ESPERIMENTI DI IDENTIFICAZIONE DELLA LINGUA PARLATA IN AMBITO GIORNALISTICO

Diego Giuliani, Roberto Gretter
Human Language Technology Research Unit – FBK – Fondazione Bruno Kessler
Via Sommarive 18, 38123, Povo (TN), Italy
giuliani@fbk.eu gretter@fbk.eu

1. SOMMARIO

Nell'ambito del riconoscimento automatico della voce, usualmente si assume di conoscere la lingua in cui un dato canale, ad esempio televisivo, trasmette i suoi telegiornali. In effetti questa assunzione viene spesso disattesa in canali internazionali (da questo punto di vista i canali italiani sono un'eccezione), dove gran parte delle interviste a persone straniere iniziano con diversi secondi dell'audio originale, che poi cala di volume quando subentra la traduzione nella lingua di riferimento. In alcuni casi, le interviste in lingua straniera vengono trasmesse direttamente con l'audio originale al quale vengono aggiunti dei sottotitoli per consentirne la comprensione. Applicare un riconoscitore automatico nella sola lingua di riferimento al flusso audio provoca quindi, inevitabilmente, lo sgradevolissimo effetto di introdurre una sequenza di errori ogniqualvolta compare del parlato in una lingua diversa.

Sorge quindi la necessità di far precedere il processo di riconoscimento automatico da un modulo di identificazione del linguaggio, capace di elaborare il flusso audio, dividerlo in segmenti di parlato ed associare ad ogni segmento di parlato la lingua identificata. Segmenti non appartenenti al linguaggio di riferimento possono quindi essere ignorati oppure essere elaborati da un riconoscitore appropriato.

L'esigenza di dover elaborare lingue per le quali non sono disponibili risorse linguistiche in quantità significativa (i cosiddetti "under resourced languages") costringe ad acquisire risorse linguistiche a basso costo, come ad esempio dati testuali raccolti via web per costruire modelli del linguaggio aggiornati e capaci di seguire giorno per giorno l'evolversi delle varie lingue, oppure registrazioni di programmi televisivi in diverse lingue. Questo tipo di dati non vengono usualmente annotati manualmente, data la difficoltà di reperire esperti nelle varie lingue ed i costi necessari ad annotare grandi basi di dati. Si preferisce quindi annotare questi dati automaticamente, e quando possibile scartare i dati che per qualche motivo risultassero poco affidabili.

In questo lavoro è stato sperimentato l'utilizzo di dati annotati in maniera completamente automatica per addestrare due diversi sistemi di identificazione della lingua parlata. Il primo sistema, puramente acustico, è basato su una mistura di Gaussiane (Gaussian Mixture Model, GMM) mentre il secondo implementa un vero e proprio riconoscitore vocale multilingua (Automatic Speech Recognition, ASR). L'unico intervento manuale in questo lavoro sta nella predisposizione di un trascrittore fonetico per ognuna delle lingue considerate.

In questo lavoro vengono considerate 6 lingue: italiano, turco, spagnolo, francese, tedesco, russo. Le prestazioni sono state valutate su tre insiemi di segmenti omogenei per durata, dove ogni segmento contiene parlato in una sola lingua. I risultati ottenuti con le due tecniche sui tre insiemi, in termini di accuratezza, sono compresi tra 87.4% e 91.0% per l'approccio GMM e tra 90.6% e tra 94.1% per l'approccio ASR.

2. I DUE APPROCCI

I sistemi per l'identificazione automatica della lingua parlata hanno lo scopo di identificare in maniera accurata e rapida la lingua parlata da un parlante sconosciuto. Gli elementi che maggiormente differenziano i diversi sistemi proposti in letteratura sono il tipo di parametri acustici utilizzati per modellare il parlato e le caratteristiche del classificatore utilizzato. In relazione a quest'ultimo aspetto, gli approcci comunemente adottati sono basati sull'utilizzo di misture di Gaussiane (note come "Gaussian Mixture Models", o GMMs), reti neurali artificiali, modelli di Markov nascosti (HMM), Support Vector Machine, etc. (Ambikairajah et alii, 2011).

