# SVILUPPO DI UN SISTEMA DI RICONOSCIMENTO PER L'ARABO: PROBLEMI E SOLUZIONI

Piero Cosi, Mauro Nicolao, Giacomo Sommavilla, Graziano Tisato

Istituto di Scienze e Tecnologie della Cognizione, Sede di Padova "Fonetica e Dialettologia",
Consiglio Nazionale delle Ricerche
via Martiri della Libertà, 2 - 35137 Padova, Italia
*{cosi, nicolao, sommavilla, tisato} @ pd.istc.cnr.it*

## 1. INTRODUZIONE

L'Arabo è attualmente una delle lingue più parlate nel mondo. Il numero di parlanti arabi è valutato intorno a 325 milioni, di cui approssimativamente 225 milioni sono parlanti L1 e 100 milioni sono parlanti L2. L'Arabo è la lingua ufficiale in più di 22 paesi e la lingua ufficiale per l'istruzione religiosa dell'Islam, di conseguenza si stima che anche molti altri parlanti ne abbiano una conoscenza almeno passiva.

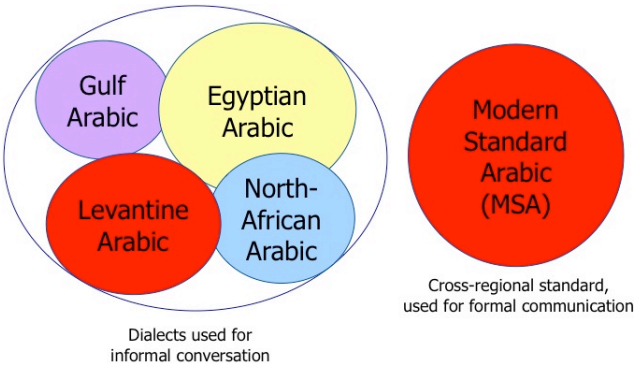


Figura 1: Le varietà linguistiche dell'Arabo (Kirchhoff et alii, 2002).

*Varietà linguistica*

Prima di tutto, è importante rendersi conto che quando ci si riferisce alla lingua Araba non s'intende una singola varietà linguistica, ma, com'è illustrato in Figura 1, più che altro, ad una collezione di dialetti linguisticamente anche molto differenti tra loro.

Esiste l'Arabo Classico, che è una lingua antica, letteraria, importante principalmente perché usata nel Corano.

Con il termine Arabo invece, normalmente, s'intende il cosiddetto Arabo Standard Moderno (MSA[1]) che è una versione dell'Arabo Classico con un vocabolario modernizzato. MSA è convenzionalmente adottata come lingua comune a tutti i paesi di cultura araba ed è utilizzata dai mezzi di informazione (giornali, radio, TV), nei discorsi ufficiali e in tutti i principali canali di comunicazione internazionali.

Tuttavia, MSA non è quasi mai usato per la comunicazione quotidiana e informale, per la quale invece vengono impiegati i vari dialetti locali. I dialetti dell'Arabo possono approssimativamente essere divisi in due gruppi:

[1] MSA: Modern Standard Arabic

- Arabo Occidentale (Western Arabic), che include i dialetti parlati nel Marocco, in Algeria, in Tunisia ed in Libia,
- Arabo Orientale (Eastern Arabic), che può ulteriormente essere suddiviso in Egiziano, Levantino e Arabo del Golfo Persico.

| Cambiamenti | MSA | ECA |
|---|---|---|
| /θ/ → /s/, /t/ | /θala:θa/ | /tala:ta/ |
| /ð/ → /z/, /d/ | /ðahab/ | /dahab/ |
| /ay/ → /e:/ | /saif/ | /se:f/ |
| Inflessioni | yatakallam(u) | yitkallim |
| Vocabolario | tawila | tarabeeza |
| Ordine delle parole | VSO | SVO |

Tabella 1: Alcune differenze tra MSA e ECA.

Questi dialetti differiscono considerevolmente sia tra di loro e sia rispetto al MSA dal quale tutti sono egualmente distanti. Le differenze interessano tutti i livelli della lingua, cioè fonetica, fonologica, lessicale, morfologica e di sintassi. La Tabella 1 illustra, ad esempio, alcune differenze tra MSA e un dialetto, ECA[2].

*La struttura della lingua araba*

L'Arabo è una lingua scritta da destra a sinistra, ha un alfabeto di ventotto lettere, con venticinque consonanti e tre lettere restanti che rappresentano le vocali cosiddette "lunghe" (/i:/, /a:/, /u:/) oppure, ove necessario, le semivocali (/y/ e /w/). In Arabo, il termine "vocale lunga" indica una vocale pronunciata con particolare intensità, di solito il termine viene utilizzato in contrapposizione con il termine "vocale breve", che indica invece un suono vocalico di breve durata e intensità. Un'analogia con questa funzione fonologica della durata, si può riscontrare, ad esempio nella lingua inglese, tra le parole *bit* e *beat* o tra *put* e *book*. Una corrispondenza con la lingua italiana è di più difficile individuazione poiché in essa le vocali di lunga o breve durata, non introducono una differenza fonologica. Per semplificare, si potrebbe identificare un'analogia nella differenza tra vocali accentate e non accentate.

