# ALCUNE CONSIDERAZIONI SULL'IMPORTANZA DEGLI ASPETTI DINAMICI NELLA PERCEZIONE, PRODUZIONE ED ELABORAZIONE DEL PARLATO

Piero Cosi
ISTC-spfd CNR
Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche - via Martiri della Libertà, 2 - 35127 Padova, Italia e
*e-mail: piero.cosi@pd.istc.cnr.it*

## 1. SOMMARIO

In questo lavoro vengono sinteticamente illustrati alcuni dei più significativi apporti tecnologici che nel corso degli ultimi anni sono stati influenzati dalla dimensione temporale del parlato nel campo dell'analisi del segnale vocale, della sintesi della voce da testo scritto e del riconoscimento automatico del segnale verbale. Per quanto riguarda la realizzazione di facce parlanti animate, sono discussi poi alcuni esempi dell'influenza degli aspetti dinamici nella percezione e nella'interpretazione delle espressioni facciali e più in generale degli intenti comunicativi, nella trasmissione di emozioni, stati d'animo e atteggiamenti, nell'interazione faccia a faccia.

## 2. INTRODUZIONE

La dimensione temporale è un elemento costitutivo non solo dei meccanismi di produzione del parlato, intervenendo, a livello segmentale, nella determinazione delle durate e nella pianificazione e nel controllo di tutti i gesti articolatori e, a livello soprasegmentale, nell'allineamento dei contorni intonativi con le parti dell'enunciato, ma anche, nella percezione del segnale verbale e, più in generale, nell'interpretazione di un qualsiasi atto comunicativo. Ad esempio, sia la configurazione delle caratteristiche facciali che la sincronizzazione delle azioni facciali sono importanti nell'espressione e nel riconoscimento delle emozioni (Cohn, 2007).
In questa breve rassegna si fa ampio riferimento a due lavori precedentemente presentati. In particolare per quanto riguarda il parlato si fa riferimento a "*50 Years of Progress in Speech and Speaker Recognition Research*", presentato da Sadaoki Furui nel 2005 e pubblicato in *ECTI Transactions on Computer and Information Technology* (Furui S., 2005) e, per quanto riguarda la percezione delle espressioni facciali e piu in generale degli intenti comunicativi, si fa riferimento a "*Foundations of human-centered computing: Facial expression and emotion*", presentato da Jordan F. Cohn nel 2007 e pubblicato in Proceedings of the InternationalJoint Conference on Artificial Intelligence (Cohn J.F, 2007).

## 3. DIMENSIONE TEMPORALE E *SPEECH TECHNOLOGY*

Per quanto riguarda il Trattamento Automatico del Linguaggio (*TAL*), e in particolare, l'analisi del segnale vocale, la sintesi della voce da testo scritto e il riconoscimento automatico del segnale verbale, gli apporti tecnologici più significativi che nel corso degli ultimi anni sono stati influenzati dalla dimensione temporale del parlato sono illustrati

schematicamente in Figura 1, dove sono evidenziate le caratteristiche specifiche che nel corso degli anni hanno reso sempre più affidabili queste tecnologie.

