DOI: 10.17469/O2102AISV000017

MIRCO RAVANELLI, LUCA CRISTOFORETTI, ROBERTO GRETTER,
MARCO PELLIN, ALESSANDRO SOSI, MAURIZIO OMOLOGO

Il corpus DIRHA-ENGLISH ed i relativi task per il riconoscimento vocale a distanza in ambienti domestici

This paper addresses the contents and the possible usage of the DIRHA-ENGLISH multi-microphone corpus, realized under the EC DIRHA project. The reference scenario is a domestic environment equipped with a large number of microphones distributed in space.
The corpus is composed of both real and simulated material, and it includes 12 US and 12 UK English native speakers' utterances. Each speaker uttered different sets of phonetically-rich sentences, newspaper articles, conversational speech, keywords, and commands. From this material, a large set of 1-minute sequences was generated, which also includes typical domestic background noise and inter/intra-room reverberation effects. Development and test sets were derived.
The paper reports a first set of baseline results obtained using different techniques, including Deep Neural Networks (DNN), aligned with the state-of-the-art at international level. Various tasks and Kaldi recipes have already been developed.

Key words: distant speech recognition, microphone arrays, corpora, Kaldi, DNN.

Introduzione

Il riconoscimento vocale è stato oggetto di molta attenzione negli ultimi anni (Yu, Deng, 2015). Come risultato, ha trovato applicazione in vari campi, come i sistemi di dettatura, il controllo in automobile, la ricerca su web. Nonostante gli sforzi, molte soluzioni sono ancora basate su un'interazione closetalk, usando un microfono vicino al parlatore, e costringendo quindi l'utente ad avvicinarsi al dispositivo. Questa limitazione non è però accettabile nel caso in cui l'utente voglia un maggiore grado di libertà nell'utilizzo di queste tecnologie. È dunque facilmente prevedibile che il riconoscimento vocale a distanza (Wölfel, McDonough, 2009) rivestirà un ruolo di primaria importanza nello sviluppo delle future interfacce uomo-macchina. Un esempio applicativo particolarmente rilevante è l'ambiente domestico, che è stato oggetto del progetto europeo DIRHA. Lo scopo era quello di sviluppare dei servizi automatizzati in ambiente domestico, controllati da un sistema con riconoscimento vocale a distanza funzionante in più lingue.

Nonostante i progressi in questo campo, le tecnologie attuali mostrano ancora una mancanza di robustezza e flessibilità, dovuta alla presenza di rumori ambientali non stazionari e alla presenza di riverbero (Hänsler, Smith, 2008). Per colmare la differenza di prestazioni rispetto ad un sistema close-talk sono necessarie

notevoli quantità di dati, registrati in condizioni adatte e trascritti appositamente. Date le innumerevoli variabili in gioco nell'ambiente domestico, si tratta di un'attività altamente complicata ed impegnativa; diventa quindi indispensabile disporre di corpora multimicrofonici realistici e di alta qualità, finalizzati all'addestramento del riconoscitore. Nonostante la disponibilità di alcuni corpora, la necessità di materiale specifico ci ha portato alla creazione di vari corpora multimicrofonici registrati in ambiente domestico.

Il corpus multi-microfonico DIRHA-ENGLISH è stato realizzato in lingua inglese assieme ad altri corpora (raccolti in quattro lingue: italiano, greco, tedesco e portoghese) nell'ambito del progetto DIRHA (Cristoforetti, Ravanelli, Omologo, Sosi, Abad, Hagmüller & Maragos, 2014). Lo scenario di riferimento è un appartamento equipaggiato con un elevato numero di microfoni, distribuiti nelle varie stanze. La scelta della lingua inglese è dettata dal fatto che questa rappresenta la lingua di riferimento nella comunità internazionale. Il corpus è composto sia da materiale simulato sia da materiale reale registrato nell'appartamento, per permettere di testare le prestazioni del riconoscimento vocale in condizioni reali.

Lo scopo di questo articolo è di descrivere il contenuto del corpus DIRHA-ENGLISH e di fornire alcuni risultati preliminari su frasi foneticamente ricche, ottenuti utilizzando il framework Kaldi (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlicek, Quian, Schwarz, Silovsky, Stemmer & Vesely, 2011). Il risultante task di tipo TIMIT può essere visto come complementare a task tipo riconoscimento WSJ o di parlato conversazionale.

L'articolo è suddiviso nel seguente modo. La Sezione 1 descrive il progetto europeo DIRHA mentre la Sezione 2 si focalizza sul contenuto e le caratteristiche del corpus DIRHA-ENGLISH. La Sezione 3 riporta una descrizione dei task definiti ed i risultati preliminari corrispondenti. La Sezione 4 fornisce alcune conclusioni.

