HLT AND REAL TIME MULTILINGUAL SUBTITLING: SOME STATE OF THE ART IN ITALY AND EXPERIENCE AT THE LAST INTERSTENO SPEECH RECOGNITION CHAMPIONSHIPS

Carlo Aliprandi[1] and Fabrizio G. Verruso[2]

[1] Synthema Srl – Pisa
carlo.aliprandi@synthema.it

[2] Assemblea Regionale Siciliana
fverruso@ars.sicilia.it

Key words

Live Subtitling, Human Language Technologies (HLT), Multilingual Speech Recognition, Assistive Technology, Natural Language Understanding (NLU)

Sottotitolazione diretta, Tecnologie del Linguaggio Naturale, Riconoscimento Vocale Multilingue, Tecnologia Assistiva, Comprensione del Linguaggio Naturale

ABSTRACT

L'articolo presenta lo stato dell'arte e alcuni recenti progressi scientifici ottenuti nel campo delle Tecnologie del Linguaggio Naturale (HLT) in relazione a Riconoscimento Vocale, Comprensione del Linguaggio e Correzione Automatica dei testi.
Si descrive Voice Subtitle, il sistema multilingue per la generazione automatica dei sottotitoli dedicato agli Operatori Televisivi. L'articolo porta inoltre il contributo scientifico e applicativo di un'esperienza professionale significativa, realizzata in occasione degli ultimi campionati mondiali di riconoscimento vocale organizzati da Intersteno, dove nuovi importanti risultati e traguardi sono stati raggiunti.

1. Introduzione

Focalizzandosi sulle recenti esperienze di ricerca, presentiamo Synthema Voice Subtitle, il sistema per la generazione automatica dei sottotitoli dedicato ai Content Provider, agli Operatori Televisivi e ai Broadcaster.
Voice Subtitle, interfacciandosi direttamente con i sistemi Teletext, permette di ridurre drasticamente i tempi e i costi di produzione e rilascio dei sottotitoli, aumentando la produttività dei sottotitolatori. Un beneficio secondario importante è la drastica riduzione dei tempi di istruzione e formazione dei sottotitolatori, che possono essere operativi molto velocemente. Presenteremo le caratteristiche dell'utilizzo di Voice Subtitle per programmi televisivi in diretta e in semidiretta (ossia programmi chiusi e rilasciati con poco anticipo sulla loro messa in onda), quali telegiornali, talk-show, dibattiti e trasmissioni sportive. Mediante un modello innovativo 'a due fasi' e mediante l'introduzione di un eventuale operatore intermedio, si possono realizzare, revisionare e trasmettere sottotitoli in tempo reale.

Viene poi introdotto Synthema Voice Suite, il sistema professionale per la resocontazione che ha consentito a Fabrizio Verruso di riconfermarsi primatista mondiale di riconoscimento vocale, stabilendo il nuovo record di velocità di dettatura (174 parole al minuto) agli ultimi campionati mondiali Intersteno, l'associazione internazionale dei professionisti e insegnanti di scritture veloci.
In tale competizione, potenziando le prestazioni del software, siamo riusciti a dimostrare come, nel panorama delle metodologie di resocontazione, la ultima nata tecnica di riconoscimento vocale s'imponga ormai con pari dignità rispetto alle ben più note stenografia e stenotipia.

2. La Sottotitolazione

La sottotitolazione è diventata un'attività fondamentale per gli operatori televisivi, anche in ragione del fatto che la legislazione nazionale ed europea richiede che sempre più programmi televisivi vengano dotati di sottotitoli per le persone disabili dell'udito. In questo senso la sottotitolazione può essere considerata a pieno titolo come una Tecnologia Assistiva, oltre che un mezzo per favorire la Digital Inclusion.

La produzione di sottotitoli può essere descritta nello stesso modo in cui vengono preparati i programmi televisivi. La trasmissione di un programma può essere effettuata durante la ripresa. In questo caso si parla di trasmissioni 'live' o in diretta. La trasmissione di un programma può essere effettuata dopo la ripresa, tramite un nastro o uno stream video registrato. Si parla in questo caso di trasmissioni 'offline' o preregistrate.
I sottotitoli possono essere preparati prima della messa in onda della trasmissione: si parla in questo caso di sottotitolazione in differita (offline subtitling). Quando i sottotitoli non sono disponibili al momento della messa in onda di una trasmissione, si parla di sottotitolazione in diretta (real-time o live subtitling).