In Arabo, la scrittura delle lettere non è univoca. Ogni grafema può apparire in quattro differenti forme a seconda che sia isolato o che occupi una posizione iniziale, centrale o finale all'interno di una parola. Inoltre, non esiste differenziazione tra lettere maiuscole e minuscole.

Una caratteristica distintiva del sistema arabo di scrittura è che le vocali brevi non sono rappresentate da lettere dell'alfabeto, grafemi veri e propri, ma sono contrassegnate dai cosiddetti segni *diacritici* che vengono scritti sopra o sotto la consonante che precede il suono vocalico. Altri diacritici sono presenti nella lingua Araba e sono utilizzati per contrassegnare altri fenomeni di pronuncia, quali ad esempio il raddoppiamento della consonante o l'assenza di suono vocalico.

Un elenco delle lettere della lingua Araba sono elencati in Tabella 2, mentre la Tabella 3 riporta i diacritici e le combinazioni di lettere particolari. Fatta esclusione per questi ultimi casi particolari, la corrispondenza tra grafemi e fonemi è uno a uno.

[2] ECA: Egyptian Colloquial Arabic

| Isolated | Beginning | Middle | End | Name | Phoneme |
|---|---|---|---|---|---|
| ا | ا | ا | ا | ’alif | /a:/ |
| ب | بـ | ـبـ | ـب | baa’ | /b/ |
| ت | تـ | ـتـ | ـت | taa’ | /t/ |
| ث | ثـ | ـثـ | ـث | thaa’ | /θ/ |
| ج | جـ | ـجـ | ـج | gym | /ʤ/ |
| ح | حـ | ـحـ | ـح | Haa’ | /ħ/ |
| خ | خـ | ـخـ | ـخ | khaa’ | /x/ |
| د | د | ـدـ | ـد | daal | /d/ |
| ذ | ذ | ـذـ | ـذ | dhaal | /ð/ |
| ز | ز | ـزـ | ـز | zayn | /z/ |
| ر | ر | ـرـ | ـر | raa | /r/ |
| س | سـ | ـسـ | ـس | syn | /s/ |
| ش | شـ | ـشـ | ـش | shyn | /ʃ/ |
| ص | صـ | ـصـ | ـص | Saad | /ʂ/ |
| ض | ضـ | ـضـ | ـض | Daad | /ɖ/ |
| ط | طـ | ـطـ | ـط | Taa’ | /ʈ/ |
| ظ | ظـ | ـظـ | ـظ | Zaa’ | /ʐ/ |
| ع | عـ | ـعـ | ـع | ’ayn | /ʕ/ |
| غ | غـ | ـغـ | ـغ | ghayn | /ʁ/ |
| ك | كـ | ـكـ | ـك | kaaf | /k/ |
| ق | قـ | ـقـ | ـق | qaaf | /q/ |
| ف | فـ | ـفـ | ـف | faa’ | /f/ |
| ل | لـ | ـلـ | ـل | laam | /l/ |
| ن | نـ | ـنـ | ـن | nuwn | /n/ |
| م | مـ | ـمـ | ـم | mym | /m/ |
| ه | هـ | ـهـ | ـه | haa’ | /h/ |
| و | و | ـو | ـو | waaw | /u:/ |
| ي | يـ | ـيـ | ـي | yaa’ | /i:/ |
| ء | أ | ئ | ئ | hamza | /ʔ/ |

Tabella 2: Lettere dell’alfabeto arabo[3]. Il nome delle lettere è dato secondo la romanizzazione Qalam[4] e i fonemi corrispondenti sono in notazione IPA.

[3] Fonte (Kirchhoff et alii, 2002).

È da segnalare comunque che i testi in lingua Araba non sono quasi mai diacriticizzati: è prassi utilizzare i diacritici solo nei casi ambigui. La mancanza di segni diacritici può portare a un numero considerevole di ambiguità del lessico. Queste devono essere risolte attraverso informazioni contestuali che presuppongono la conoscenza della lingua; di conseguenza, senza questa conoscenza, è impossibile determinare l'effettiva pronuncia di un testo non vocalizzato.

| Example | Symbol Name | Meaning |
|---|---|---|
| أَ | fatHa | /a/ |
| إِ | kasra | /i/ |
| أُ | Damma | /u/ |
| رّ | shadda | consonant doubling |
| درْس | sukuwn | absence of vowel after consonant |
| أً | tanwyn al-fatHa | /an/ |
| إٍ | tanwyn al-kasr | /in/ |
| أٌ | tanwyn aD-Damm | /un/ |
| ى | 'alif maqsuwra | /a:/ sound, historical |
| هٰذه | dagger 'alif | /a:/ sound, historical |
| آ | madda | double alif |
| في ٱلبيت | waSla | on 'alif in *al* |
| لا | laam 'alif | combination of laam and 'alif |
| ة | taa marbuwta | morphophonemic marker |

Tabella 3: Diacritici della lingua Araba[5].

