SPECIFICHE LINGUISTICHE DEL DATABASE UTILIZZATO PER LO SPEAKER RECOGNITION IN S.M.A.R.T.[1]

Luciano Romito, Tommaso Bove, Stefano Delfino, Carla Rossi, Giovanna Jona Lasinio
Università della Calabria – Laboratorio di Fonetica
luciano.romito@unical.it

1. INTRODUZIONE

All'interno dei Programmi europei OISIN I, OISIN II e AGIS 2005[2], è stato sia sviluppato un progetto di ottimizzazione di un metodo per il riconoscimento del parlatore sia in ambito forense sia, di investigazione preventiva, dal titolo SMART (Statistical Methods Applied to the Recognition of the Talker).

Il progetto, che volge ormai al termine, aveva come obiettivi: il miglioramento delle tecniche di Speaker Recognition; l'implementazione e l'elaborazione statistica dei dati vocali; e soprattutto la costruzione di un corpus etichettato e archiviato (db d'ora in poi), che offre la possibilità di selezionare e organizzare sottocorpora sulla base di variabili linguistiche, anagrafiche e tecniche quali il dialetto, l'inventario fonologico, l'età del parlante, lo zona geografica di provenienza, il canale di registrazione ecc. Tali sottocorpus potrebbero essere utilizzati come comunità linguistica di confronto (Reference Population) per una precisa comparazione fonica.

I metodi statistici utilizzati all'interno del progetto per la comparazione, sono già stati presentati in lavori precedenti come: Brutti P et al. 2002, Bove T. et al. 2002, Bove T. et al. 2003, Bove T. et al. 2004, Bove T. 2006.

Figura 1: Pagina iniziale del software SMART

[1] Lo S.M.A.R.T. (Statistical Methods Applied to the Recognition of the Talker) è un progetto di ricerca finanziato dalla Comunità Europea – Direzione Generale Giustizia, Libertà e Sicurezza.

[2] Programma OISIN II (Rif. JAI/2002/OIS/035).

2. LA COMPARAZIONE FONICA

Nel caso di una comparazione fonica, l'esperto, opera su una voce anonima, una voce nota e deve disporre di una comunità linguistica di riferimento per poter stimare il rapporto di verosimiglianza e gli errori di falsa identificazione e di mancato riconoscimento[3]. Al momento attuale l'unico sistema che utilizza un database di voci di riferimento viene utilizzato dall'Arma dei Carabinieri e prodotto dalla FUB (Fondazione Ugo Bordoni)[4]. Tale metodo utilizza un database unico per la popolazione italiana, è costruito su 4 vocali e 4 Frequenze Formantiche. Resta implicito che i segmenti scelti siano ritenuti ugualmente distribuiti su tutto il territorio.

In questa sede la nostra attenzione, è rivolta alla struttura interna del db, in particolare alle categorie linguistiche scelte e alla etichettatura di ogni singolo file. Le motivazioni di tali scelte nascono dalla convinzione che non esista un italiano standard, che la maggior parte delle intercettazione contenga voci dialettali e che i dialetti italiani sono molto differenti tra loro per segmenti vocalici e inventari fonologici utilizzati (la letteratura a riguardo è talmente vasta da risultare inutile qualunque citazione).

In definitiva crediamo che considerare gli *italiani* come appartenenti ad una unica comunità linguistica, sia un grande errore.

3. IL DATABASE

Il db è allocato in un server centrale e prevede l'aggiunta, la modifica o l'eliminazione dei dati da parte di client autorizzati distribuiti sul territorio italiano e europeo[5]. Prevede, anche, la possibilità di interrogazioni attraverso un motore di ricerca e di analisi su tutto il materiale esistente o su una sua parte. Ovviamente tutti i processi, come anche la sola consultazione sono regolati da un sistema di autorizzazioni e privilegi assegnati ad ogni singolo utilizzatore (o esperto).