Se un programma televisivo viene chiuso, nella sua parte video e audio, con un margine di anticipo di tempo sufficiente, un operatore può produrre i sottotitoli, sincronizzarli con il video e preparare il programma per la sua messa in onda (video e sottotitoli) in automatico. Tuttavia, un caso particolare e 'ibrido' di sottotitolazione è quello dei programmi che vengono chiusi dalla redazione in tempi molto vicini alla loro messa in onda (ad esempio, alcuni servizi dei telegiornali vengono chiusi 30-60 minuti prima dell'inizio del Tg). Questi programmi vengono trasmessi in una modalità ibrida tra la modalità diretta e la modalità differita, che viene definita modalità 'semidiretta'.
Per tali programmi la produzione dei sottotitoli può essere effettuata live oppure offline: nel primo caso, come precedentemente definito, si parla di sottotitolazione in diretta, mentre nel secondo caso si parla di sottotitolazione 'semidiretta'. In pratica, nella sottotitolazione in semidiretta i sottotitoli vengono prodotti prima della messa in onda del programma e un operatore procede alla sincronizzazioni manuale tra sottotitoli e video direttamente durante la messa in onda del programma. In tale caso il compito dell'operatore è quello di effettuare la sincronizzazione temporale tra sottotitoli e video, operazione non effettuata in precedenza.
Nel presente articolo parleremo delle diverse modalità di produzione dei sottotitoli e mostreremo come il sistema Voice Subtitle sia adatto per tutte le modalità definite, illustrando le caratteristiche d'uso in particolare per la sottotitolazione in diretta e in semidiretta.

3. Il sistema Voice Subtitle

Volendo realizzare un sistema specifico per la sottotitolazione abbiamo preferito studiare e sviluppare una nuova soluzione piuttosto che utilizzare o integrare un prodotto esistente sul mercato.
Con questo obiettivo abbiamo progettato Synthema Voice Subtitle, un sistema innovativo per la generazione automatica dei sottotitoli dedicato ai Content Provider, agli Operatori Televisivi e ai Broadcaster.
Volendo inoltre applicare il riconoscimento vocale multilingue alla produzione di sottotitoli abbiamo progettato un sistema flessibile e facilmente integrabile con le realtà sia operative che tecniche esistenti nelle aziende di produzione televisiva.
Nella progettazione del sistema è stata fondamentale l'esperienza precedentemente acquisita in un progetto di Data Broadcasting specifico per un servizio Teletext e la diretta collaborazione alla progettazione, sperimentazione e messa in opera della redazione del servizio Teletext del principale operatore televisivo italiano, RAI Radiotelevisione Italiana.
Abbiamo realizzato Voice Subtitle dedicando particolare attenzione al design dell'interfaccia e alla sua usabilità, in relazione alla complessità della 'catena di produzione' del sottotitolatore, che deve:

1. guardare il video della trasmissione
2. ascoltare l'audio della trasmissione
3. capire il contenuto dell'audio
4. sintetizzare il contenuto dell'audio
5. parlare in un microfono
6. verificare il sottotitolo prodotto
7. correggere il sottotitolo prodotto
8. inviare i sottotitoli in onda

Si capisce come un solo operatore/sottotitolatore difficilmente possa fare fronte a tutte queste operazioni, sebbene alcune siano opzionali o automatizzate dal sistema Voice Subtitle.

Per questo abbiamo introdotto un livello intermedio tra lo speaker ed il sistema di sottotitolazione. Tale livello è rappresentato da un operatore che può revisionare, correggere e confermare il sottotitolo prima della sua messa in onda. L'introduzione del revisore permette allo speaker di concentrarsi sulle sole fasi di produzione del testo del sottotitolo: alcune operazioni di postproduzione, quali la correzione automatica, la formattazione del testo in sottotitolo sono automatizzate da Voice Subtitle. Le operazioni di controllo e revisione finale vengono poi demandate al revisore che, lavorando fianco a fianco con lo speaker, può operare sui sottotitoli e provvedere alla messa in onda.

La Figura 1 schematizza il processo di sottotitolazione:

- Lo speaker riceve in audio (cuffia) e video (monitor TV) la trasmissione
- Lo speaker effettua la sintesi e lo rispeakeraggio della trasmissione tramite microfono
- Il Modulo ASR (Automatic Speech Recognition - riconoscimento vocale) trascrive lo speech in testo
- Il testo viene passato sottotitolo per sottotitolo al Modulo Linguistico
- Il sottomodulo Analisi effettua l'analisi morfologica e sintattica del sottotitolo. Il sottomodulo Correzione, attraverso le indicazioni ricevute dal modulo gestore della Base Dati Linguistica (Lexical Manager), effettua le correzioni automatiche
- Il sottotitolo viene passato al Modulo di Controllo
- Il revisore verifica il sottotitolo e ne conferma la messa in onda
- Il sottotitolo viene infine formattato secondo il formato Teletext e passato all'Interfaccia Teletext per la sua messa in onda

Il principale beneficio di questa architettura basata su due livelli, oltre alla suddivisione del lavoro tra speaker e revisore, consiste nell'introduzione di una fase di verifica e post-produzione del sottotitolo. Questo ovviamente introduce un ritardo rispetto alla messa in onda del sottotitolo dell'ordine di 10-12 secondi. Come vedremo in dettaglio in seguito, tale ritardo può essere compensato da una desincronizzazione tra audio e video, particolarmente utile per la modalità di sottotitolazione in semidiretta.