Per fare un esempio facilmente comprensibile, sarebbe come se la parola italiana *stucco*, fosse pronunciata normalmente, ma scritta *stcc.* Si vede immediatamente che essa sarebbe equivalente ad altre parole quali: *stucca, stacco, stacca, stecco, stecca, etc...*

Nel caso di una frase in italiano, si potrebbe dire che la frase pronunciata come:

*"L'Arabo scritto non è semplicemente leggibile poiché è una lingua senza vocali brevi"*

verrebbe scritta come:

*"L'Arb scritt non è smplcemnt lggibl pché è un ling senz vcal brev"*

Un esempio in lingua Araba, la parola كتب (ktb) ha 21 potenziali diacriticizzazioni.

Il discorso appena fatto vale per MSA, se ci si sposta nell'ambito dei dialetti arabi, le cose si possono complicare ulteriormente. Il principale motivo è che i dialetti sono delle

[4] *Qualam,* http://eserver.org/langs/qalam.txt
[5] Fonte (Kirchhoff et alii, 2002).

lingue quasi esclusivamente orali e quindi non esiste uno standard di scrittura con uniformità di convenzioni.

## 2. ROMANIZZAZIONE AUTOMATICA

Dal momento che i grafemi arabi non sono facilmente comprensibili al lettore non esperto e che la loro elaborazione, tramite processi automatizzati di elaborazione dei dati, è piuttosto complicata, quando si studia la lingua araba, nasce presto l'esigenza di tradurre i grafemi con dei simboli romanizzati. Questi tentativi di semplificare la scrittura araba si chiamano *traslitterazioni* o *trascrizioni.*

Con il termine *traslitterazioni* s'intendono tutti quei sistemi di romanizzazione dei grafemi arabi che hanno come scopo mantenere una corrispondenza uno a uno tra i caratteri arabi e l'insieme dei simboli romanizzati (normalmente ASCII standard). Molta meno importanza è data alla coerenza tra simbolo usato e l'effettivo suono che dovrebbe essere pronunciato. Compaiono quindi anche simboli come i segni di interpunzione o le parentesi.

Allo scopo di mantenere la corrispondenza tra suono e simbolo sono utilizzate le *trascrizioni*. Queste possono essere di due tipi:

- *fonemico,* che rende il lettore che non ha familiarità con la lingua originale in grado di pronunciarla con una ragionevole accuratezza;
- *fonetico,* in cui si cerca di essere ancora più precisi nel descrivere, quanto più esattamente possibile il suono che viene pronunciato nella lingua originale, utilizzando simboli fonetici (ad esempio IPA).

Un esempio di romanizzazione di una parola araba è riportato in Tabella 4 e un elenco dei principali sistemi di romanizzazione sono illustrati in Tabella 5. Il metodo che comunemente viene utilizzato in questo ambito è Buckwalter (Buckwalter).

| | | |
|---|---|---|
| **Arabico** | القاهرة | [originale] |
| **Traslitterazione** | AlqAhrp | [buckwalter] |
| **Trascrizione Fonemica** | il Cairo<br>al-Qahira<br>al-Qāhirah | [Trascrizione Primaria]<br>[Trascrizione Standard]<br>[Trascrizione Diretta] |
| **Trascrizione Fonetica** | al kæhirah | [IPA] |

Tabella 4: Esempi di romanizzazione della lingua Araba.

I fonemi della lingua araba sono in tutto trentuno, quelli illustrati in Tabella 2 più le tre vocali brevi (a, i, u) che sono la controparte delle vocali lunghe (a:, i:, u:); è da notare che la lunghezza delle vocali è fonemica.

| letter | Qalam | CAT | LDC | Karboul | Buckwalter |
|---|---|---|---|---|---|
| ’alif | aa | aa | aa | A | A |
| baa’ | b | b | b | B | b |
| taa’ | t | t | t | TH | t |
| thaa’ | th | c͡th | th | Tt | v |
| gym | j | j | j | J | j |
| Haa’ | H | h/H | h | H | H |
| khaa’ | kh | k/kh/K | kh | Kk | x |
| daal’ | d | d | d | D | d |
| dhaal’ | dh | zzh | dh | Dd | P |
| raa’ | r | r | r | R | r |
| zayn | z | z | z | Z | z |
| syn | s | s | s | S | s |
| shyn | sh | s͡sh | sh | Sc | $ |
| Saad | S | s/S | S | Ss | S |
| Daad | D | d/D | D | Sd | D |
| Taa’ | T | t/T | T | Td | T |
| Zaa’ | Z | z/Z | Z | Dt | Z |
| ‘ayn | ‘ | @ | c | Ar | E |
| ghayn | gh | g͡gh | R | G | g |
| faa’ | f | f | f | F | f |
| qaaf | q | q | q | Q | q |
| kaaf | k | k | k | K | k |
| laam | l | l | l | L | l |
| mym | m | m | m | M | m |
| nuwn | n | n | n | N | n |
| haa | h | h | h | h | h |
| waaw | w | w/uu/oo | w | W | w |
| yaa’ | y | y/ii | y | Y | y |
| hamza | ’ | ’ | C | unclear | ’ |

Tabella 5: Vari schemi di traslitterazione per l'Arabo[6].