1. *Il progetto europeo DIRHA*

Il progetto europeo DIRHA, iniziato nel gennaio 2012 e durato tre anni, aveva come obiettivi l'analisi della scena acustica e l'interazione vocale a distanza in un ambiente domestico. Seguono ora una descrizione degli obiettivi, dei task ed dei corpora raccolti.

1.1 Obiettivi e task

Lo scenario applicativo affrontato nell'ambito del progetto è caratterizzato da un sistema vocale interattivo che permette l'interazione da qualsiasi stanza e senza vincoli di posizione. Sfruttando una rete di microfoni distribuiti in varie stanze, il sistema DIRHA reagisce prontamente ai comandi impartiti da un utente. Il sistema è sempre in ascolto, aspettando una specifica parola d'ordine per iniziare un nuovo dialogo con l'utente. Il dialogo permette poi all'utente di accedere a

dispositivi e servizi, tipo l'apertura di porte e finestre, accendere o spegnere luci, controllare la temperatura o ascoltare della musica. Il sistema è inoltre caratterizzato dalla possibilità di gestire più dialoghi in parallelo in stanze diverse e dalla possibilità di fare barge-in (cioè poter interagire anche in presenza di musica o prompt acustici emessi da parte del sistema). Un'altra caratteristica molto importante è la capacità di limitare i falsi allarmi, dovuti alla non corretta interpretazione di suoni ambientali o normali dialoghi tra utenti.

Partendo da queste funzionalità, vari task sperimentali sono stati definiti in combinazione con algoritmi di front-end processing e riconoscimento vocale nelle varie lingue. La maggior parte di questi task si riferiscono ad acquisizioni vocali effettuate nell'appartamento ITEA di Trento.

1.2 I corpora DIRHA

I corpora vocali DIRHA sono stati progettati per mettere a disposizione raccolte multi-microfoniche atte ad essere utilizzate in un ampio numero di task, come menzionato sopra. Alcune raccolte sono basate su simulazioni ottenute tramite contaminazione (Matassoni, Omologo, Giuliani & Svaizer, 2002; Couvreur, Couvreur & Ris, 2000; Haderlein, Nöth, Herbordt, Kellermann & Niemann, 2005), combinando registrazioni close-talk con risposte impulsive stimate e sequenze reali di rumore di fondo (Cristoforetti et al., 2014). Altre raccolte sono invece state registrate in condizioni reali.

A parte il corpus DIRHA-ENGLISH che verrà descritto nella prossima sottosezione, gli altri corpora raccolti sono i seguenti:

- Il corpus DIRHA Sim (30 parlatori per quattro lingue) (Cristoforetti et al., 2014), che consiste in sequenze multi-canale della durata di un minuto, comprendenti vari eventi acustici e frasi;
- Una raccolta basata sul Mago di OZ (WOZ) (Brutti, Ravanelli, Svaizer & Omologo, 2014) per valutare le componenti di speech-activity-detection e di localizzazione del parlatore;
- Il corpus DIRHA AEC (Zwyssig, Ravanelli, Svaizer & Omologo, 2015) che include dati specificatamente raccolti per studiare la cancellazione dell'eco, per rimuovere interferenze acustiche note, diffuse nell'ambiente;
- Il corpus DIRHA-GRID (Matassoni, Astudillo, Katsamanis & Ravanelli, 2014) che include una raccolta multi-canale e multi-stanza di dati simulati, derivanti dalla contaminazione del corpus GRID (Cooke, Barker, Cunninghan & Shao, 2006), composto da brevi comandi in lingua inglese.

2. *Il corpus DIRHA-ENGLISH*

Come per gli altri corpora, anche il corpus DIRHA-ENGLISH è composto da una parte di dati reali ed una parte di dati simulati, questi ultimi ottenuti tramite contaminazione di parlato clean che viene descritto in seguito.

2.1 Il parlato clean

Il materiale clean è stato acquisito in una sala di registrazione in FBK, tramite un microfono di alta qualità (Neumann TLM 103) a 96kHz 24 bit. Sono stati registrati 12 parlatori nativi inglesi e 12 parlatori nativi americani, suddivisi in egual numero tra maschi e femmine. Ognuno ha letto il seguente materiale:

- 15 comandi domestici letti;
- 15 comandi domestici spontanei;
- 13 parole d'ordine (keyword);
- 48 frasi foneticamente ricche (dal corpus Harvard);
- 66/67 frasi dal WSJ-5k;
- 66/67 frasi dal WSJ-20k;
- Circa 10 minuti di parlato conversazionale (ad esempio, il parlatore doveva descrivere un film).

