EFFETTO DEL RUMORE AMBIENTALE SULLA STIMA DELL'SNR NELLE SOTTOBANDE DI INTERESSE DEGLI INDICI BIOMETRICI NELLE COMPARAZIONI FONICHE

Francesco Beritelli
Università di Catania – Dipartimento di Ingegneria Informatica e delle Telecomunicazioni
beritelli@diit.unict.it

1. SOMMARIO

Il rumore ambientale rappresenta una delle principali cause di degradazione delle prestazioni in diverse area applicative dell'elaborazione digitale della voce ed, in particolare, nel campo della codifica, riconoscimento, e della biometria vocale.

In generale, infatti, il comportamento di un sistema di elaborazione della voce dipende dal livello di potenza del rumore presente nell'ambiente, ma anche dal tipo di rumore che può essere di diversa natura (*car, bus, office, restaurant, street, stadium,* ecc) presentando, caso per caso, caratteristiche statistiche e spettrali diverse.

Nell'ambito del riconoscimento automatico o semi-automatico del parlatore, in particolare, il rumore ambientale rappresenta una delle principali cause di alterazione degli indici acustici utilizzati nella fase di identificazione/verifica biometrica.

L'attendibilità di un sistema di comparazione fonica, come è noto, dipende dalla quantità del materiale fonico disponibile, in particolare dalla numerosità delle vocali presenti nella sequenza oggetto della perizia, e dalla qualità del segnale fonico. Il primo aspetto incide sul potere risolutivo del sistema, il secondo ha un impatto sulla corretta stima degli indici biometrici. La qualità delle registrazioni è, quindi, fondamentale per avere una buona accuratezza nella identificazione del parlatore.

E' importante comprendere bene il diverso impatto che il rumore può avere sull'attendibilità di un sistema di comparazione fonica in funzione delle varie caratteristiche tempo-frequenza che esso presenta.

In questo lavoro, in particolare, si analizza il diverso effetto del rumore ambientale sulle prime tre formanti, ovvero gli indici biometrici tipicamente utilizzati nell'ambito delle comparazioni foniche. L'analisi viene effettuata attraverso il calcolo del rapporto segnale-rumore (SNR) nelle sottobande di interesse relative alle prime tre formanti al variare della vocale e del tipo di rumore.

2. CARATTERISTICHE SPETTRALI DEI RUMORI AMBIENTALI

La prima fase di questo lavoro è stata rivolta alla caratterizzazione dei rumori ambientali con lo scopo di analizzare e mettere in evidenza le diverse caratteristiche tempo-frequenza dei rumori ambientali.

Il database dei rumori raccoglie un insieme di registrazioni relative a diversi tipi di rumore ambientale, ciascuna della durata di 3 minuti, campionata a frequenza $F_s = 8\ kHz$ ed quantizzata linearmente a *16 bit*.

I rumori contenuti nel database sono riconducibili alle seguenti categorie:

- **Bus**, registrazione effettuata all'interno di un autobus durante il suo percorso;
- **Construction**, registrazione del rumore prodotto dalle apparecchiature presenti in un cantiere edile;

- **Dump**, registrazioni del rumore del tubo di scarico di un'automobile in varie condizioni di marcia;
- **Factory**, registrazione dei suoni tipici di un fabbrica;
- **Car**, registrazioni effettuate all'interno dell'abitacolo di una autovettura;
- **Office**, registrazioni effettuate durante le ore di lavoro in un ufficio;
- **Pool**, registrazione dei suoni tipici di una piscina all'aperto;
- **Restaurant**, registrazioni dei rumori tipici di un ristorante;
- **Shopping**, registrazione dei suoni avvertibili in un centro commerciale;
- **Station**, registrazione dei suoni tipici degli ambienti di cui si compone una stazione ferroviaria;
- **Street**, registrazione dei suoni che si avvertono durante una passeggiata per le strade di una città;
- **Stadium**, registrazione dei suoni che si avvertono durante lo svolgimento di una manifestazione sportiva in uno stadio;
- **Train**, registrazione dei suoni che si avvertono all'interno di uno scompartimento di un treno durante la sua marcia.

Una prima analisi riguarda la distribuzione spettrale dei vari tipi di rumore per capire la loro differente incidenza sulle tre sottobande relative ai range di interesse delle prime tre formanti.

La Figura 1, mette in evidenza la densità spettrale di potenza di 4 differenti tipi di rumore presenti nel database. La prima osservazione riguarda la non uniformità dello spettro: ad esclusione del caso di rumore "construction" che rimane pressoché costante per tutta la banda fonica (0÷4 kHz), in generale la densità spettrale decresce con la frequenza e presenta profili diversi da rumore a rumore.