[6] Fonte (Kirchhoff et alii, 2002).

## 3. DIFFICOLTÀ NELLA CREAZIONE DEI SISTEMI AUTOMATICI DI RICONOSCIMENTO

Da queste brevi note introduttive appare evidente come le peculiarità di questa lingua possano tradursi in grosse difficoltà qualora si decida di creare un sistema di riconoscimento automatico (ASR[7]).

Le caratteristiche fonetiche sono modellabili, come per le lingue maggiormente studiate, attraverso modelli statistici complessi quali le *catene di Markov nascoste dipendenti dal contesto* (CDHMM[8]).

Il passaggio che rende veramente complicata la creazione di un ASR per l'Arabo è la raccolta dell'insieme di dati necessari per l'addestramento del sistema. Questi dati sono costituiti da file audio e da file di testo che riportano quanto più fedelmente possibile ciò che viene pronunciato. Il reperimento di questo materiale trascritto e, in particolare, della sua versione diacriticizzata è molto complicato e oneroso. Molto spesso è necessario accontentarsi di trascrizioni non complete oppure bisogna provvedere a trascrivere manualmente ex-novo i file audio con l'aiuto di operatori madrelingua.

Se l'obiettivo poi è creare un ASR in grado di riconoscere il parlato continuo, spontaneo e basato su un vocabolario abbastanza grande, bisogna tener conto della presenza nel parlato spontaneo arabo delle forme dialettali. Questa variabilità aumenta enormemente la complessità del problema.

Un'ulteriore complicazione deriva dalla creazione del modello del linguaggio (LM). Infatti, l'assenza di una trascrizione diacriticizzata, con la conseguente equivalenza ortografica di molte forme in contesti diversi, determina un aumento dell'incertezza con cui le parole vengono contestualizzate. Questo implica una sostanziale diminuzione dei potere discriminativi per i modelli del linguaggio. I LM non diacriticizzati saranno quindi molto meno predittivi di quelli diacriticizzati. Per quanto riguarda invece il caso di parlato spontaneo conversazionale si deve considerare la presenza, nel materiale da riconoscere, di molte parole non inserite nel vocabolario (OOV - Out Of Vocabulary).

| Tipo di ASR | Nome ASR | Performance (WER) |
|---|---|---|
| Dettatura | IBM ViaVoice for Arabic | - |
| Broadcast | BBN Arabic Broadcast News Recognizer | 15.3% |
| Broadcast | Arabic GALE transcription system (Soltau et alii, 2007) | 18.3% |
| Parlato Conversazionale Dialettale | 1996/1997 NIST CallHome Evaluations (Zavagliakos et alii, 1997) | 55.8% |

Tabella 6: Alcuni dati sulla qualità del riconoscimento per la lingua Araba presenti in letteratura (Kirchhoff et alii, 2002).

[7] ASR: Automatic Speech Recognition.
[8] CDHMM: Context Dependent Hidden Markov Model.

Le difficoltà di creazione dei modelli acustici e dei LM si traduce perciò in una perdita di accuratezza del ASR e, a parità di condizioni, in un livello di errore sulla parola maggiore che in altre lingue.

*Stato dell'arte negli ASR per l'Arabo*

In letteratura si trovano numerosi studi su ASR che si focalizzano, in virtù di quanto detto in precedenza, sulla lingua MSA, tralasciando le componenti dialettali. Molte applicazioni derivano dallo sviluppo di sistemi per la dettatura automatica, come, ad esempio, il sistema IBM ViaVoice.

Un breve panorama dello stato dell'arte nel campo del riconoscimento automatico della lingua Araba è illustrato in Tabella 6.

## 4. IL SISTEMA DI ASR

Lo scopo di questo lavoro è quello di analizzare le problematiche e descrivere le soluzioni che permettono lo sviluppo di un sistema di riconoscimento per la lingua Araba.

Si è proceduto quindi ad analizzare ed eseguire tutti i passaggi concernenti la realizzazione del progetto.

Sono state individuate tre fasi:

- *pre-trattamento del testo*: trasformazione delle trascrizioni, secondo i criteri precedentemente enunciati, per romanizzare e fonetizzare i lemmi in lingua Araba;
- *sviluppo* di un'architettura per l'addestramento ed il test di un sistema ASR, che permetta di sfruttare due motori di riconoscimento (SONIC[9] e SPHINX[10]);
- *valutazione* dei risultati ottenuti.

*Corpus*

Come punto di partenza per lo sviluppo del nostro progetto, abbiamo scelto la lingua Araba Standard, il MSA. Il relativo materiale vocale è stato ricavato dal corpus, denominato "*West Point Arabic Speech Corpus*" (*WPA*), distribuito dal LDC (Linguistic Data Consortium[11]).