Spesso i sistemi LID si compongono di più sottosistemi che utilizzano uno a o più fonti di informazione per stimare una qualche misura di similarità rispetto alle varie lingue: le misure di similarità stimate dai vari sotto-sistemi sono quindi combinate/fuse tra loro per prendere la decisione finale (Ambikairajah et alii, 2011). Le fonti di informazione utilizzate per la classificazione sono tipicamente: puramente acustiche (per modellare le caratteristiche acustiche delle varie lingue), fonotattiche (per modellare come i fonemi presenti in una data lingua si possono susseguire nel parlato), prosodiche (che afferiscono alla variazione del pitch), morfologiche (che afferiscono alla struttura interna della parola) e sintattiche (per modellare come le parola di una data lingua si possono susseguire le une alle altre).

2.1. L'approccio GMM

L'approccio più popolare per l'identificazione del linguaggio parlato è quello basato sull'utilizzo di una mistura di Gaussiane (nota come "Gaussian Mixture Model", GMM) per modellare le proprietà acustiche di una data lingua (Ambikairajah et alii, 2011). In questo caso, per ogni lingua da riconoscere viene addestrato uno specifico GMM sui dati disponibili per quella lingua, usando l'algoritmo di aspettazione – massimizzazione (Expectation Maximization, EM) (Zissman, 1996). Questo approccio non richiede la trascrizione od annotazione dei dati acustici ma la sola conoscenza che i dati acustici siano istanze di una certa lingua.

La variante implementata per questo lavoro consiste nell'addestramento iniziale di un unico GMM utilizzando i dati di tutte le lingue: questo GMM viene chiamato modello universale di background (Universal Background Model, UBM). I GMM delle singole lingue vengono ottenuti adattando l'UBM con i soli dati di ogni lingua (Reynolds et alii, 2000, Wong and Sridharan, 2002).

In fase di test, dato un segmento di parlato, i valori di verosimiglianza che rappresentano la plausibilità che il segmento di parlato sia stato generato da ciascuno dei GMM sono confrontati tra loro, e viene dichiarato vincente il modello corrispondente al valore di massima verosimiglianza.

2.2. L'approccio ASR

Il secondo approccio che è stato implementato fa uso di un vero e proprio sistema di riconoscimento (nel seguito verrà chiamato ASR), che considera anche informazioni lessicali e sintattiche (N-grammi). Sui dati audio disponibili viene addestrato un insieme di modelli acustici multilingua (ad esempio tutte le lingue considerate condividono il modello acustico del fonema /a/), mentre i dati linguistici vengono utilizzati per addestrare un modello del linguaggio statistico anch'esso multilingua. Per implementare il sistema di riconoscimento

sono stati impiegati alcuni moduli sviluppati in FBK negli scorsi anni per ognuna delle lingue considerate:

- convertitori grafema - fonema a regole per ottenere la trascrizione fonetica delle parole;
- procedure di espansione dei numeri per elaborare e normalizzare i corpora di testo da usare per i modelli del linguaggio.

I dati audio utilizzati per addestrare i modelli acustici provengono da materiale bilanciato nelle 6 lingue, trascritto in maniera completamente automatica, senza alcun tipo di supervisione, utilizzando sistemi di riconoscimento sviluppati in FBK negli scorsi anni per le varie lingue (Falavigna & Gretter, 2011, Bisazza & Gretter, 2012).

Solo le informazioni relative ai fonemi riconosciuti sono state utilizzate per l'addestramento acustico. Tabella 1 contiene la lista dei fonemi utilizzati, ognuno con le lingue che lo utilizzano.