Un'altra considerazione riguarda la stazionarietà/non stazionarietà del rumore: nel primo caso (es. *car* e *bus* noise) la densità spettrale a lungo termine (Fig. 1) con buona approssimazione coincide con quella a breve termine, mentre nel secondo caso (es. stadium, office, restaurant noise) lo spettro a lungo termine è solo una rappresentazione che deriva da una media degli spettri a breve termine.

In generale, quindi, attraverso un'analisi per sottobande, si può osservare una differente incidenza del rumore ambientale e, quindi, a parità di potenza del segnale utile, un diverso SNR al variare delle sottobande.

Sulla base di queste considerazioni sarebbe opportuno valutare la qualità del singolo dato biometrico calcolando l'SNR nella relativa banda di interesse e, più in generale, adattare dinamicamente un sistema di elaborazione digitale della voce in funzione del livello e del tipo di rumore.

Per esempio, potrebbe essere opportuno adattare il set di parametri tipicamente estratti da un segnale in un sistema di pattern matching, i filtri di pre/post filtraggio, le soglie nei blocchi di detection, ecc.

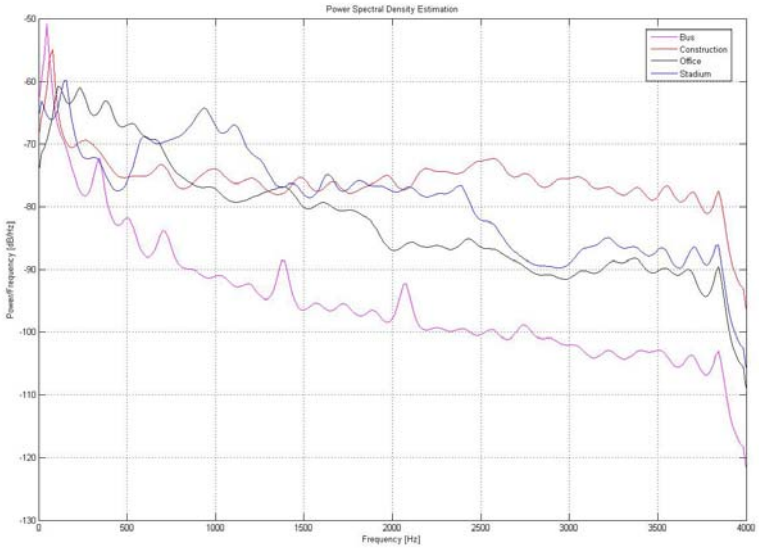


Fig. 1: Densità spettrale di potenza di alcuni tipi di rumori ambientali: bus (linea viola), construction (linea rossa), office (linea nera), stadium (linea blu).

3. INDICI BIOMETRICI E SOTTOBANDE DI INTERESSE

Gli indici biometrici maggiormente utilizzati nei metodi parametrici di identificazione automatica del parlatore sono la frequenza fondamentale o pitch (F_o) e le prime 3 formanti (F_1, F_2, F_3) calcolate per le quattro vocali dell'italiano: a, e , i ed o. In generale, uno dei parametri utilizzati per la valutazione della qualità di un segnale fonico (principalmente qui si fa riferimento a quello relativo alle intercettazioni in quanto il saggio fonico viene tipicamente registrato in ambienti non rumorosi) è rappresentato dal rapporto Segnale/Rumore (S/N o SNR) ovvero dal rapporto tra la potenza del segnale utile (la voce) e quella del disturbo (che in questo lavoro supponiamo essere solo il rumore ambientale). In genere, nelle tecniche parametriche, durante la fase di segmentazione si scartano le vocali che presentano un SNR calcolato nell'intera banda fonica inferiore a 6 dB in quanto esiste un legame tra SNR e l'alterazione dei valori degli indici biometrici rispetto ai valori nominali che essi assumono quando il segnale fonico è *clean* (ovvero quello registrato in ambienti non rumorosi).

Una importante considerazione riguarda il fatto che una misura dell'SNR su tutta la banda fonica (0÷4 kHz) non tiene conto del diverso effetto del rumore sulle bande di frequenza di interesse dei singoli indici biometrici con una possibile duplice conseguenza: lo scarto di misure "buone" o l'inserimento di dati biometrici "distorti", ovvero caratterizzati da valori che si discostano da quelli effettivi. La prima comporterebbe una riduzione del potere risolutivo, e, la seconda, l'introduzione di dati alterati. In entrambi i casi il risultato del test di identità potrebbe essere compromesso, soprattutto in quei casi in cui il test di comparazione è relativo a due voci appartenenti allo stesso parlatore (falsa esclusione).

Il presente lavoro propone di estendere l'analisi della qualità del segnale fonico ad una analisi più dettagliata rivolta ai singoli dati biometrici. In particolare, per tener conto della diversa distribuzione della densità spettrale di potenza tra varie tipologie di rumori, si calcola l'SNR nelle singole sottobande di interesse relative alle 3 prime formanti: F1, F2 e F3. La Tabella 1 mostra gli intervalli considerati per le sottobande di interesse in cui in genere cadono i valori delle 3 formanti. I valori sono quelli adottati all'interno di IDEM, il tool sviluppato dalla Fondazione Ugo Bordoni per le analisi foniche comparative.