Le operazioni che potranno essere effettuate avranno due finalità. Innanzitutto, dopo un analisi linguistica sulle voci da comparare, si estrapolerà ed identificherà il sottocorpus appartenente alla stessa comunità linguistica di parlatori ed in seguito si impiegheranno gli algoritmi per il riconoscimento del parlatore e si effettuerà la comparazione fonica.

Le specifiche linguistiche e non-linguistiche date alle etichette del db, possono essere differenziate in diverse tipologie: informazioni riguardo il supporto; informazioni riguardo la registrazione del file e le sue caratteristiche acustiche principali; le impressioni acustico-percettive sul segnale archiviato; le informazioni anagrafiche e geografiche relative al parlatore e infine le informazioni che entrano nel merito del contenuto della registrazione quindi la lingua o dialetto, l'inventario fonologico di detta lingua/dialetto, il modo di fonazione, informazioni su alcune variabili linguistico-fonologiche ecc. Sono presenti anche alcune parole chiave che riguardano le variabili diafasiche (saggio, telefonata estorsiva, ecc) o l'argomento della conversazione (droga, armi, rivendicazioni ecc). Tale scelta può risultare molto importante perché, nel caso della variabile diafasica è ovvio che

[3] Per le metodologie adottate in Italia in ambito di Speaker Recognition, si veda Romito 2006, 2007a e 2007b.

[4] Ovviamente è possibile che alcuni privati abbiano costruito in proprio un database da utilizzare in casi di Speaker recognition anche se l'indagine di Romito 2006 rivela tendenze esattamente opposte.

[5] Almeno per quanto riguarda i partner europei del progetto.

lo stato psicologico del parlatore durante una estorsione non è lo stesso di quello presente durante un saggio o una telefonata amorosa, quindi è indubbio che al variare della situazione il suo parlato risulterà differente (cfr. Romito et al. 1997 e 1998). Questo tipo di archiviazione permette di estrapolare un sottogruppo, dal db, costituito di solo voci caratterizzate dall'essere di qualità omogenea (ad esempio telefonica) e di variabile, ad esempio, estorsiva. In quest'ultimo caso, dal punto di vista sia scientifico che di indagine, potrebbe risultare interessante comparare la produzione linguistica di diversi parlatori sullo stesso argomento, per esempio per individuare gerghi o codici particolari (si pensi all'importanza di tale accertamento in relazione alle rivendicazioni).

La scheda di archiviazione comprende alcune informazioni storiche sul file come il numero del fascicolo o del protocollo, la data dell'analisi, l'ufficio richiedente, l'autorità giudiziaria, il numero di procedimento penale, il tipo di reato, la data di registrazione e la data di inizio lavori.

Informazioni sul canale come Telefono, Cellulare GPS, Cellulare UMTS, VOIP, registrazione ambientale, ambientale in auto, ambientale durante un colloquio in carcere ed è inoltre possibile aggiungere nuovi canali anche se non immediatamente previsti dagli attuali menu.

Informazioni sul supporto come Analogico e Digitale. Nel caso di supporto analogico verrà specificato il tipo di supporto: *cassetta*, *microcassetta* o *bobina*; nel caso di supporto digitale verrà segnalato se il supporto sarà un *CD*, un *DVD* o altro. Se si tratta di un nastro allora dovrà essere indicato se si tratta di cassetta *normal*, *Chrome*, *metal* e se lo stato del supporto è *eccellente*, *buono* ecc. Bisogna indicare la velocità di registrazione, se questa è analogica, la frequenza di campionamento se questa è digitale. Verrà segnalato il numero di bit e se la registrazione è *stereo* o *mono*.

Altre informazioni che non riguardano il supporto ma la registrazione sono di tipo percettivo o analitico. Nel primo caso la registrazione sarà segnalata come *distorta*, *saturata* nel secondo invece verrà segnalato il tipo di fonazione come voce *mormorata*, *bisbigliata* ecc. o il tipo di lingua utilizzata e nel caso di italiano, il dialetto utilizzato. In questo caso viene segnalato anche se tutta la registrazione è in dialetto, in italiano o se la produzione è mistilingue. All'interno delle analisi linguistiche viene segnalato il numero di segmenti vocalici presenti nella lingua o nel dialetto e la distribuzione di detti segmenti nello spazio vocalico (utilizzo di tratti come arrotondamento (F3) o nasalizzazione (FN) o durata (ms)).