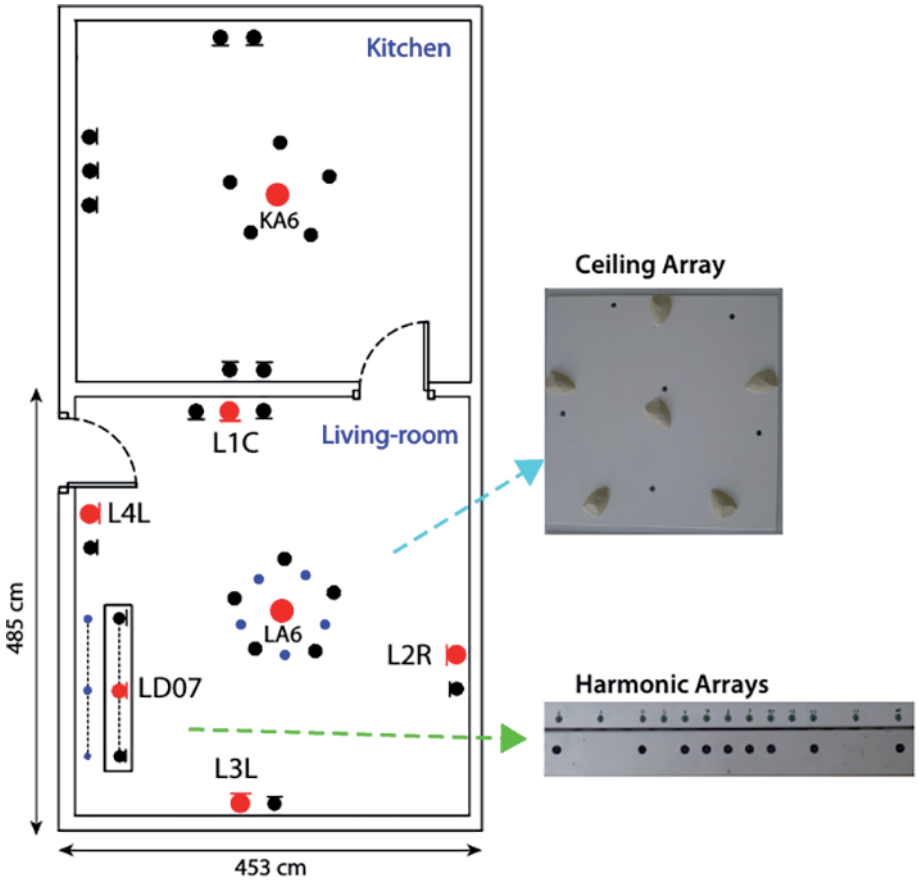
In totale sono state registrate circa 11 ore di materiale vocale e tutte le frasi sono state annotate manualmente. Le frasi foneticamente ricche sono state segmentate a livello fonetico da una procedura automatica (Brugnara, Falavigna & Omologo, 1993); un esperto ha poi controllato le trascrizioni e l'allineamento temporale.

Sei parlatori nativi inglesi e sei parlatori nativi americani sono stati assegnati al development set, mentre gli altri sono stati assegnati al test set. Le assegnazioni sono state effettuate in modo da distribuire le frasi WSJ come nel task originale (Paul, Baker, 1992). Entrambi i set sono compatibili con le specifiche TIMIT.

2.2 La rete microfonica

L'appartamento ITEA è l'appartamento di riferimento che è stato reso disponibile durante il progetto DIRHA per la raccolta di dati e lo sviluppo di prototipi. È composto da cinque stanze che sono state equipaggiate con una rete di diversi microfoni. I microfoni sono in maggior parte degli SHURE MX391 con un pattern omnidirezionale, collegati a delle schede di acquisizione RME Octamic II, campionati a 48kHz - 16 bit in maniera sincrona. Il bagno e altre due stanze sono equipaggiati con un numero limitato di microfoni organizzati in coppie o terne (in totale 12 microfoni), mentre cucina e soggiorno comprendono un numero più elevato di microfoni. Come si può vedere in Figura 1, il soggiorno include tre coppie di microfoni, una terna, due array a soffitto da sei microfoni ognuno (di cui uno composto da microfoni digitali MEMS) e due array armonici (composti rispettivamente da 15 microfoni electret e 15 microfoni digitali MEMS).