Per semplicità in questo lavoro non si distingue tra una vocale aperta o chiusa in quanto si suppone che il range di frequenza della singola sottobanda non dipende dal tipo di coarticolazione nella pronunzia della vocale. Naturalmente per tenere conto di un sistema vocalico più ampio si tratta semplicemente di considerare nello studio ulteriori range di variazione delle formanti.

I passi più significativi per lo svolgimento di tale lavoro possono essere riassunti come segue:

- Creazione dei database nelle varie condizioni di SNR e tipologie di rumore ambientale attraverso una somma digitale del rumore con il segnale *clean*;
- Segmentazione delle vocali presenti nelle sequenze *clean*;
- Stima dell'SNR nelle 3 sottobande di interesse degli indici biometrici;
- Analisi dei risultati ottenuti.

In particolare il database vocale di partenza è costituito da conversazioni *"clean"* in lingua italiana della durata media di 3 minuti con un tempo di OFF (i tratti di pausa in cui è presente solo il rumore ambientale) medio di circa 5 secondi ed un tempo medio di ON (i tratti di parlato) di circa 3 secondi (la distribuzione dei tempi di ON e OFF è di tipo esponenziale). Il segnale vocale è campionato ad una frequenza di 8kHz e quantizzato linearmente a 16 bit per campione. Il database è stato marcato individuando i tratti legati al set delle vocali *a, e, i ,o* presenti nelle conversazioni.

A partire dal database descritto, aggiungendo digitalmente le sequenze di rumore, sono stati creati un totale di 12 altri database, utilizzando tre tipi diversi di rumore (Car, Office, Restaurant), e quattro differenti rapporti segnale-rumore (0dB, 10dB, 20dB, 30dB).

Il rumore di tipo *Car* appartiene alla categoria dei rumori *stazionari*, mentre i rumori *Restaurant* e *Office* appartengono alla categoria dei rumori *non stazionari*.

Per quanto riguarda il calcolo dell'SNR questo è stato calcolato all'uscita di 3 filtri passa banda i cui range di applicazione dipendono dalla vocale in esame in base alla Tabella 1.

	Formante		
Vocale	F_1 **(Hz)**	F_2 **(Hz)**	F_3 **(Hz)**
A	478÷752	1154÷2700	2141÷3700
E	343÷700	1480÷2170	2191÷2750
I	220÷434	1746÷2450	2451÷3100
O	379÷715	720÷1202	2000÷2800

Tabella 1: Sottobande di interesse delle formanti al variare delle vocali

Partendo dalla marcatura delle singole vocali, si calcola l'SNR come rapporto tra la potenza del segnale utile e la potenza di rumore, centrando, con una finestra di 512

campioni, la vocale e il corrispondente rumore e applicando separatamente i tre filtri relativi ai range di interesse delle tre formanti.

4. RISULTATI

Le figure 2, 3 e 4 rappresentano l'andamento dell'SNR misurato per un set di 25 esempi delle vocali a, e, i, o, estratte da una conversazione caratterizzata da un SNR medio di 10 dB in presenza di rumore *car, office e restaurant*. I grafici riportano l'andamento dell'SNR relativo alla sottobanda di interesse della prima formante (spezzata in blu), alla sottobanda della seconda formante (spezzata in verde), alla sottobanda di interesse della terza formante (spezzata in rosso) e, infine, l'SNR totale (spezzata in nero) ovvero l'SNR calcolato sull'intera banda fonica 0÷4 kHz. Una prima semplice considerazione riguarda la variabilità dell'SNR nel caso di rumore *car*, che, essendo di tipo stazionario, è naturalmente causata dalla non stazionarietà del segnale vocale. In generale, gli SNR calcolati nelle sottobande di interesse relative alle tre formanti superano il valore dell'SNR totale, ma si verificano anche, in alcuni casi, delle situazioni inverse.