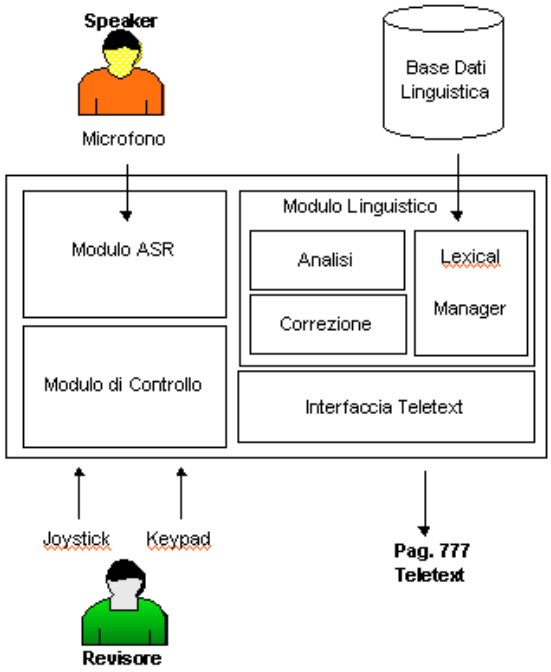

Figura 1: Architettura di Voice Subtitle

Interfacciandosi direttamente con i sistemi Teletext, Voice Subtitle permette di ridurre i tempi e i costi di produzione e rilascio dei sottotitoli. Un beneficio secondario importante è la forte riduzione dei tempi di istruzione e formazione dei sottotitolatori, che possono essere operativi molto velocemente.
Presenteremo in dettaglio nel paragrafo 5 le caratteristiche dell'utilizzo di Voice Subtitle per programmi televisivi in diretta e in semidiretta.

4. Tecnologie HLT

Le Tecnologie del Linguaggio Naturale (Human Language Technologies – HLT) hanno ormai raggiunto un grado di maturità tale che il loro impiego è quasi diventato quotidiano e naturale per molti utenti in svariate applicazioni.
Voice Subtitle utilizza tecnologie del linguaggio per il riconoscimento vocale, la comprensione del testo e la correzione automatica degli errori. Nel presente paragrafo illustriamo come tali tecnologie sono state utilizzate e qual benefici apportano nel processo di produzione di sottotitoli.

4.1 Riconoscimento Vocale

La tecnologia del riconoscimento vocale ha ormai raggiunto una maturità tale da trovare ampia applicazione in una vasta gamma di attività e

applicazioni pratiche. Tuttavia per poter produttivamente utilizzare tale tecnologia in applicazioni realtime, quali la sottotitolazione, abbiamo dovuto analizzare e risolvere alcuni problemi che, di per se, la pura tecnologia non risolveva.
Il riconoscimento vocale raggiunge livelli di qualità elevata per dizionari di base pressoché illimitati e per le principali lingue (inglese, italiano, francese, spagnolo, tedesco). La qualità dipende fortemente dal dominio di riferimento e dal livello di copertura del dizionario di base rispetto a tale dominio. Tipicamente la copertura si misura in percentuale di parole del dominio che già appartengono al dizionario di base. Le parole del dominio che non appartengono al dizionario di base costituiscono l'insieme delle cosiddette parole sconosciute (Out of Vocabulary Words - OVW). E' chiaro che, in un sistema ASR, la presenza di molte OVW implica un'alta percentuale di errori di trascrizione, in quanto il sistema ASR riconosce tutte e sole le parole presenti nei dizionari, riconducendo una OVW alla parola del dizionario acusticamente più simile.
Il dizionario di base del sistema contiene circa 100.000 parole estratte da un corpus generalista di notizie giornalistiche, documenti e lettere di vario genere. Ad esso è associato un modello del linguaggio addestrato sul corpus generalista.

Da alcuni test preliminari abbiamo verificato che, nel dominio giornalistico di riferimento, si ha una buona copertura di partenza. Tuttavia l'uso frequente di neologismi e di nuove parole, quali nomi propri, nomi di persona, nomi di organizzazioni, porta ad un peggioramento della qualità di riconoscimento. Abbiamo pertanto integrato nel sistema un meccanismo di aggiornamento e di aggiunta 'a caldo' di dizionari specifici, tecnicamente detti 'topic'.