1	TR RU	**T**	ES	**jj**	ES TR	**t`**	RU
2	FR TR	**U**	DE	**k**	IT DE FR ES TR RU	**tt**	IT TR
2:	DE	**Y**	DE	**k`**	RU	**u**	IT FR ES TR RU
6	DE	**Z**	DE FR TR RU	**kk**	IT TR	**u:**	DE
9	DE FR	**a**	IT DE FR ES TR RU	**l**	IT DE FR ES TR RU	**v**	IT DE FR TR RU
9~	FR	**a:**	DE	**l`**	RU	**v`**	RU
@	DE FR	**a~**	FR	**ll**	IT TR	**vv**	IT TR
A	FR	**b**	IT DE FR ES TR RU	**m**	IT DE FR ES TR RU	**w**	IT FR ES
B	ES	**b`**	RU	**m`**	RU	**x**	DE ES RU
C	DE	**bb**	IT TR	**mm**	IT TR	**x`**	RU
D	ES	**d**	IT DE FR ES TR RU	**n**	IT DE FR ES TR RU	**y**	FR TR
E	IT DE FR	**d`**	RU	**n`**	RU	**y:**	DE TR
E:	DE	**dd**	IT TR	**nn**	IT TR	**z**	IT DE FR ES TR RU
G	ES TR	**e**	IT FR ES TR RU	**o**	IT FR ES TR RU	**z`**	RU
H	FR	**e:**	DE	**o:**	DE	**OY**	DE
I	DE	**e~**	FR	**o~**	FR	**aI**	DE
J	IT FR ES	**f**	IT DE FR ES TR RU	**p**	IT DE FR ES TR RU	**aU**	DE
JJ	IT	**f`**	RU	**p`**	RU	**dZ**	IT DE TR
L	IT ES	**ff**	IT TR	**pp**	IT TR	**ddZ**	IT TR
LL	IT	**g**	IT DE FR ES TR RU	**r**	IT DE ES TR TR RU	**dz**	IT
N	DE FR ES	**g`**	RU	**r`**	RU	**ddz**	IT
O	IT DE FR	**gg**	IT TR	**rr**	IT ES TR	**tS**	IT DE ES TR RU
R	DE FR	**h**	DE TR	**s**	IT DE FR ES TR RU	**ttS**	IT TR
S	IT DE FR TR RU	**i**	IT FR ES TR RU	**s`**	RU	**ts**	IT DE RU
S`	RU	**i:**	DE	**ss**	IT TR	**tts**	IT
SS	IT	**j**	IT DE ES FR TR RU	**t**	IT DE FR ES TR RU	**pf**	DE

Tabella 1: I 104 fonemi usati nell'approccio ASR, ognuno con le lingue di appartenenza (46 DE, 31 ES, 37 FR, 50 IT, 42 RU, 47 TR).

I dati di testo provengono da materiale scaricato in rete da giornali on-line nelle lingue coinvolte. Tale materiale è stato ripulito in maniera automatica ed elaborato in modo da ef-

fettuare alcune normalizzazioni ed espansioni (tipicamente i numeri, ad esempio 123 diventa _cento_ _venti_ _tre_). Quindi ogni parola è stata fatta precedere da un'etichetta che individua la lingua di origine (ad es.: it:città, de:nicht, fr:attaque).

Una volta ottenuto il corpus multilingua per concatenazione dei singoli corpus monolingua, sono state utilizzate le normali procedure per costruire un modello del linguaggio a N-grammi. A parte i link di back-off che permettono ad ogni parola di essere riconosciuta in qualsiasi posizione, gli N-grammi realmente osservati in fase di addestramento sono di fatto interni alle singole lingue.

Un riconoscitore che utilizzi tali modelli potrà in effetti emettere una sequenza di parole in lingue diverse. Negli esperimenti effettuati, dovendo emettere un'unica etichetta per ogni segmento dato, a valle del riconoscitore viene applicato un filtro a maggioranza che assegna l'etichetta alla lingua le cui parole hanno durata complessiva maggiore.

3. CONTESTO APPLICATIVO E ACQUISIZIONE DATI

Negli ultimi anni sta emergendo la tendenza ad acquisire risorse linguistiche a basso costo, come ad esempio dati testuali raccolti via web per costruire modelli del linguaggio aggiornati e capaci di seguire giorno per giorno l'evolversi delle varie lingue. Come dati audio sono accessibili diverse fonti: web, canali radio o televisivi. L'audio raccolto tramite alcuni di questi canali può essere utilizzato per addestrare modelli acustici in una nuova lingua con procedure completamente non supervisionate. Viene ad esempio effettuato un primo riconoscimento con modelli acustici derivati da altre lingue, e dall'allineamento risultante è possibile addestrare dei modelli acustici imperfetti che, per passi successivi, possono essere raffinati fino ad ottenere prestazioni ragionevoli. Utilizzando questa procedura, negli anni scorsi abbiamo costruito dei riconoscitori in diverse lingue, ottenendo come sottoprodotto del materiale audio etichettato in maniera non supervisionata, con un'accuratezza di parola che, a seconda della lingua, varia tra il 50% ed il 90% lingue (Falavigna & Gretter, 2011, Bisazza & Gretter, 2012). Tale materiale è stato utilizzato per creare dei corpora in diverse lingue, omogenei per tipologia di contenuto e dimensione, poi utilizzati per addestrare i sistemi descritti nella precedente sezione.