Figura 1 - *Schema che rappresenta la distribuzione dei microfoni nel corpus DIRHA-ENGLISH. I punti blu rappresentano i microfoni digitali MEMS, i punti rossi indicano i microfoni utilizzati negli esperimenti mentre i punti neri rappresentano tutti i microfoni disponibili. Le immagini di destra mostrano l'array di microfoni sul soffitto e gli array armonici del soggiorno*



Un'attività particolarmente dispendiosa è stata dedicata a caratterizzare l'ambiente a livello acustico, attraverso più raccolte dati per stimare le risposte impulsive. In totale sono state calcolate più di 10000 risposte impulsive che descrivono come si propaga il suono da vari punti nello spazio ad ognuno dei microfoni presenti. Il metodo adottato per calcolare le risposte impulsive si basa sulla diffusione di uno sweep di frequenze esponenziale (Exponential Sine Sweep, ESS) (Farina, 2000). La rete di microfoni considerata nel DIRHA-ENGLISH (rappresentata in Figura 1) comprende solo soggiorno e cucina, ma include anche gli array armonici e gli array di microfoni MEMS che non sono disponibili negli altri corpora raccolti.

2.3 I data-set simulati

I data-set simulati derivano dal parlato clean descritto nella Sezione 2.1 e dai metodi di contaminazione descritti in (Matassoni et al., 2002; Ravanelli, Sosi, Svaizer & Omologo, 2012). Il corpus risultante consiste in un grande numero di sequenze lunghe un minuto, ognuna comprendente un numero variabile di frasi pronunciate nel soggiorno con differenti livello di rumore di fondo. Sono stati creati quattro tipi di sequenze, corrispondenti ai seguenti task:

- Frasi foneticamente ricche;
- Frasi dal WSJ-5k;
- Frasi dal WSJ-20k;
- Parlato conversazionale (comprendente anche parole d'ordine e comandi).

Sono disponibili le registrazioni di 62 microfoni per ogni sequenza, come descritto nella Sezione 2.2.

2.4 Il data-set reale

Per quello che riguarda le registrazioni di materiale dal vivo, ogni utente ha letto il materiale da un tablet, stando in piedi o seduto in soggiorno. Dopo ogni set di frasi è stato chiesto al parlatore di spostarsi in una nuova posizione con un differente orientamento. Ogni utente ha letto lo stesso materiale che aveva letto in sala di registrazione, descritto nella Sezione 2.1. Le registrazioni dei microfoni MEMS sono state allineate temporalmente con gli altri microfoni in una seconda fase, non essendo stato possibile utilizzare lo stesso clock per sincronizzare le acquisizioni.

Una volta raccolto il materiale sono state derivate sequenze da un minuto in modo da rimanere coerenti con i dati simulati.

3. *Esperimenti e risultati*

Questa sezione descrive i task sperimentali proposti ed i relativi risultati preliminari ottenuti utilizzando la parte US delle frasi foneticamente ricche del corpus DIRHA-ENGLISH.

3.1 Descrizione del contesto sperimentale

3.1.1 Corpora per test e training

In questo lavoro la fase di training è ottenuta impiegando la porzione di training del corpus TIMIT (Garofolo, Lamel, Fisher, Fiscus, Pallett & Dahlgren, 1993). Per gli esperimenti di riconoscimento a distanza il corpus originale TIMIT è stato inoltre riverberato utilizzando tre risposte impulsive misurate nel soggiorno. Inoltre sono state aggiunte alcune sequenze di rumore multi-canale, per simulare condizioni reali. Sia le risposte impulsive che le sequenze di rumore sono differenti da quelle utilizzate per generare il corpus DIRHA-ENGLISH.

La fase di test invece è stata effettuata utilizzando le frasi foneticamente ricche del corpus DIRHA-ENGLISH, sia simulate che reali. In entrambi i casi alle sequenze di un minuto è stato applicato un VAD (Voice Activity Detector) e le sequenze sono state poi sotto-campionate da 48 kHz a 16 kHz.

3.1.2 Estrazione delle feature

Alle frasi è stata effettuata un'estrazione delle feature basata su MFCCs. In particolare, il segnale è stato suddiviso in frame da 25 ms con un overlap di 10 ms; per ogni frame sono state estratte 13 feature MFCCs. Le feature sono state poi raggruppate in un vettore di 39 componenti, assieme alle loro derivate prime e seconde.

3.1.3 Training dei modelli acustici

Negli esperimenti descritti sono stati considerati tre modelli acustici differenti, con una complessità sempre maggiore. La procedura adattata per addestrare i modelli è la stessa utilizzata per la recipe Kaldi di TIMIT s5 (Povey et al., 2011). La prima baseline (mono) si riferisce ad un semplice sistema caratterizzato da 48 fonemi della lingua inglese, indipendenti dal contesto, ognuno modellato con una rete HMM (Hidden Markov Model) a tre stati (in totale usando 1000 gaussiane). La seconda baseline (tri) è basata su una modellizzazione dei fonemi che dipende dal contesto e da un addestramento che si adatta al parlatore. In totale sono utilizzati 2500 stati con 15000 gaussiane.

Per ultima, la terza baseline basata su reti neurali (Deep Neural Networks, DNN) è addestrata con la recipe Karel (Ghoshal, Povey, 2013), composta da sei stati nascosti e 1024 neuroni, su una finestra di 11 frame e un learning rate di 0,008.