Il corpus è composto da 8516 file audio, per un totale di circa 12 ore di parlato. Il segnale audio è stato registrato a 22,05 kHz e poi convertito in formato *NIST sphere*[12]. È costituito da parlato controllato, microfonico, letto in ambiente poco rumoroso. Le frasi e i lemmi pronunciati sono limitati (258 frasi e 1130 lemmi), ma l'audio è interamente e dettagliatamente trascritto. Le trascrizioni sono romanizzate con il sistema ArabTex e sono, inoltre, interamente diacriticizzate.

Con il corpus è fornito anche un dizionario con una trascrizione fonetica dei lemmi molto precisa.

I parlatori sono principalmente di madrelingua araba, ma è presente anche una piccola porzione di non madrelingua:

- 7200 delle registrazioni sono state pronunciate da madrelingua, per un totale di 10,5 ore
- 1200 file sono stati prodotti da parlatori non madrelingua, per un totale di poco più di 1 ora.

---

[9] http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html
[10] http://cmusphinx.sourceforge.net
[11] http://www.ldc.upenn.edu/
[12] http://www.nist.gov

*Preparazione del corpus*

Il materiale audio è stato adeguato agli standard richiesti per la realizzazione del modello acustico. Sono stati, prima di tutto, convertiti dal formato NIST sphere al normale PCM a 16 bit e successivamente sono stati sotto campionati da 22.05 kHz a 8kHz. Da segnalare che sono stati eliminati i file audio in cui le frasi da pronunciare sono state prodotte in modo incompleto o con errori macroscopici di riproduzione.

Le trascrizioni sono state convertite dalla codifica ArabTex alla romanizzazione Buckwalter per poter ottenere una corrispondenza biunivoca tra grafema arabo e simbolo romanizzato, peculiarità che l'originale sistema di trascrizione non garantisce.

Da queste trascrizioni, inoltre, sono state estratte le singole parole, in modo da creare un dizionario di lemmi ammissibili che costituisce l'insieme di parole che il sistema ASR può riconoscere. Per ogni parola del dizionario, è stata fornita una descrizione fonemica della pronuncia, ricavata da quella originale presente nel corpus.

Infine si è provveduto ad isolare singoli eventi non testuali, detti *filler,* che servono a descrivere le parti di audio che non possono essere catalogate come parlato. Un esempio di eventi extratestuali sono eventi come: risata, rumore, musica o il semplice silenzio.

*Architettura del sistema di riconoscimento.*

Sono stati utilizzati due sistemi di riconoscimento automatico del parlato: SONIC ver. 2 beta 3 e CMU SPHINX 3 ver. 0.7. Sono stati scelti questi due sistemi perché sono alcuni tra i software che meglio rappresentano lo stato dell'arte nel campo del riconoscimento automatico del parlato, tra quelli che permettono il libero utilizzo per scopi di ricerca.

I due sistemi si basano entrambi sul riconoscimento di sequenze di fonemi tramite l'utilizzo di un Modello Acustico statistico (AM) bastato su CDHMM (Context Dependent Hidden Markov Models) e di un Modello del Linguaggio a trigrammi (LM), basato su normali catene di Markov.

Il risultato del riconoscimento è la successione di parole che ha la probabilità più alta di essere stata pronunciata. Questa si ricava calcolando, tramite l'algoritmo di Viterbi, quale sia il percorso con punteggio più alto tra tutti quelli possibili, all'interno di una struttura detta *lattice* (reticolo o traliccio). Questa contiene tutti le possibili successioni fonetiche e, quindi, di parole che, con probabilità superiore ad un certo valore di soglia, possono essere state pronunciate nel file audio analizzato. Essa viene creata attraverso i valori risultanti dalla combinazione delle probabilità provenienti dal AM e dal LM.

Nonostante le numerose caratteristiche comuni, i due sistemi differiscono parecchio (Pellom & Hacioğlu, 2003; Chan et alii, 2007). Le principali differenze sono:

- *Estrazione delle features*. Le features sono dei vettori di coefficienti che identificano l'andamento dell'inviluppo spettrale del segnale audio da riconoscere. SONIC utilizza un metodo di calcolo di questi vettori più sofisticato rispetto a SPHINX. Quest'ultimo, infatti, utilizza i comuni MFCC (Mel Frequency Cepstrum Coefficients), mentre il primo adotta i coefficienti PMVDR (Perceptual Minimum Variance Distortionless Response). Il calcolo di questo tipo di vettori prevede un prefiltraggio del segnale per meglio modellare la curva percettiva dell'orecchio umano rispetto al sistema basato sulla scala Mel.
- *Adattamento del riconoscimento*. Nei processi di ASR, dopo un primo passaggio di riconoscimento, detto *baseline*, si utilizzano le informazioni così ricavate per reiterare il processo applicando delle tecniche di adattamento del modello statistico al segnale audio in esame. Ci sono varie strategie che possono essere applicate. In

SONIC sono implementate il MLLR[13], il VTLN[14] e SAT[15], mentre SPHINX permette solamente di utilizzare il metodo MLLR e la stima dei parametri MAP[16]. Questo si traduce, dopo l'adattamento, in un miglioramento delle prestazioni di SONIC nei confronti di quelle di SPHINX anche quando le percentuali della baseline sono sfavorevoli.