Esistono diversi canali TV che trasmettono notizie internazionali in diverse lingue, fornendo un flusso ininterrotto di dati vocali potenzialmente attraenti come fonte di dati vocali paralleli. In uno di questi canali, Euronews, che trasmette via satellite, le notizie vengono trasmesse in oltre 10 lingue diverse (il numero di lingue cambia nel tempo): inglese, francese, tedesco, italiano, spagnolo, portoghese, polacco, greco, ungherese, ucraino, russo, turco, arabo e persiano.

Ogni lingua viene trasmessa su un canale audio differente nello stesso flusso digitale, e poiché il contenuto video è lo stesso per tutti i canali audio, ogni singola notizia può essere considerata temporalmente allineata nelle varie lingue. In sostanza, Euronews trasmette uno schema ciclico che si ripete ogni 30 minuti ed è composto da: notizie principali della giornata (politica, attualità); musica e spot pubblicitari, servizi specialistici (economia, tecnologia, storia, natura); musica e spot pubblicitari. Durante il giorno, ogni notizia ha un ciclo di vita: può essere ripetuta più volte, può essere espansa, accorciata od aggiornata.

Tali dati non sono realmente omogenei, nel senso che avvengono diversi fenomeni: in caso di interviste, ma non solo, vengono trasmessi alcuni secondi di registrazione nella lingua originale prima che inizi la traduzione; alcuni spot sono spesso in inglese, vi è la pre-

senza di musica; a volte una particolare notizia non è stata ancora tradotta in tutte le lingue, e in questo caso alcuni canali potrebbero contenere l'audio originale (in un'altra lingua).

Da maggio 2009 in FBK registriamo digitalmente ogni giorno un'ora di flusso audio: prima viene estratto l'elenco dei canali audio e poi, per ciascun lingua, la sua traccia audio che viene salvata a 16 kHz, insieme al codice della lingua (eng fra deu ita spa por pol gre hun ukr rus tur ara per).

Data la particolare struttura del flusso è abbastanza facile segmentare le diverse tracce audio in notizie, sfruttando il fatto che i confini tra le notizie sono caratterizzati da rumore di fondo (silenzio) su tutti i canali. Quindi, da un punto di vista pratico, un semplice approccio è quello di rilevare le pause separatamente su ogni canale, e considerare come confini probabili di notizia le pause che sono comuni a tutti i canali, come evidenziato in Figura 1.

Una nota per chi fosse interessato ad usare questi dati per traduzione automatica: l'allineamento delle notizie nelle varie lingue è reale, ma i testi in lingue diverse non sono l'esatta traduzione l'uno dell'altro. A volte la stessa notizia è affrontata da diversi punti di vista, a volte un linguaggio dà maggiori dettagli rispetto agli altri.