Le informazioni anagrafiche sul singolo parlatore invece, richiedono il nome, l'età, il grado di scolarizzazione, provincia di provenienza, la provincia di residenza, ecc.

4. INTERVENTI POSSIBILI SUL DATABASE

Il database come già descritto prevede l'inserimento di dati fissi e di dati dinamici. È possibile inserire un file wav contenente una registrazione completa (per esempio una conversazione telefonica tra due interlocutori) o una singola voce presente in una registrazione più ampia (per esempio solo la voce appartenente all'interlocutore intercettato o sotto controllo). Il file viene etichettato e catalogato attraverso informazioni fisse (utenza telefonica, giorno mese ed anno nonché ora durata della registrazione, ecc) e informazioni dinamiche come nuove trascrizioni, nuove operazioni di filtraggio e di ripulitura, nuove misure acustiche effettuate in seguito ad un accertamento di riconoscimento del parlatore con algoritmi diversi o altri accertamenti attraverso analisi dialettali, fonetiche e fonologiche, morfologiche o sintattiche e stilistiche, ecc.

È possibile modificare o manipolare le **informazioni dinamiche** presenti nel db. Aggiungendo informazioni e analisi su più livelli, misure acustiche o relative alla durata, parametri come *articulation rate*, *speech rate*, o *fluency* o anche parametri che al momento attuale non vengono considerati come *andamento storico della Frequenza Fondamentale o delle Frequenze Formantiche* ecc. Ovviamente per effettuare qualunque modifica l'operatore[6] deve essere effettuata dall'amministratore del db.

5. IL RISULTATO DELLA COMPARAZIONE

Una volta che la comunità linguistica è stata identificata e la stessa viene *trasferita* sul *client* (in modo da non appesantire le operazioni del server), è possibile effettuare la comparazione fonica.

Usando l'approccio del massimo likelihood, viene stimata la soglia di confidenza cioè la peculiare caratteristica dello speaker analizzato, l'errore A-priori di *Falsa Reiezione* e di *Falsa Accettazione*.

Se la distanza tra la voce nota e la voce anonima è maggiore della soglia stimata allora le due voci non appartengono allo stesso parlatore, viceversa se la distanza è minore allora le due voci appartengono allo stesso parlatore o quanto meno sono compatibili. In entrambi i casi vengono presentati anche i valori di errore di Falsa Identificazione e di Falsa Accettazione.

I risultati dell'analisi comprendono una completa valutazione prevedendo un grafico rappresentante la distanza intra e inter parlatore con la soglia massima di likelihood (LR); un *report* contenente due tipi di errori a-priori e tutte le informazioni legate all'analisi; un *grafico* rappresentante la distanza intra e inter speaker all'interno della attuale distanza tra voce nota e voce anonima; un *report* contenente due tipi di errori a-posteriori e tutte le informazioni relative alla comparazione tra voce anonima e voce nota (per approfondimenti sul metodo parametrico si vedano i lavori già citati di Brutti P et al. 2002, Bove T. et al. 2002, Bove T. et al. 2003, Bove T. et al. 2004, Bove T. 2006.

6. CONCLUSIONI

Il progetto SMART non vuole essere solo un metodo per la comparazione di voci note e anonime, ma anche un software di supporto per l'esperto nella complessa operazione di comparazione. L'individuazione della comunità linguistica della voce nota e della voce anonima, l'identificazione di caratteristiche fonetico-fonologiche simili, l'identificazione di una sottocomunità ristretta di voci prodotte da locutori non solo appartenenti alla stessa comunità linguistica ma anche con caratteristiche simili quali l'età, il canale di registrazione e addirittura la tipologia diafasica rendono più agevole e sicuramente più preciso il risultato. Il database può anche essere consultato per studiare gerghi, codici ristretti o variabili stilistiche particolari delle solo estorsioni o delle sole rivendicazioni, ecc.