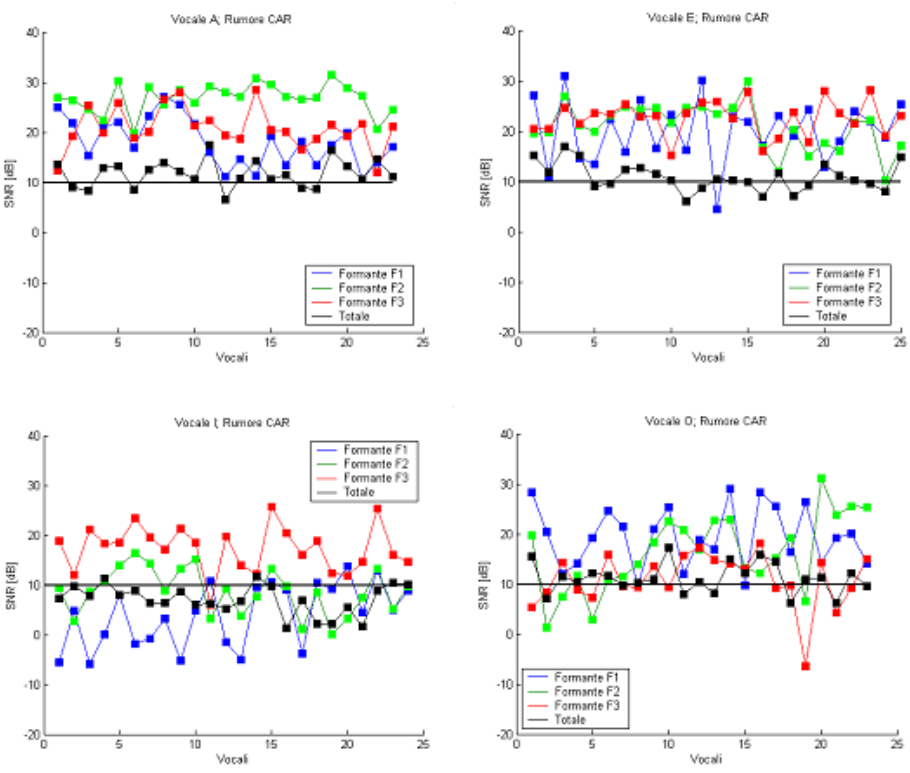


Fig. 2: SNR per un set di 25 esempi di vocali (rumore CAR, 10 dB)

Si veda, in particolare, l'SNR della prima formante per la vocale "i" e l'SNR della seconda e terza formante per la vocale "o" nel caso di rumore *restaurant*. In generale, quindi, si possono osservare una serie di casi in cui sebbene l'SNR totale superi la soglia

critica dei 6dB, l'SNR relativo ad alcuni indici biometrici è inferiore alla soglia critica e una serie di casi in cui si verifica una situazione opposta in cui l'SNR totale è inferiore alla soglia critica e mentre quello calcolato nelle sottobande relative ad alcune formanti risulta sopra soglia. Nel primo caso l'approccio proposto porterebbe ad uno scarto dei dati biometrici con SNR inferiore alla soglia critica mentre nel secondo caso potrebbero recuperare alcuni indici biometrici che da una analisi dell'SNR totale verrebbero scartati. Si analizzi ad esempio il grafico relativo alla vocale "i" nel caso di rumore *car*: il primo dato della serie presenta un SNR totale che supera i 6 dB e quindi, secondo l'approccio tradizionale, si utilizzerebbero tutti i tre dati delle formanti nelle successive fasi di matching, quando la misura dell'SNR relativo alla prima formante presenta addirittura un SNR negativo e, quindi, questo dato andrebbe sicuramente scartato.

Se si analizza il sedicesimo dato della stessa serie si osserva invece che, sebbene l'SNR totale sia inferiore alla soglia critica dei 6 dB, e quindi si scarterebbe l'intera vocale, i valori dell'SNR nelle tre sottobande di interesse delle formanti sono superiori alla soglia critica e, quindi, questa metodologia salverebbe i relativi dati aumentando il potere risolutivo, ovvero l'affidabilità del test. Analoghi ragionamenti possono essere fatti analizzando i grafici relativi alle altre vocali e agli altri tipi di rumore.