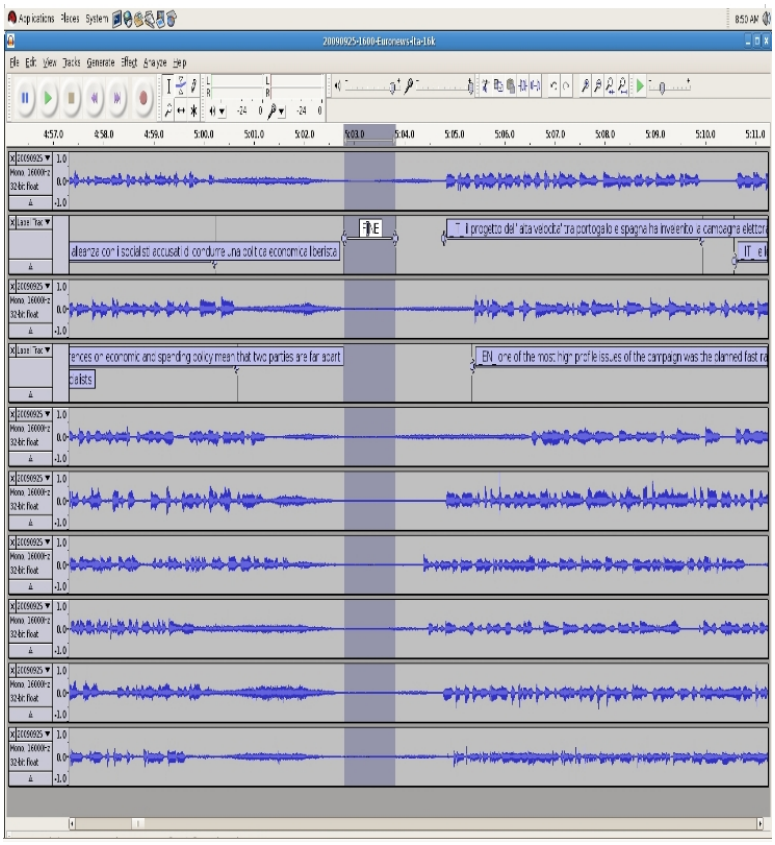

Figura 1: Evidenziazione del confine di notizia in alcuni canali audio di Euronews.

[S e l e z i o n a r e l a d a t a]

4. ESPERIMENTI

Per confrontare i due approcci proposti a parità di condizioni, abbiamo definito un insieme di dati che sono stati utilizzati per addestrare i rispettivi modelli e per testarne le capacità di classificazione. Naturalmente l'approccio ASR fa uso di tutti i dati (inclusi i corpora di testo) mentre l'approccio GMM usa solo il materiale audio, quindi è lecito aspettarsi che il primo fornisca prestazioni migliori in termini di classificazione, come in effetti accade. In questo lavoro considereremo 6 lingue: italiano, turco, spagnolo, francese, tedesco, russo, che fanno parte dell'insieme delle lingue trasmesse da Euronews.

4.1. Dati di addestramento

I nostri esperimenti hanno considerano data set abbastanza contenuti. Come materiale di addestramento abbiamo utilizzato, per ognuna delle lingue coinvolte:

- 3 ore di materiale audio, proveniente da Euronews. Notare che, come già accennato, tale materiale non è "puro" ma può contenere segmenti audio in lingue diverse da quella ipotizzata. Oltre al segnale audio è disponibile una segmentazione in fonemi, ottenuta in maniera automatica da un riconoscitore vocale nella lingua ipotizzata;
- 10 milioni di parole (running words) di testi scritti raccolti in ambito giornalistico, dal web; tale materiale è stato elaborato e normalizzato in maniera completamente automatica. Come accade per il materiale audio, anche quello testuale può contenere frasi provenienti da lingue diverse da quella ipotizzata;
- un lessico composto dalle 5000 parole più frequenti presenti nel corpus appena descritto, trascritto foneticamente utilizzando un sistema a regole nella lingua ipotizzata.

4.2. Dati di test

Dai test set predisposti negli scorsi anni per valutare le prestazioni dei riconoscitori vocali, ottenuti trascrivendo manualmente alcune ore di materiale di Euronews in diverse lingue, abbiamo estratto del materiale utilizzabile per valutare le prestazioni di sistemi di LID. Per valutare le prestazioni dei vari sistemi implementati abbiamo definito 3 insiemi di test, costituiti da segmenti audio. Ogni segmento audio contiene parlato in una sola tra le 6 lingue considerate, ed è caratterizzato da una durata prestabilita (ad esempio, tra 3 e 7 secondi), come indicato in Tabella 2.

identificativo	numero di segmenti	durata minima	durata massima
TF1_TT5	541	1 secondo	5 secondi
TF3_TT7	522	3 secondi	7 secondi
TF5_TT9	481	5 secondi	9 secondi

Tabella 2: Caratteristiche dei tre insiemi di test utilizzati.