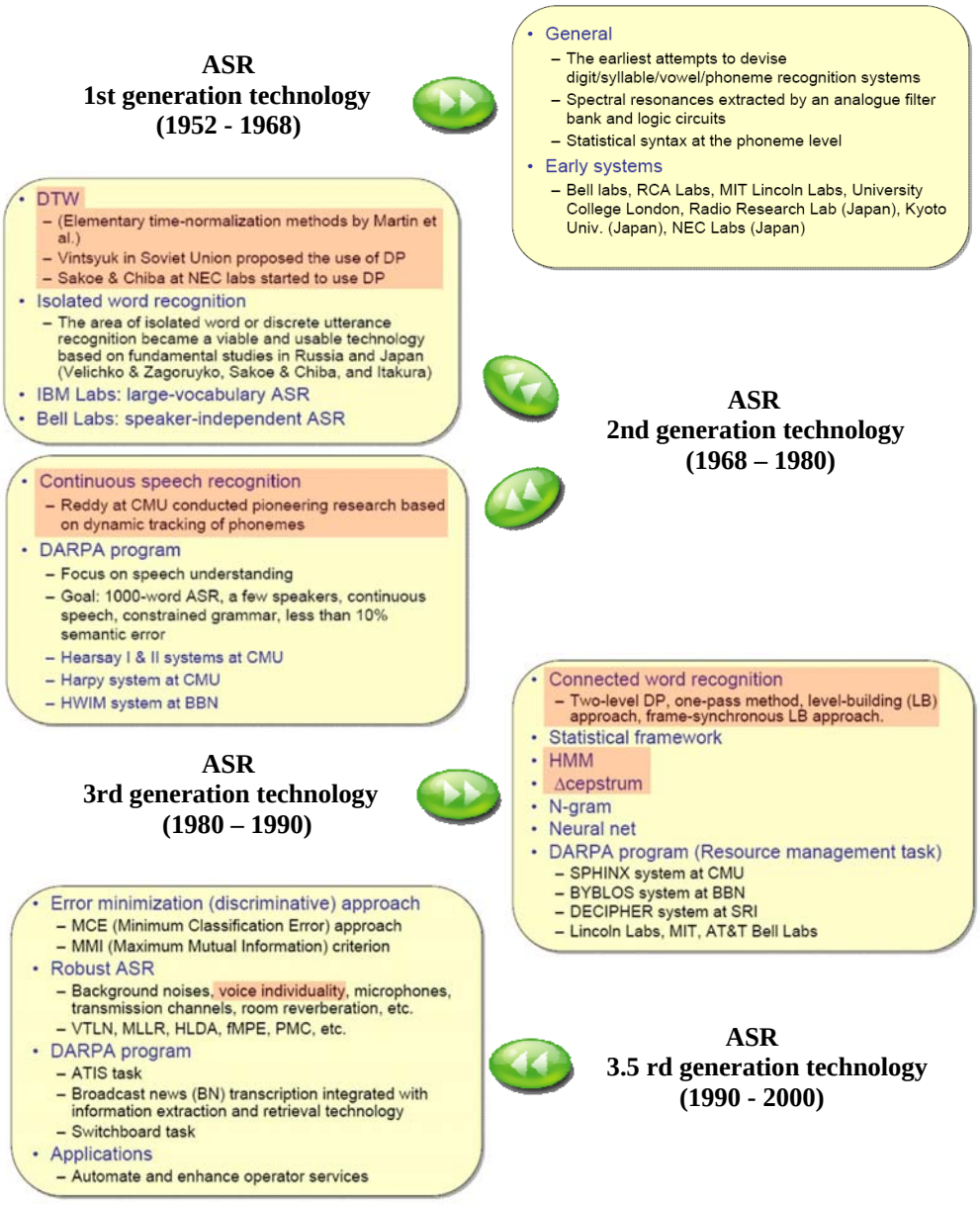


Figura 1. Apporti tecnologici più significativi influenzati dalla dimensione temporale del parlato nel campo del riconoscimento automatico: sono evidenziate le caratteristiche specifiche che nel corso degli anni hanno reso sempre più affidabili queste tecnologie.

Nonostante enormi progressi, di seguito illustrati in Figura 2, manca ancora molto però all'utilizzazione diffusa di queste tecnologie soprattutto a causa della loro inaffidabilità

quando vengono utilizzate realmente "*sul campo*", quando cioè si devono superare i problemi relativi ad esempio al riconoscimento automatico del parlato in situazioni rumorose (cocktail party, rumori sovrapposti, rumori di canale…) oppure al riconoscimento automatico di parlato spontaneo.