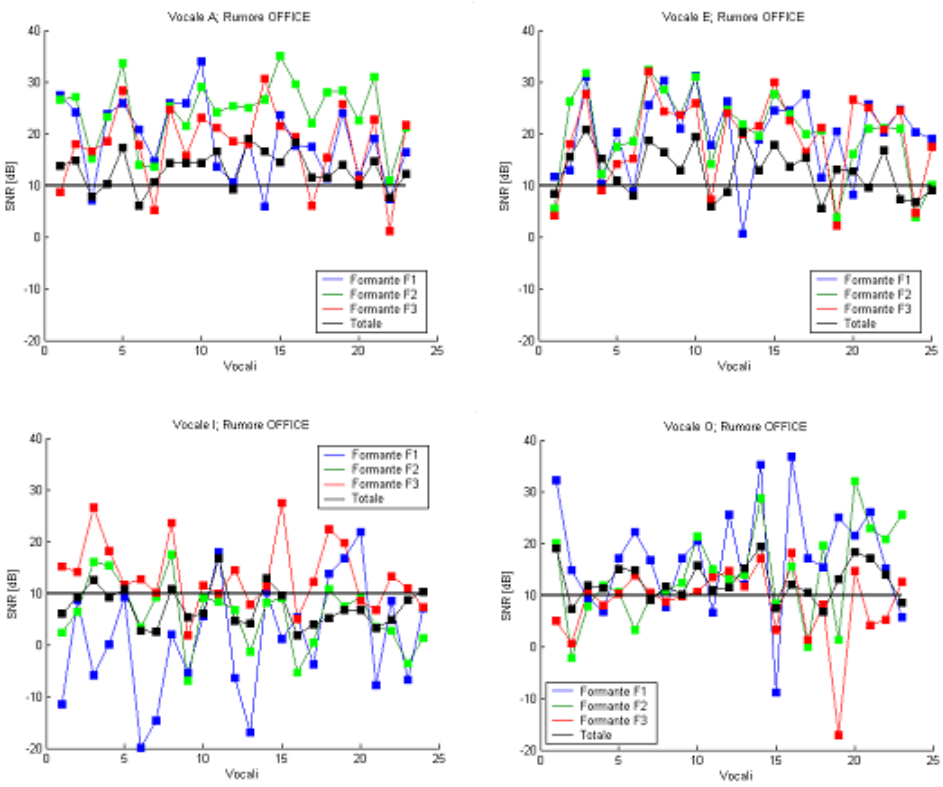


Fig. 3: SNR per un set di 25 esempi di vocali (rumore OFFICE, 10 dB)

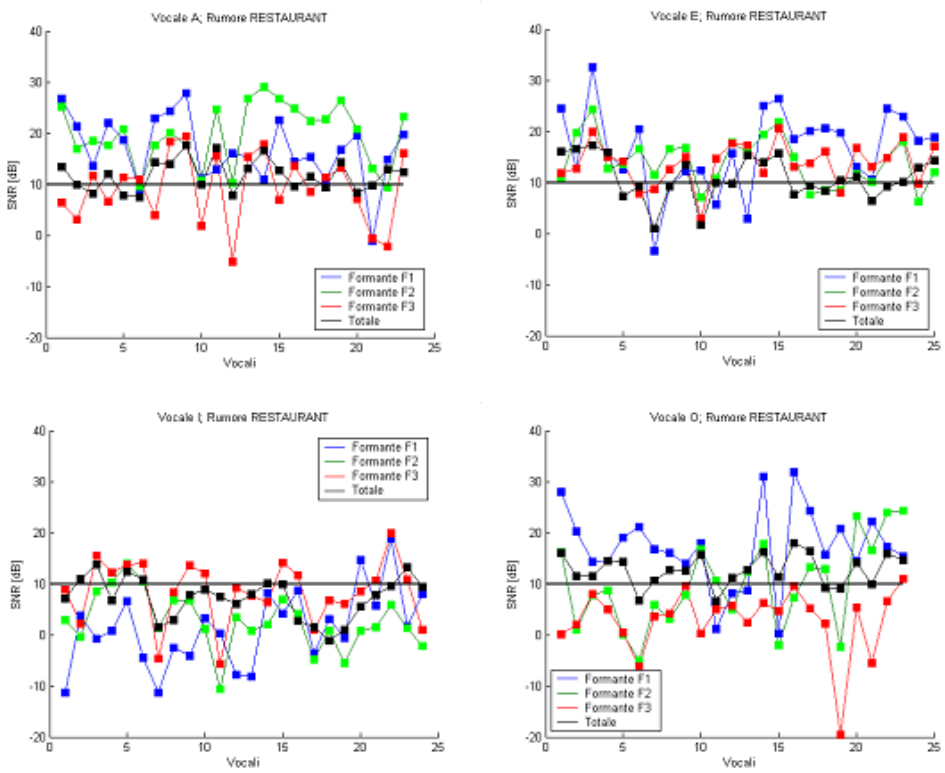
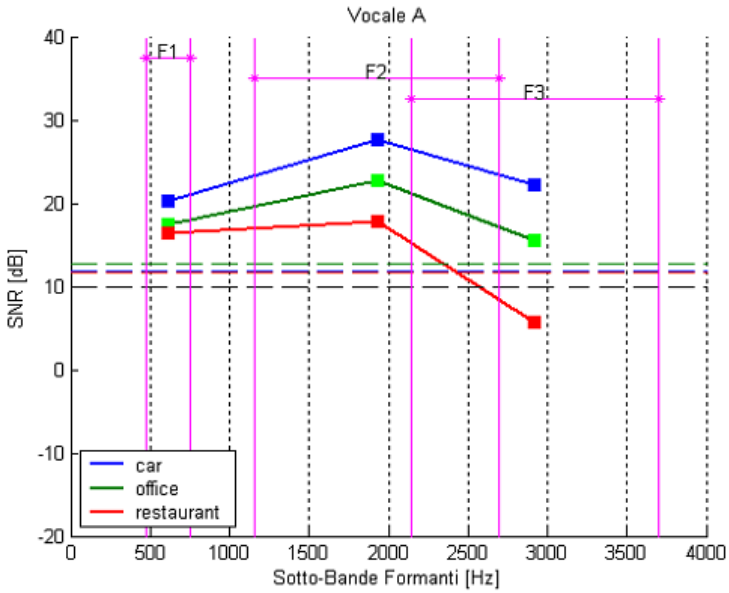