4.3. Risultati

I due approcci sono stati implementati, addestrati sui dati descritti sopra, ed infine applicati ai diversi test set. Ricordiamo che ogni segmento è stato etichettato con una sola lingua e che entrambi i sistemi dovevano restituire una sola etichetta, quindi la scelta dei seg-

menti da elaborare era data a priori e nel caso ASR è stato applicato un filtro a maggioranza per decidere la lingua riconosciuta. L'etichetta "linguaggio non conosciuto" non è stata considerata in questo insieme di esperimenti. I risultati complessivi in termini di accuratezza, ottenuti con le due tecniche, sono riportati in Tabella 3, per ognuno dei test set definiti. L'approccio ASR ha fornito prestazioni migliori, come era lecito aspettarsi in quanto fa uso anche di informazioni lessicali e sintattiche mentre l'approccio GMM usa solo informazioni acustiche.

identificativo	GMM	ASR
TF1_TT5	87.4%	90.6%
TF3_TT7	91.0%	**94.1%**
TF5_TT9	90.0%	92.5%

Tabella 3: Percentuale di segmenti correttamente identificati coi due approcci nei tre test set.

Infine, Tabella 4 mostra la matrice di confusione ottenuta con l'approccio ASR sul test set TF3_TT7, corrispondente ad una percentuale di segmenti correttamente identificati pari a 94.06%. Si può notare come le due lingue meno identificate siano il turco ed il russo, lingue estremamente morfologiche per le quali il modello del linguaggio è più debole.

	tur	**deu**	**ita**	**esp**	**fra**	**rus**	
tur	**69**	**0**	**4**	**2**	**0**	**0**	**92.0%**
deu	**1**	**45**	**0**	**0**	**1**	**0**	**95.7%**
ita	**1**	**0**	**96**	**1**	**2**	**0**	**96.0%**
esp	**2**	**1**	**0**	**97**	**0**	**0**	**97.0%**
fra	**2**	**0**	**0**	**0**	**98**	**0**	**98.0%**
rus	**8**	**0**	**0**	**5**	**1**	**86**	**86.0%**

Tabella 4: Matrice di confusione ottenuta nel caso ASR TF3_TT7, (94.1%).

5. DISCUSSIONE E CONCLUSIONI

In questo lavoro sono stati messi a confronto due sistemi per l'identificazione automatica della lingua parlata. Il primo sistema, più semplice, è basato su GMMs mentre il secondo è un vero e proprio riconoscitore del parlato multilingua con una post-elaborazione sull'output generato. Gli esperimenti di identificazione delle lingua sono stati condotti su un insieme di dati che include segmenti di parlato in 6 lingue. Gli esperimenti hanno evidenziato la superiorità del sistema basato sul sistema di riconoscimento del parlato multilingua. Questo non è un risultato sorprendente in quanto quest'ultimo sistema è molto più complesso ed utilizza sia informazione acustica sia informazione linguistica. Il tasso di identificazione ottenuto varia dal 87% al 94%, in relazione alla lunghezza dei segmenti di parlato. Questo risultato è stato ottenuto addestrando il sistema di identificazione della lingua utilizzando dati acustici di training non trascritti manualmente. La trascrizione automatica automatica dei dati di training è stata ottenuta utilizzando dei sistemi di ASR già dispo-

nibili che per francese, russo, spagnolo e turco sono stati a loro volta addestrati in maniera non supervisionata.

RINGRAZIAMENTI

Questo lavoro è stato in parte finanziato dalla Comunità Europea nell'ambito del progetto del Settimo programma quadro EU-BRIDGE, contratto numero 287658.

BIBLIOGRAFIA

Ambikairajah, E., Li, H., Wang, L., Yin, B. & Sethu, V. (2011), Language identification: A tutorial, Circuits and Systems Magazine, IEEE, 11(2):82–108.

Bisazza, A. & Gretter, R. (2012), Building a Turkish ASR system with minimal resources, in Proceedings of LREC Workshop on Language Resources and Technologies for Turkic Languages, Istanbul, Turkey, May 21, 6-10.

Falavigna, D. & Gretter, R. (2011), Cheap bootstrap of multi-lingual hidden markov models, in Proceedings of Interspeech, August 28-31, pages 2325–2328.

Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000), Speaker verification using adapted gaussian mixture models, Digital Signal Processing, pages 19–41.

Wong, E. & Sridharan, S. (2002), Methods to improve gaussian mixture model based language identification system, in Proceedings of Int. Conf. Spoken Language Processing (ICSLP), Denver, Colorado, USA, September 16-20, 93–96.

Marc A. Zissman, M. A. (1996), Comparison of four approaches to automatic language identification of telephone speech, IEEE Transactions on Speech and Audio Processing, 4(1):31.