La stessa creazione del database risulta essere utile sia agli esperti sia agli operatori, infatti è da stimolo e da volano per procedure di standardizzazione sui metodi di misura,

[6] Per operatore, in questo caso si intende anche un dialettologo che effettuata analisi e misure su un corpus di parlato spontaneo telefonico o ambientale di una certa zona geografica. I suoi risultati scientifici potranno essere utilizzati da un esperto perito in fase di indagine.

sugli algoritmi da utilizzare, sulle statistiche decisionali e sicuramente sulla compartecipazione di diversi metodi all'interno dello stesso ambiente informatico.

Allo scopo sono già iniziati lavori di analisi per studiare fenomeni quali la variabilità interparlatore o intercomunità rispetto a quella intra parlatore ecc. si veda Romito Lio 2008.

BIBLIOGRAFIA

Brutti P, Fabi F. Jona Lasinio G. (2002), Una proposta di meta-analisi basata sulla combinazione di classificatori per il problema del riconoscimento del parlatore, *Statistica*, 3 pp. 455:473.

Bove T., Brutti P., Fabi F., Jona Lasinio G., Giua P.E., Forte A. e Rossi C., (2002), Tecnologie Informatiche nella Promozione della Lingua Italiana, 25-26 giugno 2002, Progetto di ricerca S.M.A.R.T. 2, Conferenza: *T.I.P.I.*, pp. 121-124.

Bove T., Brutti P., Fabi F., Jona Lasinio G., Giua P.E., Forte A. e Rossi C. (2003), Three approaches to the speaker identification problem for forensic use, in *atti del Convegno Cladag* 22-24 settembre 2003 (relazione invitata).

Bove T. Jona Lasinio G., Rossi C.. (2004), The speaker recognition problem, XLII Riunione Scientifica della Società Italiana di *Statistica*, pp. 429:440, (relazione invitata)

Bove T., (2006), SMART II: Analisi della voce Conferenza T.A.L.–in *Trattamento Automatico Linguaggio* – Sessione 3^: Intelligence 9-10 marzo 2006.

Trumper J., Romito L., Maddalon M., Mendicino A., Belluscio G. M. G., (1993), Stime manuali: un esperimento. Atti del convegno *Teoria e Sperimentazione: Parametri, tratti e segmento*, Calabria, 28-29 novembre, 1991, Esagrafica Roma:Roma, Vol. XIX, pp. 61-79.

Trumper J., Romito L., Mendicino A., Li J., (1994), Stime vocaliche problematiche. Atti del convegno *Giornate di Studio del Gruppo di Fonetica Sperimentale*, Padova, 1992, Vol. XX, pp. 49-59.

Romito L., Maddalon M., Trumper J., (1996), La parametrizzazione nei test di riconoscimento. Atti del convegno *VIe Giornate di Studio del Gruppo di Fonetica Sperimentale (G.F.S.)*, Roma, 1996, pp. 87-93.

Romito L., Maddalon M., Trumper J., (1996), Atteggiamento della Magistratura nei confronti delle perizie foniche. Atti del convegno *VIe Giornate di Studio del Gruppo di Fonetica Sperimentale (G.F.S.)*, Roma, 1996, Roma, pp. 34-45.

Romito L., *Manuale di Fonetica articolatoria, acustica e forense*, (2000), Università degli Studi della Calabria: centro editoriale e Librario.

Romito L., (2003), Passato Presente e Futuro nelle Analisi di Speaker Recognition. In *Voce Canto Parlato*, Zamboni A. (a cura di), Padova: Unipress, pp. 237-246.

Romito L., Galata' V., (2004), Towards a protocol in speaker recognition analysis. *Forensic Science International*, pp. 105-113.

Romito L., (2004), La misura dell'intelligibilità e il rapporto segnale-rumore. *Atti del convegno "AISV* (Associazione Italiana di Scienze della Voce)", Padova, 2004.