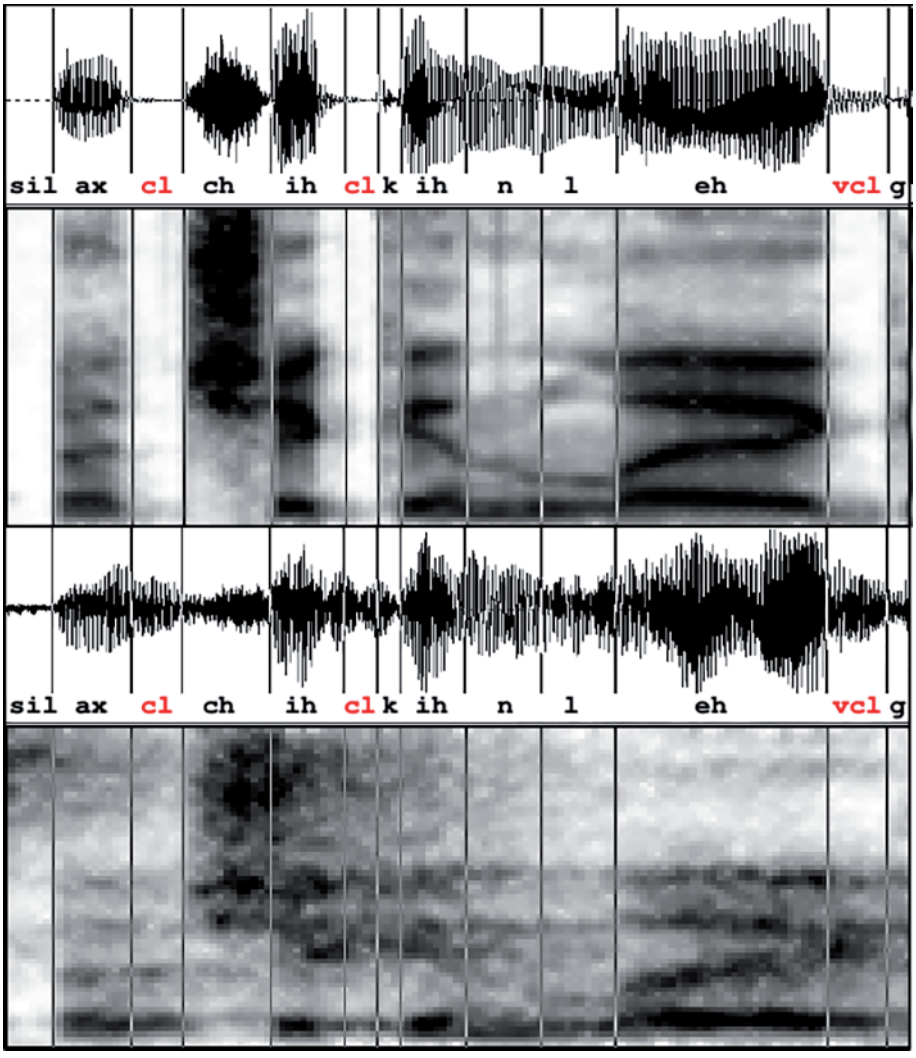
3.1.4 Task proposto e valutazione

La recipe Kaldi si basa sull'impiego di un modello del linguaggio a bigrammi, stimato utilizzando le trascrizioni fonetiche disponibili con il training set. Al contrario, qui proponiamo l'adozione di un puro task phone-loop (zero-grammi) in modo da evitare qualsiasi possibile alterazione dovuta a non-linearità o ad artefatti causati dal modello del linguaggio. I nostri lavori precedenti (Ravanelli et al., 2012; Ravanelli, Omologo, 2014; Ravanelli, Omologo, 2015), infatti, suggeriscono che, sebbene l'impiego di modelli del linguaggio sia certamente utile al fine di aumentare le prestazioni di riconoscimento, l'utilizzo di un semplice task phone-loop è più adatto nel caso di esperimenti che focalizzano l'attenzione sull'informazione acustica.