Il processo di riconoscimento in un ASR è composto di varie fasi, illustrate in Figura 2:

1. il file audio da riconoscere viene preelaborato per renderlo compatibile con il modello statistico creato, cioè PCM a 16 bit, campionato a 8kHz;
2. da questo file trasformato vengono estratte i vettori di features (MFCC o PMVDR) che saranno la base per il riconoscimento vero e proprio;
3. con le informazioni statistiche ricavate dal modello acustico e dal modello del linguaggio si cerca di catalogare i vettori provenienti dal passaggio precedente per associarli ai fonemi della lingua. Dalla successione di tali fonemi tramite il dizionario dato, si risalirà alle parole pronunciate;
4. Il risultato di questo processo è la sequenza di parole che con maggior probabilità, secondo i modelli dati, corrisponde all'audio in ingresso. Normalmente l'output è composto di una sola ipotesi; all'occorrenza, è possibile estrarre anche le N migliori ipotesi (*N-best*) o addirittura l'intero *lattice* di riconoscimento. Quest'ultimo può essere usato come ingresso per ulteriori passaggi basati su altre strategie di riconoscimento più ad alto livello (ad esempio un riconoscimento semantico, o le *confusion networks,* (Bertoldi & Federico, 2005).

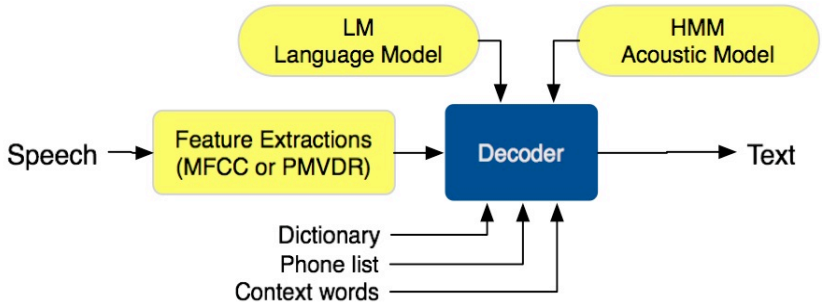


Figura 2: schema del sistema di riconoscimento di un file audio, comune ai software SONIC e SPHINX.

*Architettura del sistema di addestramento*

Per poter utilizzare la struttura ASR precedentemente illustrata, è necessario creare i modelli acustici e linguistici adatti. Questa è la parte più delicata di tutto il processo e anche quella che permette più margini di sviluppo.

I passaggi necessari per la creazione del modello acustico sono illustrati in Figura 3. L'input del sistema è costituito, come detto in precedenza, da un insieme di audio

[13] MLLR: Maximum Likelihood Linear Regression, adattamento senza supervisione.
[14] VTLN: Vocal Tract Length Normalization, normalizzazione rispetto alla lunghezza del tratto vocale del parlatore.
[15] SAT: Speaker Adaptive Training, adattamento ad un parlatore specifico
[16] MAP: Maximum A Posterior, adattamento con supervisione.

omogeneamente organizzato e dalle relative trascrizioni. Normalmente, è necessaria una ingente mole di dati, oltre le 10 ore di parlato, per ottenere un modello abbastanza stabile.

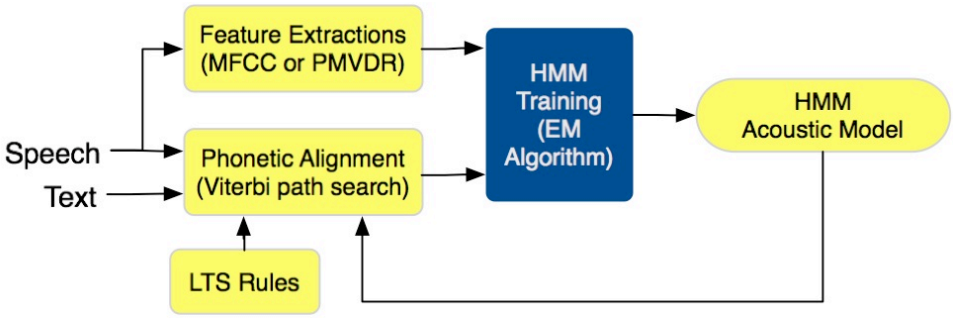


Figura 3: schema del sistema di addestramento comune ai software SONIC e SPHINX.

L'audio e le trascrizioni costituiscono quindi l'input del primo passaggio dell'addestramento: il cosiddetto *allineamento*. In questa fase il sistema, sulla base delle informazioni provenienti dalle trascrizioni, unite con quelle provenienti dal dizionario fonetizzato, cerca di far corrispondere, minimizzando la probabilità di errore, ogni segmento di file audio ad un fonema.