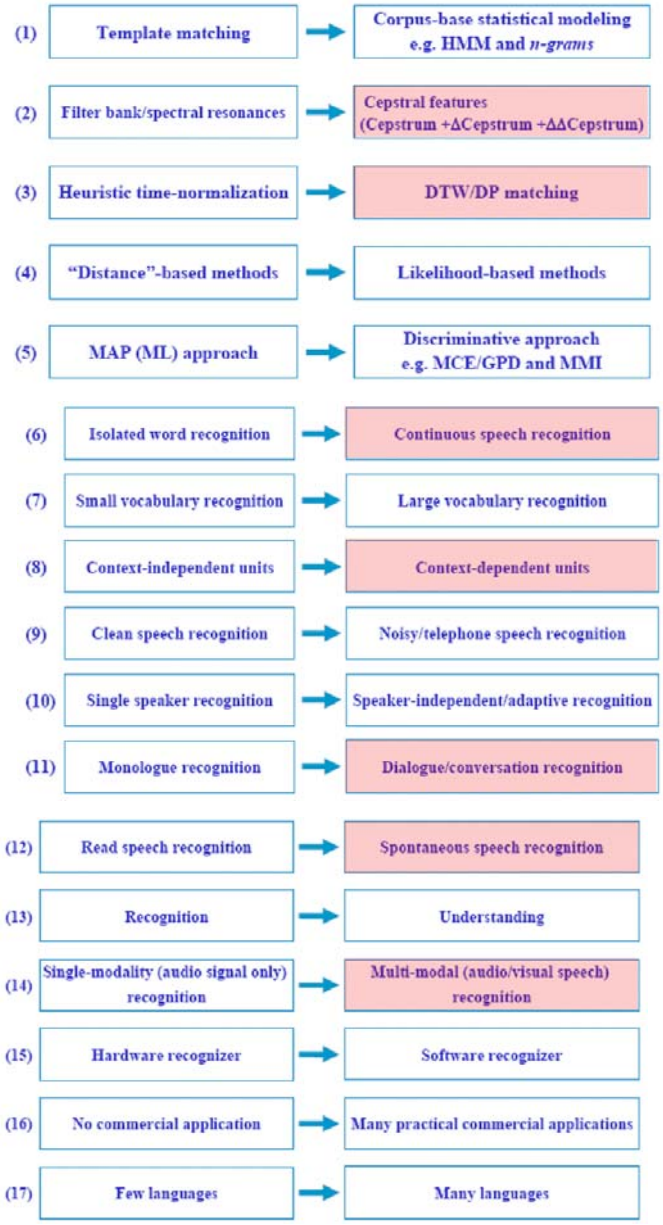


Figura 2. Principali innovazioni tecnologiche in cui la dimensione temporale gioca un ruolo fondamentale.

A titolo di esempio si sottolineano gli enormi miglioramenti dovuti al passaggio fra l'utilizzazione di caratteristiche basaste sull'analisi effettuata da banchi di filtri in frequenza e l'utilizzazione di caratteristiche basate sull'analisi del Cepstrum e delle relative velocità ed accelerazioni ($\Delta$, $\Delta\Delta$), oppure fra l'utilizzazione della normalizzazione temporale euristica utilizzata agli inizi degli anni 70 per uniformare i confronti fra il parlato target e i modelli di parola memorizzati e l'utilizzazione della tecnica di programmazione dinamica (*Dynamic Time Warping*), oppure l'introduzione della modellizzazione basata sulla teoria delle catene di *Markov* nascoste.

Riassumendo, in Figura 3, è graficamente illustrato l'avvicendarsi temporale delle varie generazioni dei sistemi di riconoscimento., dalla preistoria (1920) agli anni recenti (3.5G). L'interrogativo fondamentale dei prossimi anni, per risolvere i problemi rimasti per un'utilizzazione diffusa di queste tecnologie è:

QUALI SARANNO LE CARATTERISTICHE DELLA QUARTA GENERAZIONE DEI SISTEMI ASR?

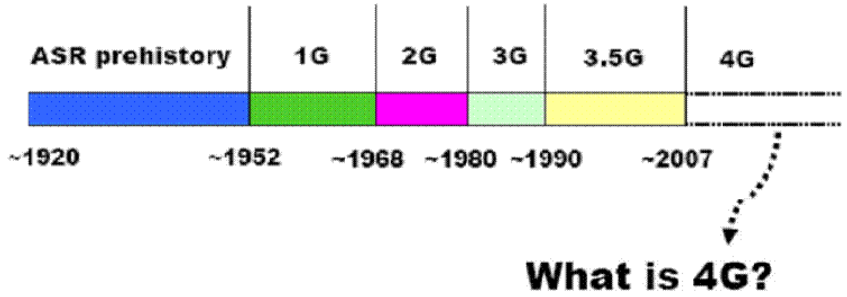


Figura 3. ASR dalla preistoria ai giorni nostri,.dalla prima generazione alla terza, alla terza e mezzo ed alla futura quarta.

Le maggiori difficoltà che devono a tutt'oggi essere ancora risolte sono illustrate in Figura 4. In particolare le più rilevanti sono quelle relative alla variabile velocità di eloquio, all'estrema variabilità dell'accento, dello stile ed in generale della prosodia, tutte caratteristiche in cui la dimensione temporale risulta di fondamentale importanza.

In conclusione, per quanto riguarda il TAL, negli ultimi 50 anni sono stati fatti passi giganteschi e le maggiori innovazioni tecnologiche sono state focalizzate al miglioramento dei sistemi di riconoscimento soprattutto in termini di aumento della loro robustezza. Tuttavia il 60% (16/28) dei "problemi irrisolti" elencati da Beek et al. nel 1977 non sono ancora stati risolti.