Un'altra differenza rispetto alla recipe originale Kaldi riguarda la valutazione di silenzi e closures. In fase di valutazione, la recipe standard Kaldi (basata su sclite) mappa le 48 unità fonetiche inglesi in un set ridotto a 39 unità, come originariamente fatto in (Lee, Hon, 1989). In particolare, sei closures (*bcl,dcl,gcl,kcl,pcl,tcl*) vengono mappate come "silenzio opzionale", e possibili cancellazioni di queste unità non vengono considerate come errori in fase di valutazione. Errori relativi a tali unità sono molto frequenti, ed il loro contributo al calcolo complessivo del tasso di errore può introdurre un "bias". Questo aspetto risulta rilevante soprattutto nel caso del riconoscimento vocale a distanza dai microfoni, in cui le code del riverbero ren-

dono l'individuazione di tali unità praticamente impossibile, come evidenziato in Figura 2. Per questa ragione, in questo lavoro proponiamo semplicemente di eliminare silenzi e closures sia dalla sequenza di riferimento che dalla sequenza fonetica prodotta dal sistema di riconoscimento. Questa scelta porta ad un peggioramento nelle prestazioni del sistema, in quanto tutti i silenzi opzionali inclusi aggiunti nel caso della recipe originale non vengono più considerati. Allo stesso tempo, si ottiene una stima più coerente delle prestazioni del sistema, per quel che riguarda unità fonetiche di maggiore importanza, quali ad esempio vocali e burst di occlusive.

Figura 2 - *La frase "a chicken leg" registrata tramite close-talk in studio (in alto) e con microfono distante nell'ambiente reale (in basso). Le closures (in rosso) sono coperte dalla coda di riverbero nella registrazione ambientale*



3.2 Risultati sperimentali

Questa sezione fornisce alcuni risultati *baseline*, che possono risultare utili come riferimento per altri ricercatori che intendessero utilizzare questo corpus. Nelle prossime sezioni vengono presentati risultati ottenuti sia nel caso di input close-talk che di input da microfono posto a distanza dal parlatore.

3.2.1 Performance nel caso di input close-talk

Come riportato in Tabella 1, le prestazioni ottenute decodificando sequenze vocali clean (ovvero acquisite nello studio di registrazione di FBK) attraverso l'impiego di un modello del linguaggio a bigrammi di fonemi o di un semplice loop di unità fonetiche (phone loop).

I risultati sono stati ottenuti attraverso la recipe standard Kaldi s5 e attraverso l'alternativa recipe basata sulla nostra proposta di valutazione degli errori, in modo da poter evidenziare le discrepanze nei risultati fra le due diverse condizioni sperimentali.

Tabella 1 - *Phone Error Rate (PER%) ottenuta applicando differenti recipes Kaldi alle frasi fonetiche acquisite nella sala registrazioni di FBK*

Recipe	LM type	Mono	Tri	DNN
Standard Kaldi s5	Bigram LM	36.4	23.2	20.1
Standard Kaldi s5	Phone-loop	39.4	26.3	22.4
Proposed Evaluation	Bigram LM	42.7	28.6	24.6
Proposed Evaluation	Phone-loop	46.7	32.5	27.5

Come prevedibile, i risultati evidenziano che le prestazioni cambiano significativamente quando si passa dal semplice caso di GMM che modellano unità indipendenti (monofoni) dal contesto al caso di DNN. Inoltre, come evidenziato in Sezione 3.1.4, applicando la recipe Kaldi originale si osserva una riduzione relativa prossima al 20% del tasso di errore rispetto alla procedura da noi proposta, che, di fatto, non corrisponde ad alcun miglioramento nelle capacità del sistema, ma esclusivamente al metodo di valutazione.

Gli esperimenti di riconoscimento a distanza dal microfono che vengono descritti nelle prossime sezioni si riferiscono al solo caso di phone loop e di valutazione basata sulla procedura da noi proposta.

3.2.2 Prestazioni del sistema nel caso di singolo microfono distante

In questa sezione vengono riportati e discussi i risultati che sono stati ottenuti nel caso in cui l'input del riconoscitore corrisponde ad uno fra quattro possibili microfoni (LA6, L1C, LD07, KA6) posti a distanza dal parlatore. La Tabella 2 riporta la lista completa di questi risultati.

Tabella 2 - *Phone Error Rate (PER%) ottenuta con un singolo microfono a distanza dal parlatore. "Sim" si riferisce ai dati simulati, mentre "Real" si riferisce ai dati reali*

	Sim Mono	Sim Tri	Sim DNN	Real Mono	Real Tri	Real DNN
LA6	67.0	57.7	51.6	70.5	60.9	55.1
L1C	67.4	58.5	52.4	70.3	61.7	55.6
LD07	67.5	58.1	53.2	71.5	62.6	57.3
KA6	76.7	67.3	64.0	80.5	73.6	70.3

Come evidenziato in tabella, nel caso di microfono a distanza dal parlatore le prestazioni risultano nettamente peggiori rispetto al caso di input close-talk. Come già osservato nel caso di input close-talk, si osserva inoltre che l'impiego della DNN migliora le prestazioni in modo significativo rispetto a quanto è possibile ottenere con gli altri modelli acustici di riferimento. Questa evidenza sperimentale è confermata con tutti e quattro gli input microfonici che sono stati considerati, sia in caso di segnali simulati che di segnali reali. In realtà, le prestazioni nel caso di segnali reali sono leggermente peggiori rispetto al caso di dati simulati, a causa di un inferiore rapporto segnale rumore che caratterizza le registrazioni in ambiente reale.