Fig. 4: SNR per un set di 25 esempi di vocali (rumore RESTAURANT, 10 dB)

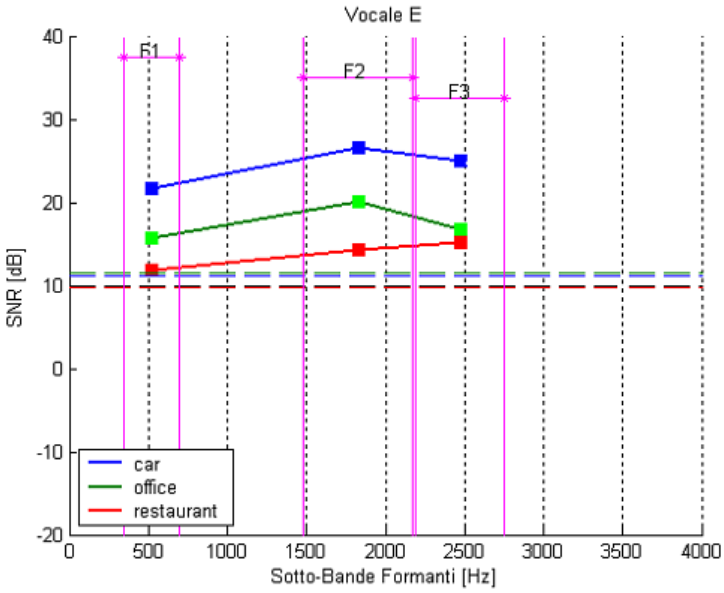
Da questa analisi si deduce che è opportuno effettuare la misura dell'SNR per ogni singola sottobanda di interesse relativa alle tre formanti in quanto si ha una migliore valutazione della bontà del dato biometrico.

La Fig. 5 (a,b,c,d) illustra i valori medi di SNR, espressi in dB, calcolati nelle sottobande di interesse degli indici biometrici nel caso di SNR medio calcolato su tutta la sequenza pari a 10dB e al variare delle quattro vocali e del tipo di rumore. In particolare le curve blu rappresentano il caso con rumore *car*, quelle verdi il caso con rumore *office* e infine le curve in rosso il caso di rumore *restaurant*.

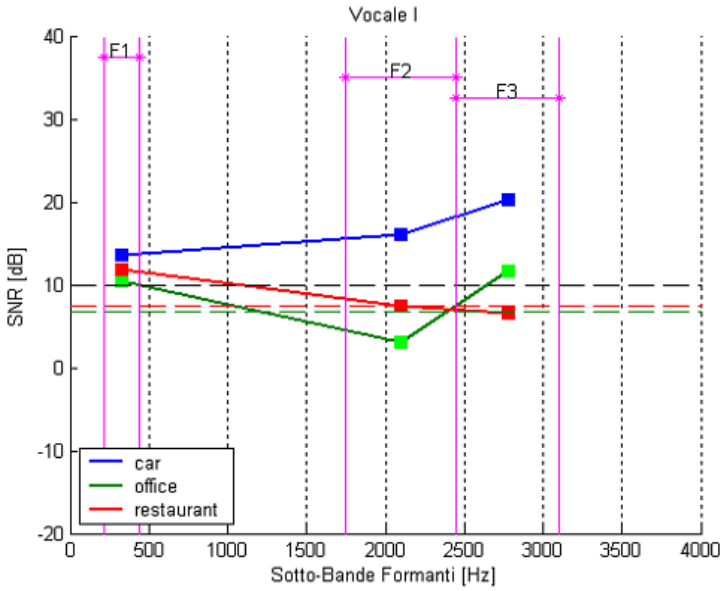
Si nota che uno stesso SNR medio su tutta la conversazione porta a un diverso impatto del rumore a livello di sottobande di interesse degli indici biometrici sia in funzione del tipo di rumore che della specifica sottobanda e vocale.