Abbiamo costruito due topic, uno per il dominio *News* e uno per il dominio *Calcio*. Questi topic sono stati creati a partire da corpora di testi che riflettono il lessico e il modello del linguaggio di ciascun dominio. Un topic definisce quindi un dizionario specifico, in pratica permette di aggiungere al dizionario di base le nuove parole contenute nel corpus analizzato. L'uso dei topic ha consentito di minimizzare le OVW e di standardizzare per tutti gli speaker la qualità del riconoscimento, migliorando il tasso di riconoscimento del sistema fin da principio.
Oltre alle nuove parole, un topic integra nel sistema di riconoscimento le relative trascrizioni fonetiche e il modello del linguaggio. Questo permette, oltre ad aggiungere le OVW per tutti gli speaker senza che debba essere inserita –tipicamente dettandola- la loro trascrizione fonetica, di adattare e migliorare la qualità del riconoscimento. Il modello del linguaggio, infatti, permette di risolvere eventuali ambiguità distinguendo sempre le forme corrette in base al topic in uso.
Inoltre i topic, a differenza dei dizionari di base, possono essere attivati anche contemporaneamente e sono utilizzabili da subito da ogni speaker.
I test che abbiamo effettuato prima e dopo la creazione e integrazione dei topic, hanno mostrato un innalzamento della copertura dal 90%-92% al 97%-98% e un miglioramento della qualità di riconoscimento vocale di circa il 5%.

4.2 Natural Language Understanding

Un altro contributo significativo di tecnologie HLT in Voice Subtitle è rappresentato dalle funzioni di comprensione automatica del linguaggio (Natural Language Understanding – NLU).
Abbiamo implementato un meccanismo di interpretazione automatica del parlato in base a strutture predefinite. Tali strutture, definite *grammatiche contestuali,* permettono allo speaker di semplificare notevolmente il linguaggio da usare durante lo speakeraggio.

Figura 2: Esempio di Dizionari NLU

Lo speaker, infatti, può pronunciare poche parole o frasi sintetiche, le quali vengono espanse dal Modulo ASR in frasi complete di senso compiuto. In tali

frasi di senso compiuto vengono inoltre inserite le parole pronunciate dallo speaker che corrispondono ai 'segnalibri'. Tali parole sono riconosciute non sull'intero dizionario di base ma su dizionari limitati di parole detti *Dizionari NLU* , quali, nel dominio *Calcio* le squadre di calcio. I Dizionari NLU sono attivabili in base al calendario delle partite disponibile nel sistema.

La Figura 3 mostra le frasi disponibili per il dominio *Calcio.* Ad esempio, la frase

```
(Squadra)(al xx°) gol vantaggio Nome
```

permette allo speaker di pronunciare le seguenti frasi sintetiche:

```
gol vantaggio Maldini
al 30° gol vantaggio Kakà
Milan gol vantaggio Kakà
```

che verranno espanse come segue:

```
MALDINI REALIZZA IL GOL DEL VANTAGGIO
KAKA' REALIZZA IL GOL DEL VANTAGGIO AL 30°
GRAZIE AL GOL DI KAKA', MILAN IN VANTAGGIO
```

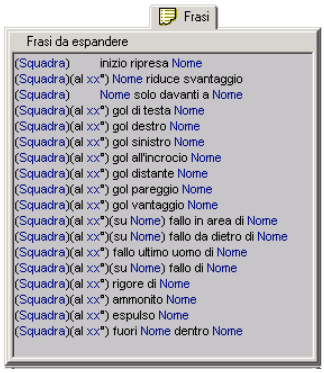

Figura 3: Speech NLU

I dizionari NLU contengono la liste di parole che il Modulo ASR riconoscerà: ad esempio il dizionario 'Squadra' contiene l'elenco dei nomi delle squadre di calcio e il dizionario 'Milan' contiene l'elenco dei nomi dei giocatori della squadra. Le grammatiche, che contengono delle parti fisse e delle parti variabili (le porzioni di testo contenute tra parentesi tonde), permettono di ridurre la variabilità delle frasi da pronunciare, velocizzando il tempo di produzione del sottotitolo e standardizzandone il testo.

Occorre sottolineare che il meccanismo è attivabile dinamicamente, personalizzabile e modificabile sia nella parte delle grammatiche che dei dizionari.

Attualmente sono disponibili due domini:

- **News** - per la sottotitolazione di telegiornali e rubriche di informazione
- **Calcio** - per la sottotitolazione di incontri di calcio, contiene 20 dizionari relativi alle squadre e ai giocatori

5. Sottotitolazione con Voice Subtitle

In questo paragrafo presenteremo l'utilizzo di Voice Subtitle per la sottotitolazione multilingue di programmi televisivi in diretta, quali trasmissioni sportive, e in semidiretta quali telegiornali, talkshow e dibattiti.