È altrettanto importante rilevare che il trend generale delle prestazioni risulta sufficientemente coerente al variare dei modelli acustici esaminati. Solo il caso del microfono installato in cucina (KA6) è caratterizzato da un netto peggioramento delle prestazioni, causato da una generale riduzione del rapporto segnale rumore, dovuta al fatto che tutte le frasi sono state lette in salotto.

3.2.3 Prestazioni nel caso di delay-and-sum beamforming

In questa sezione viene esaminato l'andamento delle prestazioni nel caso di impiego della tecnica di delay-and-sum beamforming (Brandstein, Ward, 2000) per la combinazione di segnali acquisiti rispettivamente dall'array di microfoni posto nel soffitto e dall'array armonico, entrambi installati nel salotto.

Tabella 3 - *Phone Error Rate (PER%) ottenuta applicando il delay-and-sum beamforming agli array di microfoni presenti nel soggiorno dell'appartamento ITEA*

	Sim Mono	Sim Tri	Sim DNN	Real Mono	Real Tri	Real DNN
Array a soffitto	66.2	55.9	50.4	65.9	55.9	50.6
Array armonico	66.2	56.0	51.8	66.2	56.2	51.5

I risultati, riportati in Tabella 3, dimostrano che il beamforming risulta utile nel migliorare le prestazioni del sistema, per es. dal 55.1% PER osservato nel caso di singolo microfono al 50.6% PER che si ottiene applicando questa tecnica combinata con DNN, ai segnali acquisiti attraverso l'array installato nel soffitto.

Sebbene quest'ultimo array consista di soli sei microfoni, le prestazioni che esso offre risultano migliori rispetto al caso di array armonico, comprendente 13 microfoni. Questo risultato sperimentale potrebbe essere dovuto al posizionamento del primo dei due array, il quale spesso acquisisce un maggiore contributo in termini di propagazione diretta del suono rispetto all'array armonico. Una seconda motivazione per questo miglioramento di prestazioni è legata alla migliore qualità dei microfoni.

Si osserva inoltre che il miglioramento di prestazioni introdotto dalla tecnica di delay-and-sum beamforming risulta più evidente nel caso di segnali reali. Ciò conferma il fatto che il filtraggio spaziale risulta particolarmente utile nei casi in cui le condizioni acustiche sono meno stazionarie e quindi meno predicibili.

3.2.4 Prestazioni basate su selezione automatica del microfono

Il corpus DIRHA-ENGLISH può essere utilizzato anche in esperimenti basati sull'impiego di tecniche di selezione automatica del microfono. A questo proposito, risulta quindi interessante esaminare alcune prestazioni di riferimento che possono costituire un *upper-bound* per tali tecniche. La Tabella 4 fornisce un confronto tra i risultati ottenuti nel caso di selezione casuale del microfono e quelli ottenuti nel caso (Oracle) in cui per ciascuna frase viene selezionato come input il microfono che assicura il minimo tasso di errore. L'esperimento è stato condotto utilizzando i sei microfoni del salotto indicati in rosso in Figura 1.

Tabella 4 - *Phone Error Rate (PER%) ottenuta applicando una selezione casuale del microfono (random) oppure attraverso una selezione operata con modalità oracolo (Oracle)*

	Sim Mono	Sim Tri	Sim DNN	Real Mono	Real Tri	Real DNN
Random	67.6	57.7	52.4	70.3	61.0	55.4
Oracle	56.6	47.1	42.0	60.3	49.6	44.0

I risultati dimostrano che una opportuna selezione automatica dinamica del microfono può risultare determinante nel miglioramento delle prestazioni di un sistema DSR. Si può osservare una significativa differenza fra l'upper bound indicato nella riga *Oracle* e il lower bound basato su una selezione random del microfono. Ciò conferma l'importanza da attribuire alla tematica di ricerca riguardante la selezione automatica del microfono, la quale è potenzialmente in grado di fornire risultati migliori rispetto all'impiego di delay-and-sum beamforming. Per esempio, un 50.6% PER ottenuto applicando il beamforming all'array del soffitto va confrontato con il corrispondente 44.0% ottenibile nel caso di selezione automatica del microfono ideale.

4. *Conclusioni e lavori futuri*

Questo articolo descrive il corpus multi-microfonico DIRHA-ENGLISH ed alcuni esperimenti preliminari relativi all'utilizzo delle frasi foneticamente ricche. In generale i risultati sperimentali mostrano le prestazioni che ci si aspettava, allineate con altri lavori in questo campo.