Vocale A
F1
F2
F3
SNR [dB]
Sotto-Bande Formanti [Hz]
car
office
restaurant

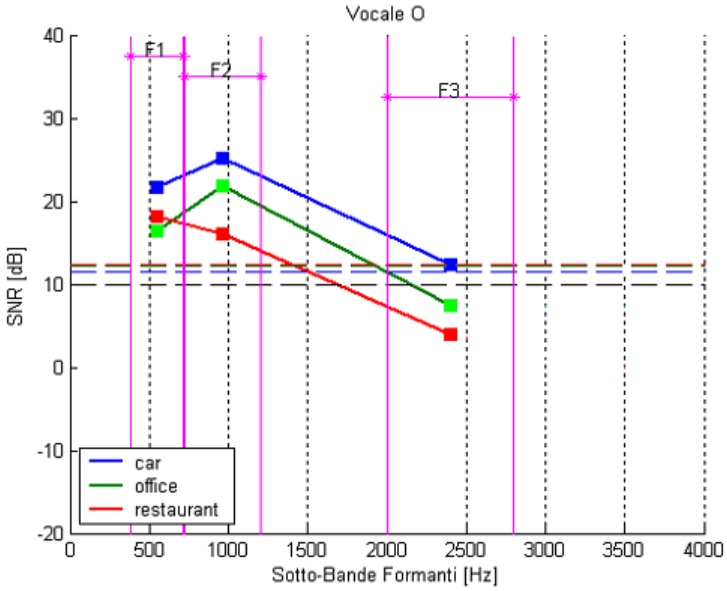
(a)


Vocale E
F1
F2
F3
SNR [dB]
Sotto-Bande Formanti [Hz]
car
office
restaurant

(b)



(c)



(d)

Fig. 5: Valori medi dell'SNR (in dB) nelle sottobande di interesse degli indici biometrici nel caso di SNR medio su tutta la banda pari a 10dB (linea tratteggiata)

In generale è possibile verificare che l'effetto del rumore di tipo *car* non altera il segnale nelle sottobande di interesse come evidenziato in figura 6. La figura, in particolare, riporta, nel caso della vocale *"a"*, la distribuzione dello scarto delle frequenze formati rispetto al valore nominale del caso clean, una volta aggiunto rumore *car* a 10 dB di SNR. Come si evince le variazioni dei valori di frequenza sono minimi per tutte le 4 formanti e questo conferma i risultati mostrate in figura 5 (a) relativi alle misure di SNR per la vocale *"a"* che superano sempre i 20 dB.

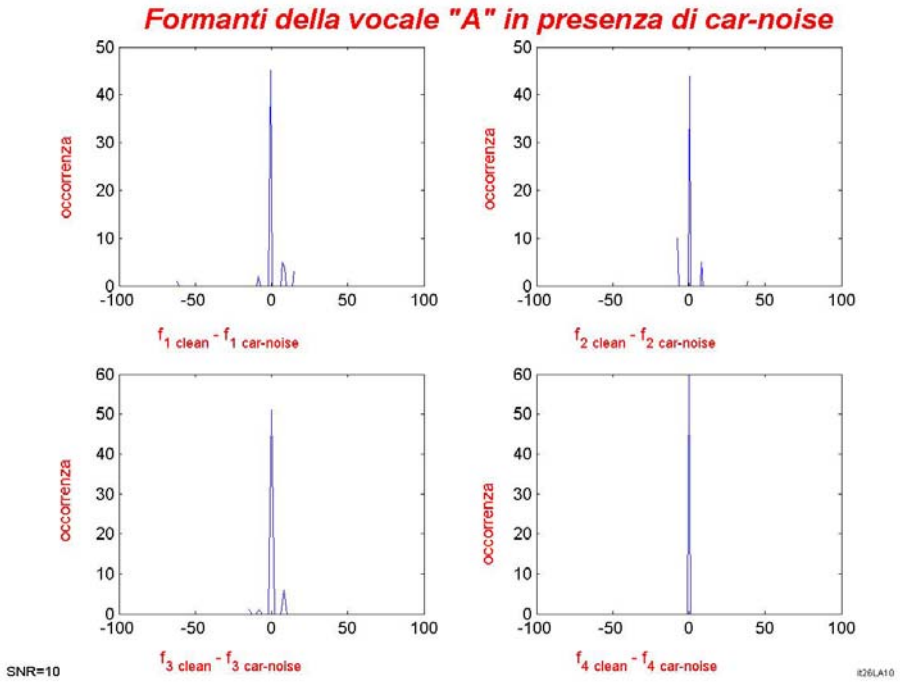


Fig. 6: Sensibilità delle formanti al rumore *car*: distribuzione delle variazioni di frequenza subite a causa del rumore

La figura 7 illustra le distribuzione dello scarto dei valori delle formanti nel caso di rumore *restaurant* a 10 dB di SNR nel caso della vocale *"i"*. In questo caso si ha una certa sensibilità alla presenza del rumore e si notano delle variazioni di frequenza principalmente di piccola entità ma che possono superare in alcuni casi i 100 Hz. Tra le 4 formanti sicuramente la seconda è la più sensibile al disturbo in quanto presenta una più ampia varianza della distribuzione. Anche in questo caso, guardando la figura 3 nel caso della vocale *"i"* possiamo dedurre che esiste un certo legame tra entità delle variazioni subite dalle frequenze formanti e l'SNR misurato in quanto l'SNR è intorno ai 10 dB se misurato nelle bande di interesse della prima e terza formante e vale circa 0 dB nel caso in cui viene calcolato nella sottobanda di interesse della seconda formante.