Voice Subtitle offre due modalità di lavoro: la modalità diretta e la modalità differita.
In base alla modalità di utilizzo l'interfaccia di Voice Subtitle risulta leggermente diversa, dato che in modalità differita l'Interfaccia Teletext non viene utilizzata online.

5.1 Sottotitolazione diretta

Questa modalità viene utilizzata per creare in tempo reale i sottotitoli di trasmissioni in diretta. Esempi di trasmissioni sono i collegamenti in diretta durante i telegiornali, edizioni straordinarie, partite di calcio.
Prima di attivare una sessione di sottotitolazione in diretta è necessario impostare il protocollo e le modalità di comunicazione con il servizio Teletext. L'operatore deve pertanto collegare Voice Subtitle all'interfaccia Teletext, specificando il tipo di protocollo di comunicazione (TCP/IP oppure collegamento seriale - COM), identificando il computer remoto sul quale è in esecuzione il servizio Teletext. L'operatore deve poi specificare se inviare in modo automatico o manuale i sottotitoli creati al servizio Teletext.
Se l'operatore specifica di inviare in modo automatico i sottotitoli, durante la produzione dei sottotitoli non sarà possibile effettuare revisioni e correzioni. In questa opzione è possibile specificare anche un eventuale 'Ritardo di Invio', utile per distanziare temporalmente l'invio e per dare intervalli ragionevoli di persistenza in onda a ciascun sottotitolo.
Se invece l'operatore specifica di inviare in modo manuale i sottotitoli, il revisore può revisionare ed eventualmente correggere i sottotitoli prima di inviarli.

I sottotitoli creati verranno inviati direttamente al servizio Teletext, che provvederà alla loro messa in onda in automatico.

Un esempio pratico, nella Figura 4, mostra il funzionamento del sistema per una telecronaca calcistica.
Nell'Area della Dettatura lo speaker vede la trascrizione in testo della propria dettatura, eventualmente segmentata in sottotitoli su righe. Ciascun sottotitolo viene accodato nell'Area della Revisione, da dove il sistema automaticamente provvede all'invio al servizio Teletext, con un Ritardo di Invio, in questo caso, di 300 millisecondi. Nell'Area Teletext, infine, vengono visualizzati i sottotitoli ricevuti e messi in onda dal servizio Teletext.

Nell'esempio mostrato si nota che il sottotitolo dettato dallo speaker:

```
Milan gol vantaggio Kakà
```

è stato automaticamente espanso dalle grammatiche contestuali in:

```
GRAZIE AL GOL DI KAKA', MILAN IN VANTAGGIO
```

e messo in onda.
Il sottotitolo:

```
gattuso passa una buon pallone
```

è stato automaticamente corretto dal sistema, eliminando un errore di concordanza in genere tra l'articolo `una` e il nome `pallone`. Il sottotitolo inviato al servizio Teletext e messo in onda è:

```
GATTUSO PASSA UN BUON PALLONE
```

Tutti i dati intermedi per la produzione dei sottotitoli vengono gestiti da Voice Subtitle e possono essere salvati nella sessione di dettatura, conservando la storia dei sottotitoli dettati, corretti, inviati e messi in onda dal servizio Teletext.

5.3 Sottotitolazione semidiretta

La semidiretta è una sottomodalità della diretta, utilizzabile per creare con un leggero ritardo i sottotitoli di trasmissioni in diretta.
Possono essere sottotitolati in questa modalità i servizi dei telegiornali, i talkshow e i news magazine.
Abbiamo messo a punto un modello innovativo 'a due fasi' che prevede l'introduzione intermedia di un revisore. Il revisore lavora in parallelo allo speaker e opera la fase di verifica dei sottotitoli.
Selezionando l'opzione di invio manuale dei sottotitoli al servizio Teletext, si permette al revisore di correggere, se necessario, i sottotitoli. Il revisore può inoltre scegliere, per ciascun sottotitolo, se inviarlo al servizio Teletext oppure cestinarlo.
E' necessario specificare il 'Ritardo di Invio', che definisce il tempo utilizzabile per effettuare le operazioni di revisione.

