mark-up delle labbra che dovrebbe favorire l'applicazione di software specifici progettati al fine di memorizzare automaticamente i movimenti labiali.

Il corpus AV dovrà essere diviso in almeno 5 sezioni (Tabella 1), corrispondenti a differenti e sempre più complessi compiti, relativamente ai quali dovranno essere via via progettati e implementati diversi sistemi di riconoscimento.

Tabella 1

	n. soggetti	compito	vocabolario	n. parole
1	50 (25m+25f)	set di parole isolate	(CVCV)	150
2	50 (25m+25f)	sequenze di lettere connesse	26	1500
3	50 (25m+25f)	sequenze di numeri connessi	10	1500x8
4	50 (25m+25f)	numeri telefonici e carte di credito		100x1
5	50 (25m+25f)	frasi		500x1

(N.B. Le quantificazioni sono indicative e dovranno essere definite, assieme al contenuto dettagliato delle varie sezioni, nella fase iniziale di preparazione del corpus)

Per quanto riguarda il primo compito, ogni soggetto pronuncerà un gruppo (1 su 5 gruppi) di 30 parole. L'inventario delle parole dovrà contenere quante più possibili coppie minime in termini sia di ***fonem****i* che di ***visem****i*, esemplificare i casi più importanti di corrispondenza tra fonemi e visemi e di coalescenza di più fonemi in un unico visema e permettere la valutazione degli effetti coarticolatori al fine di individuare le caratteristiche e i problemi di un corretto accesso al lessico dell'italiano per via unimodale visiva e bimodale uditivo-visiva.

Questo tipo di materiale dovrà essere registrato secondo tutte le modalità precedentemente indicate e costituirà un corpus di riferimento indispensabile tanto per gli esperimenti di riconoscimento automatico bimodale quanto per i programmi di sintesi audio-visiva del parlato.

L'attività proposta è articolata in tre anni, con sotto-obiettivi progressivi che possono essere così riassunti:

-I anno: il conseguimento di una preliminare definizione integrata e coerente dei requisiti e delle metodologie, cioè definizione di una bibliografia multidisciplinare di riferimento e redazione di una relazione sullo stato dell'arte; definizione e standardizzazione delle metodologie riguardo l'acquisizione del corpus audio-video;

-II anno: registrazione del corpus audio-video secondo le modalità specificate nel corso della prima fase;

-III anno: valutazione e analisi conclusiva del materiale registrato, produzione del prodotto finale sotto forma di Cdrom..

Oltre alla realizzazione del prodotto finale, cioè il corpus audio-video, un altro importante risultato del progetto sarà la messa a punto e la condivisione di strategie comuni da parte di più gruppi di ricercatori italiani per l'acquisizione e il riconoscimento del segnale audio-video e la standardizzazione di tutte le metodologie coinvolte.

E' infatti stato evidenziato, sulla base di imprese simili già realizzate in altre nazioni tecnologicamente avanzate, che alla disponibilità di questi corpora nei vari laboratori ed enti di

ricerca consegue sempre un notevole sviluppo di tutte le metodologie associate alle problematiche della sintesi e del riconoscimento bimodali.
Infine, visto il notevole sforzo scientifico ed economico necessario per la realizzazione di simili corpora, l'unione di più laboratori favorirà un migliore sfruttamento delle risorse che altrimenti potrebbero risultare parcellizzate, duplicate e disorganizzate.

5. UNITA' OPERATIVE

Le unità operative che intendono cooperare a questo progetto attualmente sono:

-ISTC-CNR
Istituto di Scienze e Tecnologie della Cognizione - Sezione di Fonetica e Dialettologia
Consiglio Nazionale delle Ricerche (coordinatori: Emanuela Magno Caldognetto, Piero Cosi) WWW: http://nts.csrf.pd.cnr.it/ o http://www.csrf.pd.cnr.it/

-ITC-IRST
Istituto Trentino di Cultura - Istituto per la Ricerca Scientifica e Tecnologica
www: http://www.itc.it/IRST/index.htm (responsabile: Gianni Lazzari)

-DEI
Dipartimento di Elettronica e Informatica - Università degli Studi di Padova
www: http://www.dei.unipd.it (responsabile: Antonio Mian)

-DIST
Dipartimento di Informatica, Sistemistica e Telematica- Università degli Studi di Genova
www: http://www.dist.unige.it (responsabile: Fabio Lavagetto)

DIS
Dipartimento di Informatica e Sistemistica - Università di Roma "La Sapienza"
www: http://www.dis.uniroma1.it/ (responsabile: Catherine Pelachaud)

6. CONSIDERAZIONI FINALI SU APPLICAZIONI E POSSIBILITÀ DI SFRUTTAMENTO

Nella comunità scientifica internazionale vi è un notevole interesse verso le nuove tecnologie multimediali e multimodali poiché è unanimemente riconosciuto che una loro introduzione nelle normali applicazioni di sintesi e di riconoscimento automatici del parlato potrebbe portare ad un'efficace robusta e sempre più *human-like* interazione uomo-macchina.