Nella ricerca sul riconoscimento del parlato a distanza ci sono vari vantaggi nell'utilizzo di materiale foneticamente ricco con un così vasto numero di microfoni.

Il corpus contiene anche parti del WSJ e parlato conversazionale che potrebbero essere oggetto di una distribuzione pubblica ed oggetto di future competizioni relative al riconoscimento vocale a distanza. Il parlato conversazionale, in particolare, potrebbe essere utile per investigare altri aspetti chiave, come ad esempio la combinazione di ipotesi multi-microfoniche basate su reti di confusione, reticoli multipli e rescoring.

I futuri lavori prevedono lo sviluppo di baseline e relative recipe per l'utilizzo dei microfoni MEMS digitali, per frasi del WSJ e conversazionali, e per l'inglese britannico.

5. *Rilascio del corpus*

Alcune sequenze di un minuto di durata l'una sono disponibili a questo indirizzo: http://dirha.fbk.eu/DIRHA_English. L'accesso ai dati utilizzati in questo articolo ed i relativi documenti saranno possibili tramite i server FBK, con le modalità che saranno riportate nel sito http://dirha.fbk.eu. In futuro altri dati saranno resi disponibili, corredati da documentazione e recipe, ed istruzioni per poter effettuare dei paragoni tra sistemi differenti.

Ringraziamenti

Il lavoro presentato è stato parzialmente finanziato dalla Comunità Europea nell'ambito del Settimo Programma Quadro (FP7/2007-2013), con il contratto 288121-DIRHA.

Riferimenti bibliografici

Brandstein, M., Ward, D. (2000). *Microphone arrays*. Berlin: Springer.

Brugnara, F., Falavigna, D. & Omologo, M. (1993). Automatic segmentation and labeling of speech based on hidden markov models. In *Speech Communication*, 12, 4, 357-370.

Brutti, A., Ravanelli, M., Svaizer, P. & Omologo, M. (2014). A speech event detection/localization task for multi-room environments. In *Proc. of HSCMA*, 157-161.

Couvreur, L., Couvreur, C. & Ris, C. (2000). A corpus-based approach for robust ASR in reverberant environments. In *Proc. of INTERSPEECH*, 397-400.

Cooke, M., Barker, J., Cunningham, S. & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. In *Journal of the Acoustical Society of America*, 120, 5, 2421-2424.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M. & Maragos, P. (2014). *The DIRHA simulated corpus.* In *Proc. of LREC*, 2629-2634.

Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. of the 108th AES Convention*, 18-22.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S. & Dahlgren, N.L. (1993). *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*.

Ghoshal, A., Povey, D. (2013). Sequence discriminative training of deep neural networks. In *Proc. of INTERSPEECH*.

Haderlein, T., Nöth, E., Herbordt, W., Kellermann, W. & Niemann, H. (2005). Using Artificially Reverberated Training Data in Distant-Talking ASR. In *Lecture Notes in Computer Science*, 3658, 226-233. Springer.

Hänsler, E., Schmidt, G. (2008). *Speech and Audio Processing in Adverse Environments*. Springer.

Lee, K.F., Hon, H.W. (1989). Speaker-independent phone recognition using hidden markov models. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37, 11, 1641-1648.

Matassoni, M., Omologo, M., Giuliani, D. & Svaizer, P. (2002). Hidden Markov model training with contaminated speech material for distant-talking speech recognition. In *Computer Speech & Language*, 16, 2, 205-223.

Matassoni, M., Astudillo, R., Katsamanis, A. & Ravanelli, M. (2014). The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones. In *Proc. of INTERSPEECH*, 1616-1617.

Paul, D.B., Baker, J.M. (1992). The design for the wall street journal-based csr corpus. In *Proc. of the Workshop on Speech and Natural Language*, 357-362.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proc. of ASRU*.

Ravanelli, M., Omologo, M. (2014). On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In *Proc. of INTERSPEECH*, 1028-1032.

Ravanelli, M., Omologo, M. (2015). Contaminated speech training methods for robust DNN-HMM distant speech recognition. In *Proc. of INTERSPEECH*.

Ravanelli, M., Sosi, A., Svaizer, P. & Omologo, M. (2012). Impulse response estimation for robust speech recognition in a reverberant environment. In *Proc. of EUSIPCO*, 1668-1672.

Yu, D., Deng, L. (2015). *Automatic Speech Recognition - A Deep Learning Approach*. Springer.

Wölfel, M., McDonough, J. (2009). *Distant Speech Recognition*. Wiley.

Zwyssig, E., Ravanelli, M., Svaizer, P. & Omologo, M. (2015). A multi-channel corpus for distant-speech interaction in presence of known interferences. In *Proc. of ICASSP*, 4480-4485.