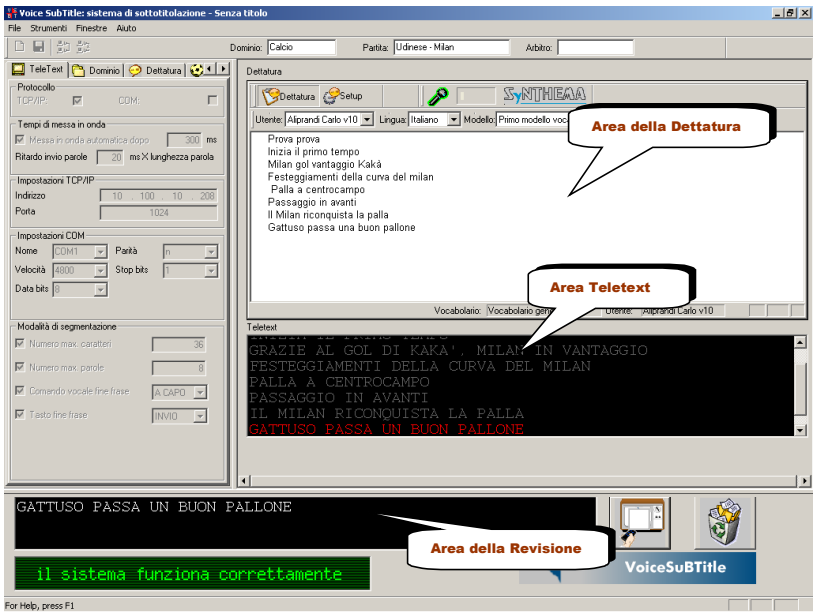

Figura 4: Sottotitolazione diretta

Il revisore riceve i sottotitoli prodotti dallo speaker con un ritardo di produzione, corrispondente all'incirca al tempo di dettatura.
Un Ritardo di Invio sufficiente, tipicamente dell'ordine di 6-8 secondi, permette al revisore di:

- verificare il sottotitolo prodotto
- correggere il sottotitolo prodotto
- inviare i sottotitoli in onda

Il revisore riceve nell'Area della Revisione un solo sottotitolo alla volta che, dopo essere inviato al servizio Teletext o cestinato, viene sostituito da un nuovo sottotitolo, se disponibile, prodotto dallo speaker.
Poiché la velocità di speakeraggio tipicamente è diversa dalla velocità di revisione e non necessariamente lo speaker e il revisore devono lavorare in sincronia, l'Area della Revisione dispone di un buffer nel quale vengono accodati i sottotitoli prodotti dallo speaker.

Nell'esempio mostrato in Figura 5 si nota che lo speaker ha dettato i sottotitoli:

```
Si gioca una importante partita
politica per gli stati uniti.
Gli americani sono chiamati in ad
eleggere 435 membri della camera.
Anche 33 di cento senatori devono
essere rieletti.
Si tratta dell'ultimo test prima
delle elezioni presidenziali del
2007
Il presidente potrebbe non
```

il revisore ha verificato il sottotitolo:

```
Anche 33 di cento senatori devono
```

correggendo un errore di riconoscimento, `di` con `dei`, e un errore stilistico, `cento` con `100`.
Il sottotitolo finale e corretto, pronto per l'invio al servizio Teletext è:

```
ANCHE 33 DEI 100 SENATORI DEVONO
```

Il buffer di accodamento dell'Area della Revisione ha memorizzato i successivi sottotitoli. All'invio al servizio Teletext, il sottotitolo

```
ESSERE RIELETTI.
```

viene presentato al revisore.

Occorre sottolineare che, anche in questa modalità, è disponibile la funzione automatica di correzione morfologica e sintattica. Tale funzione ha trattato e modificato, ad esempio, il sottotitolo:

```
SI GIOCA UN'IMPORTANTE PARTITA
```

Nostri test di laboratorio hanno mostrato che una desincronizzazione tra audio e video dell'ordine di 10-12 secondi può compensare il ritardo necessario ad effettuare le operazioni di revisione, permettendo la messa in onda dei sottotitoli in sincronia.