La raccolta e l'utilizzazione del corpus proposto costituiranno il primo, indispensabile passo nell' elaborazione di nuovi sistemi per il tracking automatico della posizione della faccia del soggetto e dei movimenti labiali e mandibolari, dati che potranno successivamente essere usati come parametri ausiliari a quelli acustici per un riconoscimento bimodale e come informazioni indispensabili alla messa a punto di Facce Parlanti.

BIBLIOGRAFIA

[1] Summerfield Q.,1987, Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception, in Dodd B. and Campbell R.(Eds.), Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 3-51.

[2] Massaro D.W., 1987, Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry, in Dodd B. and Campbell R.(Eds), Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 53-83.

[3] Dodd B. and Campbell R., (Eds), 1987, Hearing by Eye: The Psychology of Lip-Reading, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[4] Erber N.P.,1975, Auditory-Visual Perception of Speech, Journal of Speech and Hearing Disorders, 40, 481-492.

[5] MacLeod A. and Summerfield Q., 1987, Quantifying the Contribution of Vision to Speech Perception in Noise, British Journal of Audiology, 22, 131-141.

[6] Mohamadi T. et Benoit C., 1992, Apport de la vision du locuteur à l'intelligibilité de la parole bruite en français , Bulletin de la Communication Parlée, n. 2, 32-41.

[7] Magno Caldognetto E. e Vagges K, 1990a, Il riconoscimento visivo dei movimenti articolatori da parte di soggetti normali e ipoacusici. In Scritti in onore di Lucio Croatto, Padova, 153-166.

[8] Magno Caldognetto E. e Vagges K., 1990b, Il riconoscimento delle consonanti in un test di lettura labiale, Atti del Congresso Nazionale della Società Italiana di Acustica, l'Aquila, 94-99.

[9] Magno Caldognetto E., Vagges K. and Zmarich C., 1995, Visible Articulatory Characteristics of the Italian Stressed and Unstressed Vowels, Proc. of ICPhS 95, Stochkolm, 14-19 August 1995, Vol. 1, 366-369.

[10] Cosi P. and Magno Caldognetto E., 1996, Lip and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications, in D.G. Stork and M.E. Henneke (Eds.), 'Speechreading by Humans and Machine: Models, Systems and Applications', NATO ASI Series, Series F: Computer and Systems Sciences, Vol. 150, Springer-Verlag, 291-313.

[11] Magno Caldognetto E., Zmarich C., Cosi P. and Ferrero F., 1997, Italian Consonantal Visemes: Relationships Between Spatial and Temporal Articulatory Characteristics and Coproduced Acoustic Signal, Proceedings of AVSP-97, Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational and Cognitive Science Approaches, Rhodes (Greece), 26-27 September 1997, 5-8.

[12] Magno Caldognetto E., Zmarich C. and Cosi P., 1998, Statistical Definition of Visual Information for Italian Vowels and Consonants, in D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (eds.), Proceedings AVSP '98, 4-6 December 1998, Terrigal (AUS), 1998, pp. 135-140.

[13] Brooke N.M. and Petajan E.D., 1986, Seeing Speech: Investigations into the Synthesis and Recognition of Visible Speech Movements Using Automatic Image Processing and Computer Graphics, Proceedings of the International Conference on Speech Input and Output: Techniques and Applications, 24-26.

[14] Benoit C., Lallouache T., Mohamadi T., and Abry C., 1992, A Set of French Visemes for Visual Speech Synthesis, in Bailly G., Benoit C., and Sawallis T.R. (Eds.), Talking machines: Theories, Models, and Designs, North-Holland, Amsterdam, 485-504.

[15] Lavagetto F.and Lavagetto P., 1995, A New Algorithm for Visual Synthesis of Speech, 4th European Conference on Speech Communication and Technology, Madrid, 18-21 settembre 1995, Vol. 1, 303-306.

[16] Garcia, O., Goldschen, A.J., and Petajan, E.D., 1992, Feature Extraction for Optical Automatic Speech Recognition or Automatic Lipreading, George Washington University: IIST-92-32, November, 1992.

[17] Bregler C. and Konig Y., Eigenlips for Robust Speech Recognition,1994, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 669– 672.

[18] Brooke N.M., Tomlinson M.J. and Moore R.K., 1994, Automatic Speech Recognition That Includes Visual Speech Cues, Proceedings of the Institute of Acoustics-1994 Autumn Conference (Speech and Hearing), Vol. 16, Part 5, 15 – 22.

[19] Cosi P., Dugatto M., Ferrero F.E., Magno Caldognetto E. and Vagges K., 1996, Phonetic Recognition by Recurrent Neural Networks Working on Audio and Visual Information, Speech Communication, North Holland, Vol. 19, No. 3, 245-252.

[20] Chibelushi C.C., Deravi F. and Mason, J.S.D., 1996, Survey of Audio Visual Speech Databases, Internal Report - Speech and Image Processing Research Group, Dept. of Electrical and Electronic Engineering, University of Wales Swansea, [URL: http:// faith.swan.ac.uk/SIPL/david/survey.html].