Anche in questo caso vi è una differenza tra SONIC e SPHINX. Il primo crea questa segmentazione sulla base anche di informazioni ricavate dal modello acustico di un'altra lingua (di solito l'inglese) e di corrispondenze tra i fonemi delle due lingue, che vengono decise manualmente in fase di preparazione della procedura di addestramento. Utilizzando il secondo sistema, invece, si parte da una segmentazione uniforme del file audio di ingresso e con successivi passaggi di allineamento basati sui modelli acustici così ottenuti, viene perfezionata la segmentazione.

Si è potuto constatare che, nel nostro caso, le segmentazioni risultanti sono grossomodo equivalenti.

Per quanto riguarda la creazione del modello del linguaggio è stato utilizzato un *toolkit* sviluppato presso la Carnegie Mellon University, il CMU-Cambridge Statistical Language Modeling toolkit[17]. Con esso è possibile estrarre, da grandi moli di testo scritto, le informazioni statistiche che permettono di creare la struttura a trigrammi, bigrammi e unigrammi che il sistema ASR utilizzerà per il riconoscimento. Nel nostro caso abbiamo utilizzato per creare il modello del linguaggio l'insieme delle trascrizioni di addestramento.

Questo crea un LM abbastanza polarizzato verso il tipo di conversazioni che dovranno essere riconosciute in fase di test.

## 5. RISULTATI SPERIMENTALI

Le performance del sistema con queste specifiche sono molto elevate, si ha infatti meno del 2% di errore di riconoscimento sulla singola parola (WER) e sono in linea con i risultati ottenuti in altri laboratori. In (Alotaibi et alii, 2007) le prestazioni di un sistema di riconoscimento, bastato sul sistema HTK (Young et alii, 2005) e addestrato su parlatori madrelingua del WPA, che processa file provenienti da madrelingua, utilizzando un

[17] http://www.speech.cs.cmu.edu/SLM_info.html

modello del linguaggio a bigrammi, raggiunge una correttezza del 99,05%. Mentre lo stesso sistema allenato anche con parlatori non madrelingua riconosce l'audio dei madrelingua con una correttezza del 93,98%.

I nostri test sono stati svolti utilizzando diverse configurazioni:

- con entrambi i sistemi di riconoscimento;
- con l'addestramento fatto su tipologie diverse di parlatore:
  - solo parlatori di madrelingua araba
  - sia parlatori madrelingua che non madrelingua;
- modificando i vari parametri di regolazione dei sistemi: ampiezza del *lattice*, peso del LM, etc;

I test sono stati effettuati sullo stesso insieme di file di test: circa mezz'ora di parlato continuo controllato, pronunciato da 5 parlatori diversi di madrelingua araba, proveniente dal WPA, ma non utilizzato per l'addestramento dei modelli.

I risultati ottenuti sono stati tutti soddisfacenti e non si differenziano tra loro in modo eclatante. Per chiarezza, illustriamo i due migliori risultati ottenuti per i due diversi software di riconoscimento denominati test A e test B.

La Figura 4 e la Tabella 7 mostrano i risultati finali del riconoscimento.

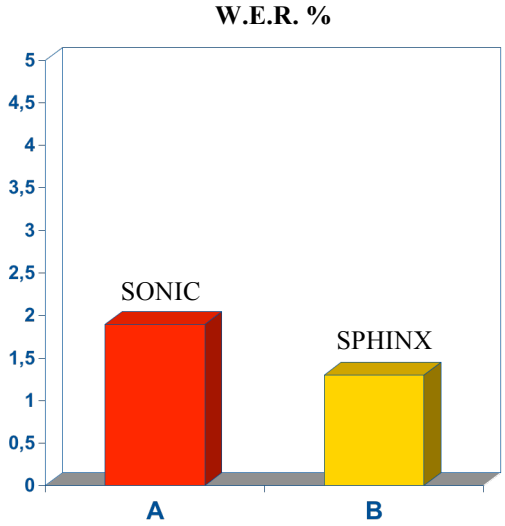


Figura 4: Grafico di confronto tra la percentuale di Word Error Rate (WER) per il sistema di riconoscimento basato su SONIC (A) e su SPHINX (B)

| | **Correttezza** | **Sostituzione** | **Cancellazione** | **Inserimento** | **WER** |
|---|---|---|---|---|---|
| **A** | 98.4 % | 0.9 % | 0.7 % | 0.3 % | **1.9 %** |
| **B** | 98.9 % | 0.5 % | 0.6 % | 0.2 % | **1.3 %** |

Tabella 7: Percentuali di riconoscimento sull'insieme di file audio di test.

Il test A è stato svolto facendo riconoscere i file di test dal sistema ASR SONIC allenato con un insieme di file audio di soli parlatori madrelingua. Per il test B, invece è stato utilizzato il riconoscitore SPHINX addestrato con tutto il materiale del corpus WPA meno i file di test.

Le elevate prestazioni di questi test di riconoscimento sono spiegate facilmente attraverso l'analisi del problema che stiamo affrontando. Si tratta infatti di:

- un riconoscimento di parlato controllato, microfonico e non rumoroso, quindi contenente pochi elementi difficilmente catalogabili dal AM;
- il vocabolario è interamente noto e limitato (1130 lemmi), quindi l'ambito semantico è molto focalizzato e l'incertezza sui vocaboli è molto bassa;
- il modello del linguaggio è anch'esso molto semplice e al suo interno presenta molti schemi che si trovano anche nelle conversazioni di test.