Romito L., (2005), La competenza linguistica nelle trascrizioni Forensi: l'intelligibilità, l'oggettività e il rapporto segnale/rumore. *Detective And Crime*.

Romito L., Blefari M., (2005), Towards a new parameter in Recognition. *Speech Language and the Law*.

Romito L., Galata' V., Lio R., (2006), Fluency Articulation and Speech Rate as new parameters in the Speaker Recognition. Atti del convegno "*III Congreso de Fonética Experimental*", Santiago de Compostela, 26-24 ottobre, 2005, Xunta de Galicia:Santiago de Compostela, pp. 537-549.

Romito L., Galata' V., (2007), Speaker Recognition: Stato dell'arte in Italia. Atti del 3° Convegno Nazionale AISV, "*Scienze Vocali e del Linguaggio Metodologie di Valutazione e Risorse Linguistiche*", Trento, 29/11-1/12, 2006, EDK Editore SRL:RN, Vol. 3.

Romito L., Galata' V., (2008), Speaker Recognition in Italy: evaluation of methods used in forensic cases. *Atti del IV Convegno CFE*, Granada - Spagna, 11-14/02, 2008.

Romito L. , Tucci M. , (2008), Verso un formato standard nelle intercettazioni e una proposta per l'archiviazione e la conservazione delle registrazioni. Atti del IV convegno AISV - *La fonetica sperimentale. Metodo e applicazioni*, Cosenza, 3-5/12/2007, 2007, EDK s.r.l.:RN, Vol. IV.

Romito L., Scullari V., (2008), Un protocollo delle procedure di restauro all'interno dell'Archivio sonoro calabrese. *Atti del IV convegno AISV*, Università della Calabria, 3-5/12, 2007, EDK Editore SRL:RN, Vol. 4.

Romito L., Lio R., (2008), Stabilità dei parametri nello Speaker Recognition: la variabilità intra e interparlatore. *Atti del IV convegno AISV* (Associazione Italiana Scienze della voce), Unical, Campus di Arcavacata di Rende (CS), 3-5/12, 2007, EDK Editore s.r.l.:RN, Vol. IV.

G. Cavarretta L. Romito, M. Tucci, (2008), Verso un formato standard nelle intercettazioni: archiviazione, conservazione, consultazione e validità giuridica della registrazione sonora, in *Atti del Convegno Internazionale Ass.I.Term*, Università degli studi della Calabria 6-8/6/2008.