Una comprensione assai più dettagliata del processo di produzione e percezione del parlato sarà necessariamente richiesta in futuro prima che i sistemi di riconoscimento vocale automatico possano avvicinarsi alla prestazione umana e di sicuro gran parte degli avanzamenti significativi in questo campo verranno dalla estesa collaborazione fra questa necessaria conoscenza e l'elaborazione della conoscenza basata invece sulle architetture e sulle teorie del riconoscimento di pattern basati sulla statistica.

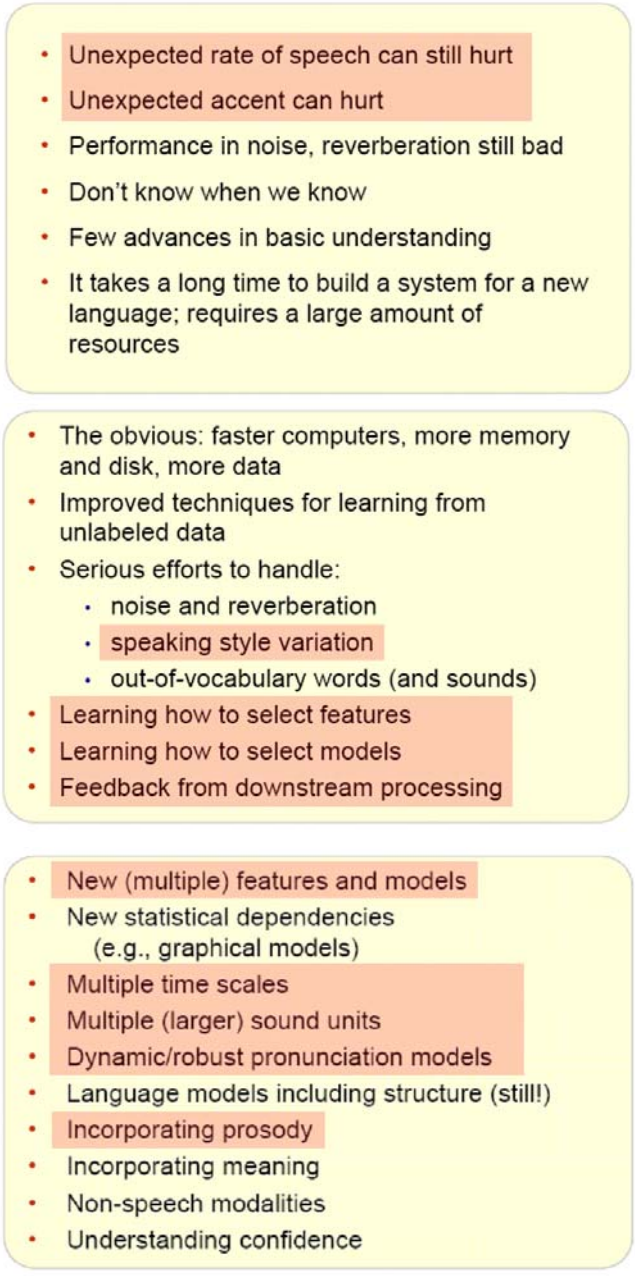


Figura 4. Elenco delle maggiori difficoltà che a tutt'oggi devono essere ancora risolte per una diffusione completa delle tecnologie TAL ed in particolare del riconoscimento automatico del parlato.

## 4. DIMENSIONE TEMPORALE E ESPRESSIONI FACCIALI

Già nel 1921, Flach sosteneva che solo la dinamica di un movimento è non-ambigua e convincente ("*only the dynamics of a movement is unambiguous and convincing*"). La configurazione delle azioni facciali (espressioni relative sia a specifiche emozioni sia ad unità di azione individuali) rispetto alle emozioni ed all'intenzione comunicativa è un importante tema di ricerca. Meno invece si conosce circa la sincronizzazione delle azioni facciali, anche perché la misurazione manuale della sincronizzazione è assai complicata e laboriosa. Tuttavia, sappiamo che (Cohn, 2007) noi siamo altamente sensibili alla sincronizzazione delle azioni facciali nelle interazioni sociali (Edwards, 1998). Le azioni facciali più lente (vedi Figura 5), ad esempio, sembrano essere più genuine e naturali (Krumhuber & Kappas, 2005), come pure lo sembrano essere quelle più sincrone nei loro movimenti (Frank & Ekman, 1997). In particolare, le espressioni facciali più sottili diventano visibili soltanto quando le informazioni di movimento sono a disposizione di chi le percepisce (Ambadar, Schooler, & Cohn, 2005).