Un elemento interessante è che il sistema non viene per nulla peggiorato se si introducono, in fase di addestramento, anche dei "disturbi" derivanti da pronunce non perfette di parlatori non madrelingua. Questo indica che il modello del linguaggio, quando è molto focalizzato, è in grado di supplire alle carenze di modellazione acustica dei fonemi.

## 6. CONCLUSIONI E SVILUPPI FUTURI

Per quanto riguarda i possibili sviluppi futuri, è nostra intenzione considerare l'estensione di quanto sperimentato in questo contesto semplificato in un ambito più complesso come il parlato arabo spontaneo e conversazionale.

Il riconoscimento automatico in queste condizioni rappresenta un *task* molto più difficile. Il vocabolario utilizzato è molto più ampio, di solito è nell'ordine delle 40.000 parole. La costruzione della frase non è sempre regolare, sono presenti false partenze, interruzioni e ripetizioni. La qualità del segnale audio di solito è peggiorata dalla presenza di disturbi quali il rumore di fondo, altre voci o espressioni extra testuali.

Altro problema poi sta nel reperimento del materiale per creare l'allenamento dei modelli statistici. Generalmente questi corpora sono forniti esclusivamente con trascrizioni non diacriticizzate e non contengono informazioni sulle vocali brevi. Di conseguenza in una prima fase, per l'addestramento del sistema, sarà necessario basarci soltanto sull'informazione ortografica non vocalizzata, su cui potremo applicare eventualmente semplici regole di sostituzione grafema-fonema.

Nonostante l'oggettiva difficoltà di addestrare un sistema di riconoscimento sulla base delle sole trascrizioni ortografiche invece delle complete trascrizioni fonetiche, i risultati ottenuti in letteratura sono molto promettenti e consentono ampi margini di miglioramento soprattutto considerando il fatto che qualora si riesca a disporre della diacriticizzazione o della trascrizione fonetica precisa di una buona parte del corpus l'addestramento del sistema risulterà sicuramente più preciso ed affidabile.

## RINGRAZIAMENTI

Parte di questo lavoro è stato possible grazie ad una collaborazione con la ditta Expert System Spa[18] di Modena.

[18] www.expertsystem.it

**BIBLIOGRAFIA**


Alotaibi, Y. A., Selouani, S., O'Shaughnessy, D., (2007) *Experiments on Automatic Recognition of Nonnative Arabic Speech,* accepted for EURASIP Journal on Audio, Speech, and Music Processing, Hindawi Publishing Corporation

Chan, A., Gouvea, E., Singh, R., Ravishankar, M., Rosenfeld, R., Sun, Y., Huggins-Daines, D., Seltzer, M., (2007) *The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources,* March 11, 2007, *http://www.cs.cmu.edu/~archan/.*

Bertoldi, N., Federico, M., (2005) *A new decoder for spoken language translation based on confusion networks,* Automatic Speech Recognition and Understanding, IEEE Workshop, 27 Nov-1 Dec, 86–91.

Buckwalter, *http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/buckwalter-about.html*; *http://www.qamus.org/transliteration.htm.*

Kirchhoff, K., Bilmes, J., Henderson, J., Schwartz, R., Noamany, M., Schone, P., Ji, G., Das, S., Egan, M., He, F., Vergyri, D., Liu, D., Duta, N., (2002) *Novel Speech Recognition Models for Arabic,* Final Report, Johns-Hopkins University Summer Research Workshop, *http://www.clsp.jhu.edu/ws02.*

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., Stolcke, A., (2004), *Morphology-based language modeling for conversational Arabic speech recogniton*, INTERSPEECH - ICSLP, International Conference on Spoken Language Processing, Korea, October 4-8, 2245-2248.

Pellom, B., Hacıoğlu, K. (2003) *Sonic: The University of Colorado Continuous Speech Recognizer - Technical Report TR-CSLR-2001-01,* Center for Spoken Language Research University of Colorado, Boulder.

Soltau, H., Saon, G., Kingsbury, B., Kuo, J.; Mangu, L., Povey, D., Zweig, G., (2007), The IBM 2006 Gale Arabic ASR System, *Acoustics, Speech and Signal Processing 2007. ICASSP 2007*, Volume 4, April 15-20, 349 - 352.

Young, S., Evermann, G., Gales, M. et alii, (2005) *The HTK Book (for HTK Version. 3.3)*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/

Zavaliagkos, G., McDonough, J., Miller, D., El-Jaroudi, A., Billa, J., Richardson, F., Ma, K., Siu, M., Gish, H. (1998), The BBN Byblos 1997 large vocabulary conversational speech recognition system. *Acoustics, Speech and Signal Processing 1997. Proceedings of the IEEE International Conference*, Volume 2, May 12-15, 905-908.