7. APPENDICE

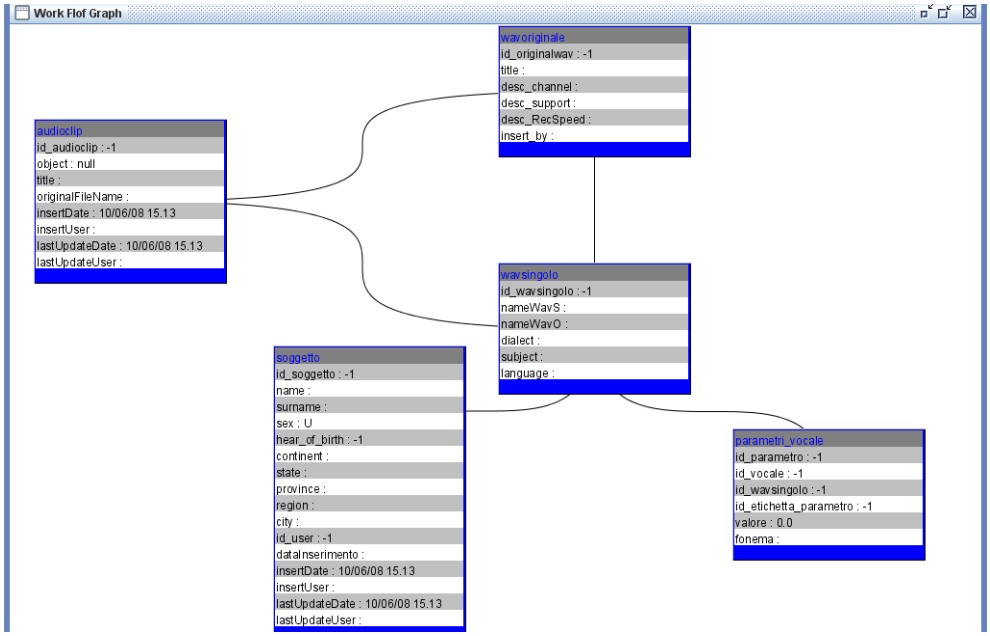

Figura 2: Diagramma di flusso del software SMART

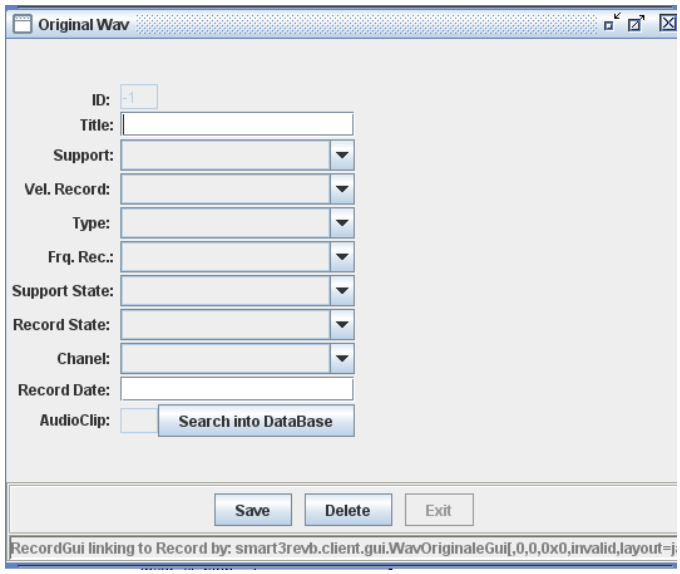

Figura 3: Maschera per l'inserimento dei dati riguardanti il supporto e il file.

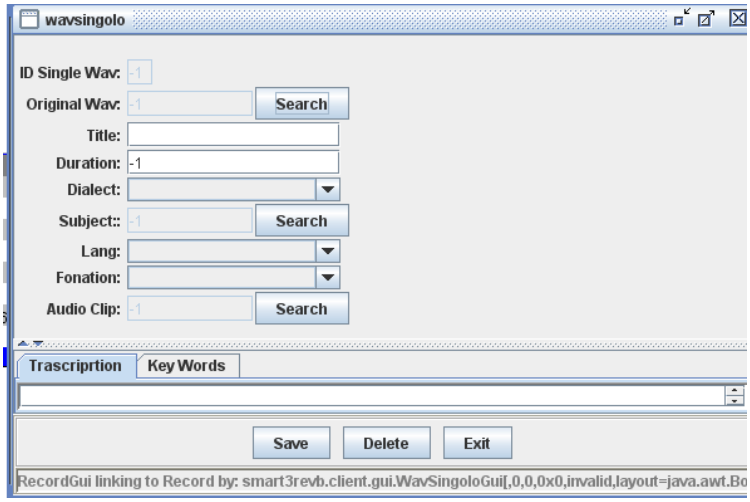

Figura 4: Maschera per l'inserimento dei dati riguardanti la registrazione.