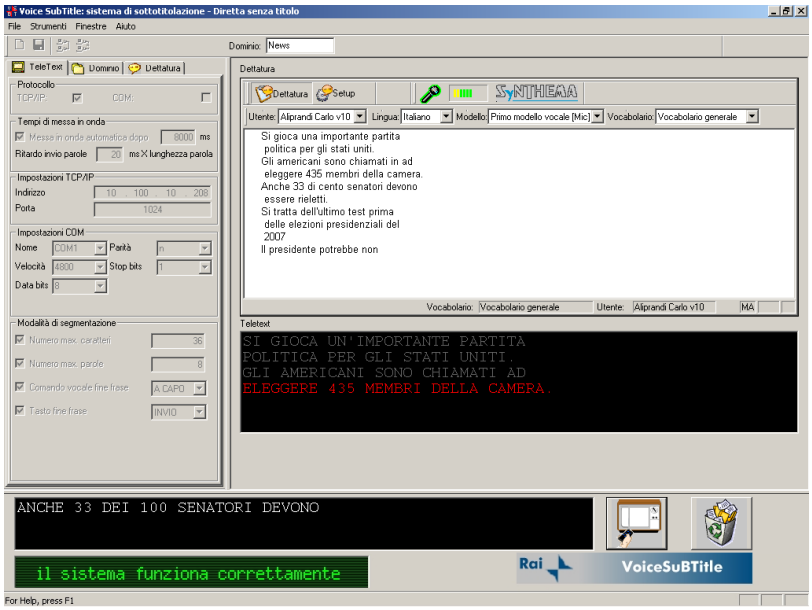

Figura 5: Sottotitolazione semidiretta

6. Resocontazione e Intersteno

Intersteno, fondata nel 1887 a Londra, è l'Associazione che mantiene quale scopo statutario quello di aumentare gli scambi di esperienze derivanti dall'utilizzo e dall'insegnamento di strumenti di ripresa del parlato e di riproduzione di testi, dalla stenografia alle moderne tecniche di riconoscimento vocale.
Intersteno, affiliata all'UNESCO, annovera tra i soci resocontisti provenienti da tutti i Continenti, rivestendo, pertanto, dimensioni mondiali.
Stenografia e dattilografia hanno da sempre rappresentato strumenti preziosi e probabilmente intramontabili, quanto a formazione del bagaglio di competenze del resocontista. Tali competenze, se un tempo si limitavano, per lo più, a mansioni di mero segretariato, oggi sono diventate tecniche insostituibili nella professione del resocontista e hanno iniziato a influenzare anche la professione del sottotitolatore. Tali professioni, infatti, hanno storicamente utilizzato tecniche simili e i progressi tecnologici di una hanno influenzato l'altra. Si pensi, ad esempio, quanto la 'catena di produzione' del resocontista sia simile a quella del sottotitolatore (cfr. paragrafo 3):

1. assistere all'evento/ascoltare l'audio
2. capirne il contenuto
3. ripresa integrale/sintesi del contenuto
4. riprodurre il testo
5. verificare il testo prodotto
6. correggere il testo prodotto

Si pensi inoltre, alla comune necessità di ottenere risultati qualitativamente eccellenti, in termini di comprensione del testo e di limitazione del numero di errori di trascrizione e scrittura, spesso da coniugare con significative velocità nella trascrizione diretta del parlato: la sottotitolazione diretta di un telegiornale deve produrre risultati simili a quelli della resocontazione di un dibattimento in aula.

Nell'ambito dei Congressi mondiali di Intersteno hanno luogo delle sessioni competitive. Alle sessioni competitive, a partire dall'edizione del 2003, è stata ammessa la tecnologia di riconoscimento vocale fra quelle utilizzabili oltre alle 'tradizionali' stenografia, stenotipia e scrittura alla tastiera.
Per ciascuna specialità di gara, Intersteno assicura parità di valutazione per qualunque tecnologia utilizzata, consolidando l'opinione largamente condivisa che lo strumento di ripresa del parlato sia ininfluente, purché si ottenga un output di qualità, nel minor tempo possibile.
Tra le sessioni competitive, il riconoscimento vocale ha trovato spazio in quattro specialità di gara:

a. Ripresa Veloce in lingua madre. Si tratta della ripresa, sotto dettatura a velocità crescente, e successiva trascrizione, di un testo della durata complessiva di 15 minuti.
b. Corrispondenza e Resoconto Sommario. Tipologia di gara in cui confluisce l'originaria 'funzione segretariale' della scrittura veloce, unitamente alle abilità di resocontazione sommaria. Consiste in una prima ripresa e trascrizione integrale di una lettera, della durata di tre minuti, che introduce il tema della seconda parte del testo di gara, in cui il concorrente dovrà realizzare un resoconto sommario di un brano della durata di 7 minuti.
c. Competizione Poliglotta. I concorrenti devono classificarsi in almeno due lingue straniere, con ripresa e trascrizione integrale dei testi di gara, della durata di 3 minuti, dettati a velocità crescente.
d. Competizione Fast. I concorrenti devono dimostrare l'abilità di ripresa del parlato e la successiva trascrizione entro tempi di consegna assolutamente contenuti, comunque non oltre 24 minuti dal termine della dettatura.

Nel corso degli anni i Campionati Intersteno hanno consentito l'emersione di significativi e importanti risultati, utilissimi nella valutazione e comparazione del 'saper fare' che deve contraddistinguere la trasposizione del discorso parlato in testo scritto. Attraverso tali appuntamenti, inoltre, gli operatori del settore hanno beneficiato di un costante aggiornamento rispetto alle tecnologie e hanno saputo rapportarle e associarle alle capacità umane che devono saper governare potenzialità e limiti degli strumenti tecnologici.