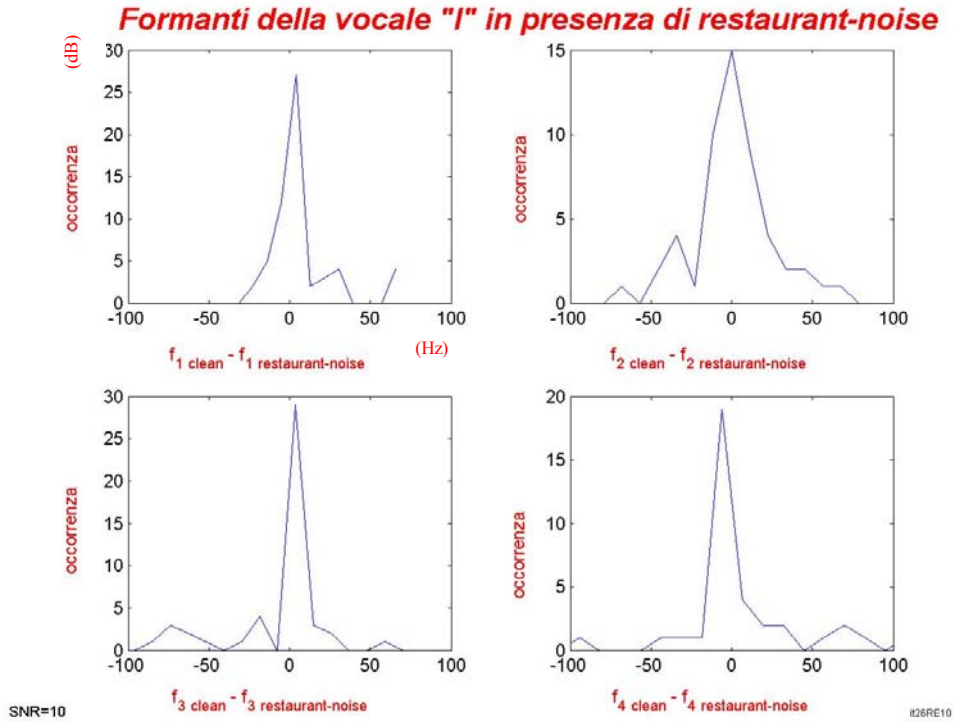


Fig. 7: Sensibilità delle formanti al rumore *restaurant*: distribuzione delle variazioni di frequenza subite a causa del rumore

L'analisi della distribuzione delle variazioni dei valori formatici di sequenze rumorose rispetto al caso clean, suggerisce un criterio per la determinazione degli soglie critiche di SNR al di sotto delle quali è opportuno scartare la singola misura biometrica. Per ogni tipo di rumore, per ogni vocale, e per ogni formante si impone il valore della deviazione standard massima che può subire una formante a causa del rumore. In corrispondenza di tale valore si determina la corrispondente soglia critica.

5. CONCLUSIONI

In conclusione, questo lavoro ha messo in evidenza il diverso effetto del rumore ambientale sul calcolo dell'SNR nelle sottobande di interesse degli indici biometrici adottati nelle comparazioni foniche. I risultati ottenuti indicano un diverso impatto del rumore ambientale al variare della vocale e della singola formante. La stima dell'SNR nella sottobanda di interesse di ogni indice biometrico permette una più accurata selezione dei dati biometrici da utilizzare nelle successive fasi relative al test di identità e di verosimiglianza per il calcolo dell'errore di falsa identità o di falsa esclusione.

BIBLIOGRAFIA

Hollien, H. (2001), *Forensic Voice Identification*, London: Academic Press.

Nolan, F. (1997), Speaker Recognition and Forensic Phonetics, in *A Handbook of Phonetics Science* (W. Hardcastle and J. Laver, editors), Oxford: Blackwell, 146-153.

Falcone, M., Paoloni, A., De Sario, N. (1995), IDEM: A Software Tool to Study Vowel Formant in Speaker Identification, in *Proceedings of the ICPhS '95*, Stoccolma, 145-150.

Paoloni, A. (2003), Note sul riconoscimento del parlante nelle applicazioni forensi con particolare riferimento al metodo parametrico IDEM, *Rivista Italiana di Acustica*, 27, n. 3-4, 113-128.

Reynolds, D.A. (1995), Automatic Speaker Recognition Using Gaussian Mixture Speaker Model, *The Lincoln Laboratory Journal*, 8, 173-191.

Rose, P. (2002), *Forensic Speaker Identification*, London: Taylor and Francis.