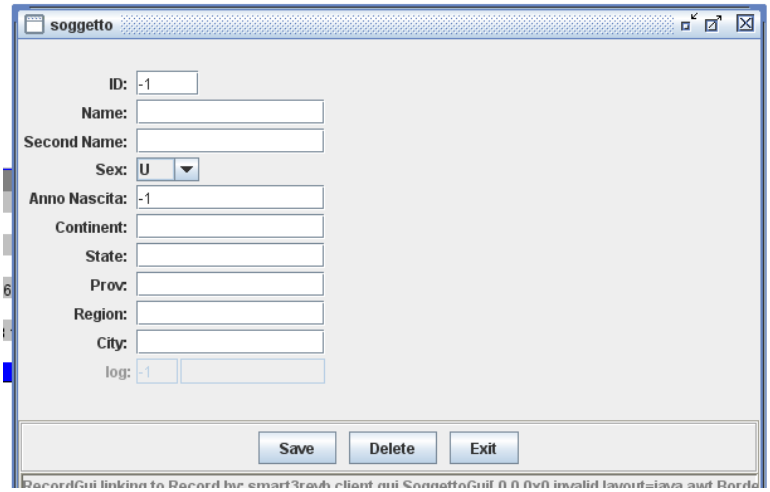

Figura 5: Maschera per l'inserimento dei dati anagrafici del locutore.

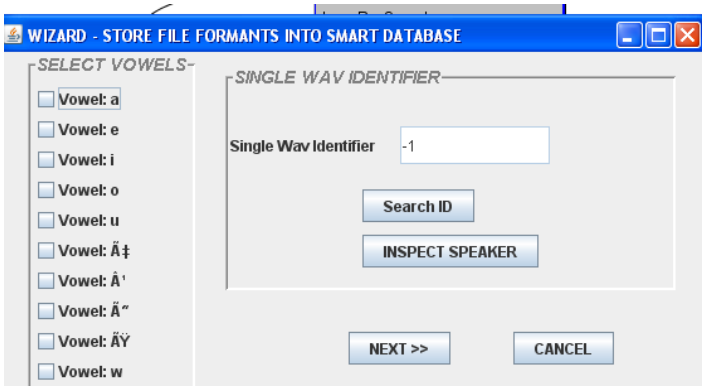

Figura 6: Maschera per l'inserimento dei dati acustici

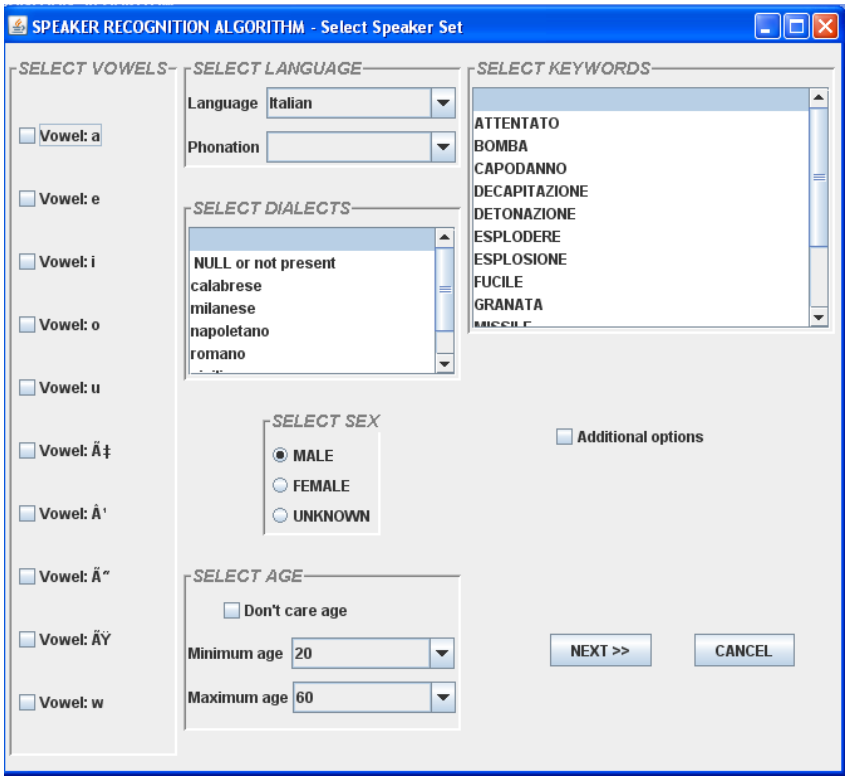

Figura 7: Maschera per l'inserimento e per la ricerca di parole chiave nelle trascrizioni allegate alle registrazioni