In occasione del Congresso Intersteno 2003, i risultati ottenuti dai concorrenti con tecnologia di riconoscimento vocale nella gara di ripresa veloce sono:

Pos.	Nome	Lingua	Penalità	Sillabe
1	Verruso	italiano	33	380[1]
2	Magee	inglese	29	307
3	Di Nepi	italiano	32	286

[1] Lo strumento utilizzato per la gara è IBM ViaVoice.

Il numero di penalità evidenziato rende conto degli errori commessi durante la trascrizione. Ciascuna penalità può corrispondere sia ad errori di uso della tecnologia, quali parole non riconosciute, sostituite o errate, sia ad errori 'umani', quali segni di punteggiatura mancanti o non posti correttamente. Si tratta, comunque, di una componente relativamente trascurabile, dal momento che i regolamenti di gara tollerano una soglia non superiore a circa il 4%. Mediamente, a velocità interessanti a fini professionali (per esempio, intorno a 140/150 parole al minuto), non si va oltre 5/6 penalità per minuto. La percentuale di errori è tale, quindi, da non pregiudicare l'intelligibilità del testo.

In occasione del Congresso Intersteno 2005, questi i risultati:

Pos.	Nome	Lingua	Penalità	Sillabe
1	Verruso	italiano	33	393[2]
2	Green	inglese	34	301
3	Smith	inglese	22	250

Il primato, tutto italiano, nelle gare con riconoscimento vocale, dimostra l'affidabilità dei software soprattutto in termini di accuratezza, parametro significativo per una lingua complessa e ricca di forme flesse tra loro simili. Conferma inoltre il livello qualitativo raggiunto per la lingua italiana sia dei componenti ASR che dei componenti NLU sottostanti, in particolare, per il sistema Synthema Voice Suite.

Il risultato del 2005 nella gara di ripresa veloce è inoltre significativo in quanto stabilisce il nuovo record mondiale di velocità di dettatura nella sezione di riconoscimento vocale. Tale record, pari a 393 sillabe al minuto, corrisponde a ben 174 parole al minuto, un risultato formidabile ottenuto grazie alle caratteristiche professionali dello strumento opportunamente adattato alle esigenze di competizione.
Anche il risultato della Competizione Fast è interessante. Per la prima volta nella storia di Intersteno un concorrente riesce a classificarsi in graduatoria usando la tecnica del riconoscimento vocale, precedendo diversi colleghi che concorrevano con tecniche classiche quali stenotipia e stenografia.

[2] Lo strumento utilizzato per la gara è Synthema Voice Suite.

Tali risultati sono stati ottenuti, oltre che grazie alla validità e all'adattabilità dello strumento tecnologico, soprattutto grazie ad un percorso formativo e ad un allenamento mirati. Il raggiungimento di adeguate velocità professionali non prescinde infatti da percorsi didattici specifici e affidabili, per serietà e organizzazione.
L'apprendimento dei software di riconoscimento vocale è possibile in tempi senz'altro più rapidi rispetto alle tecniche di ripresa stenografica o stenotipica: il raggiungimento di elevate velocità di ripresa, infatti, procede non già attraverso il consolidamento delle regole abbreviative dei sistemi di scrittura veloce, ma con naturali esercizi di dettatura, progressivamente più intensi e veloci, fino ad acquisire l'abilità di "star dietro" all'oratore. Tali esercizi portano parallelamente all'accrescimento dei margini di accuratezza dei software. Occorre agire su due fronti: dettatura di testi e inclusione di nuovi vocaboli, soprattutto di quelli che fanno parte dei topics di riferimento rispetto al proprio settore lavorativo (medico, giuridico o giornalistico, ad esempio). Si determina, così, un progressivo miglioramento del profilo vocale dell'utente che, per esperienza, diventa interessante dopo un consistente esercizio paziente e quotidiano.
Il miglioramento consente un abbattimento degli errori, portando la copertura al 96%-97%, una percentuale che, come discusso nel paragrafo 4.1, determina una qualità più che accettabile da parte dell'operatore. Anche in questo caso, come per la sottotitolazione, il grado di soddisfazione si rivela determinante, perché altrimenti prevarrebbe il desiderio di ricorrere ad altre tecniche.

L'esperienza e i risultati Intersteno dimostrano, pertanto, come la neonata tecnica del riconoscimento vocale s'imponga con pari dignità rispetto alle note stenografia e stenotipia. Ma adeguati percorsi formativi precedono sempre l'affermazione dei risultati migliori.

Autori:

Dr. Carlo Aliprandi
Synthema Srl
Via Malasoma 24, Pisa
carlo.aliprandi@synthema.it

Dr. Fabrizio Gaetano Verruso
Assemblea Regionale Siciliana
P.zza del Parlamento 1, Palermo
fverruso@ars.sicilia.it