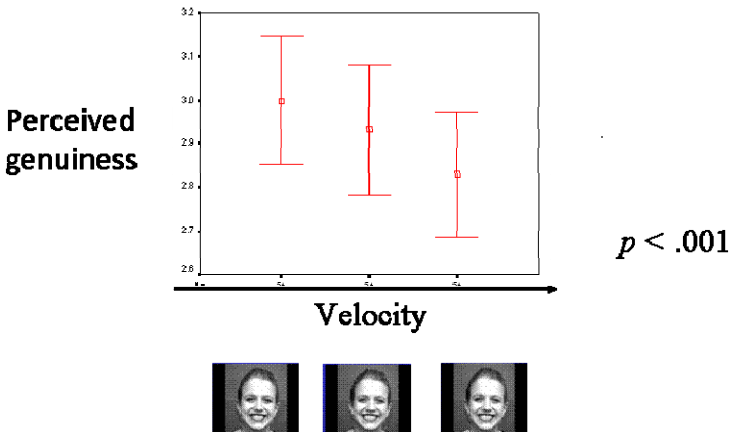


Figura 5. Risultati sperimentali a sostegno dell'ipotesi che la genuinità di un espressione (il sorriso in questo caso) è fortemente correlata alla lentezza della sua realizzazione articolatoria.

La dinamica è cioè particolarmente importante per inferire l'intenzione comunicativa. Alcuni studi condotti dal gruppo di ricerca di CMU utilizzando tecniche automatiche di analisi di immagini facciali per misurare la sincronizzazione delle azioni facciali, hanno provato (vedi Figura 5) che le caratteristiche dinamiche riescono a discriminare fra i sorrisi intenzionali e quelli spontanei con un livello di precisione dell' 89% (Cohn & Schmidt, 2004). Usando caratteristiche simili, il divertimento, l'imbarazzo ed il sorriso "gentile" sono stati discriminati con una precisione dell' 83% (Kanade, Hu, & Cohn, 2005), che è paragonabile a quella umana. Lavori più recenti suggeriscono inoltre che la coordinazione multimodale dell'espressione facciale, del movimento della testa e dei gesti sono caratteristiche specifiche dell'imbarazzo (Keltner, 1995).

## 5. OSSERVAZIONI CONCLUSIVE

L'unica osservazione che può essere fatta sulla base di queste brevi note è che senza una completa conoscenza della dimensione temporale dei meccanismi di produzione del parlato

e, più in generale, nell'interpretazione di un qualsiasi atto comunicativo, una diffusione capillare e completa delle tecnologie TAL ed in particolare del riconoscimento automatico del parlato non potrà mai avvenire in maniera soddisfacente.

## 6. BIBLIOGRAFIA


Ambadar, Z., Schooler, J., & Cohn, J. F. (2005). Deciphering the enigmatic face: The impor-tance of facial dynamics to interpreting subtle facial ex-pressions. *Psychological Science*, 16, 403-410.

Beek, B. (1977). An assessment of the technology of automatic speech recognition for military applications, *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-25, 1977, 310-322.

Cohn, J.F. (2007). Foundations of human-centered computing: Facial expression and emotion. *Proceedings of the InternationalJoint Conference on Artificial Intelligence* (IJCAI'07), Hyderabad, India, 5-12.

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and sponta-neous smiles. *International Journal of Wavelets, Mul-tiresolution and Information Processing*, 2, 1-12.

Edwards, K. (1998). The face of time: Temporal cues in facial expressions of emotion. *Psychological Science*, 9(4), 270-276.

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stakes lies. *Journal of Personality and Social Psychology*, 72(6), 1429-1439.

Furui S. (2005), 50 Years of Progress in Speech and Speaker Recognition Research, *ECTI Transactions on Computer and Information Technology*, Vol.1, no.2 November 2005, 64-74.

Kanade, T., Hu, C., & Cohn, J. F. (2005). Facial expression analysis. Paper presented at the *IEEE International Workshop on Modeling and Analysis of Faces and Gestures*, Beijing, China.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement and shame. *Journal of Personality and Social Psychology*, 68(3), 441-454.

Krumhuber, E., & Kappas, A. (2005). Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior*, 29, 3-24.