1

Multi-channel Feature Enhancement for Robust Speech Recognition

Rudy Rotili, Emanuele Principi, Simone Cifani, Francesco Piazza and Stefano Squartini
Università Politecnica delle Marche
Italy

1. Introduction

In the last decades, a great deal of research has been devoted to extending our capacity of verbal communication with computers through automatic speech recognition (ASR). Although optimum performance can be reached when the speech signal is captured close to the speaker's mouth, there are still obstacles to overcome in making reliable distant speech recognition (DSR) systems. The two major sources of degradation in DSR are distortions, such as additive noise and reverberation. This implies that speech enhancement techniques are typically required to achieve best possible signal quality. Different methodologies have been proposed in literature for environment robustness in speech recognition over the past two decades (Gong (1995); Hussain, Chetouani, Squartini, Bastari & Piazza (2007)). Two main classes can be identified (Li et al. (2009)).

The first class encompasses the so called model-based techniques, which operate on the acoustic model to adapt or adjust its parameters so that the system fits better the distorted environment. The most popular of such techniques are multi-style training (Lippmann et al. (2003)), parallel model combination (PMC) (Gales & Young (2002)) and the vector Taylor series (VTS) model adaptation (Moreno (1996)). Although model-based techniques obtain excellent results, they require heavy modifications to the decoding stage and, in most cases, a greater computational burden.

Conversely, the second class directly enhances the speech signal before it is presented to the recognizer, and show some significant advantages with respect to the previous class:

- independence on the choice of the ASR engine: there is no need of intervening into the (HMM) of the ASR since all modifications are accomplished at the feature level, which has a significant practical mean;
- ease of implementation: the algorithm parameterization is extremely simpler than in the model-based case study and no adaptation is requested to find the optimal one;
- lower computational burden, surely relevant in real-time applications.

The wide variety of algorithms in this class can be further divided based on the number of channels used in the enhancing stage.

Single-channel approaches encompass classical techniques operating in the frequency domain such as Wiener filtering, spectral subtraction (Boll (1979)) and Ephraim & Malah (logMMSE STSA) (Ephraim & Malah (1985)), as well as techniques operating in the feature domain such

as the MFCC-MMSE (Yu, Deng, Droppo, Wu, Gong & Acero (2008)) and its optimizations (Principi, Cifani, Rotili, Squartini & Piazza (2010); Yu, Deng, Wu, Gong & Acero (2008)) and VTS speech enhancement (Stouten (2006)). Other algorithms belonging to the single-channel class are feature normalization approaches as cepstral mean normalization (CMN) (Atal (1974)), cepstral variance normalization (CVN) (Molau et al. (2003)), higher order cepstral moment normalization (HOCMN), histogram equalization (HEQ) (De La Torre et al. (2005)) and parametric feature equalization (Garcia et al. (2006)).

Multi-channel approaches use the benefits of the additional informations carried out by the presence of multiple speech observations. In most cases the speech and noise sources are in different spatial locations, thus a multi-microphone system is theoretically able to obtain a significant gain over single-channel approaches, since it may exploit the spatial diversity.

This chapter will be devoted to illustrate and analyze multi-channel approaches for robust ASR in both the frequency and feature domain. Three different subsets will be addressed highlighting advantages and drawbacks of each one: beamforming techniques, bayesian estimators (operating at different level of the feature extraction pipeline) and histogram equalization.

In ASR scenario, beamforming techniques are employed as pre-processing stage. In (Omologo et al. (1997)) the delay and sum beamformer (DSB) has been successfully used coupled with a talker localization algorithm but its performance are poor when the number of microphones is small (less than 8) or when it operates in a reverberant environment. This motivated the scientific community to develop more robust beamforming techniques e.g. generalized sidelobe canceler (GSC) and transfer function GSC (TF-GSC). Among the beamforming techniques, likelihood maximizing beamforming (LIMABEAM) is an hybrid approach that uses informations from the decoding stage to optimize a filter and sum beamformer (Seltzer (2003)).

Multi-channel bayesian estimators in frequency domain has been proposed in (Lotter et al. (2003)) where both minimum mean square error (MMSE) and maximum a posteriori (MAP) criteria were developed. The feature domain counterpart of the previous algorithms has been presented in (Principi, Rotili, Cifani, Marinelli, Squartini & Piazza (2010)). The simulations conducted on the Aurora 2 database showed performance similar to the frequency domain ones with the advantage of a reduced computational burden.

The last subset that will be addressed, is the multi-channel variant of histogram equalization (Squartini et al. (2010)). Here the presence of multiple audio channels is exploited to better estimate the histograms of the input signal and so making the equalization processing more effective.

The outline of this chapter is as follows: section 2 describe the feature extraction pipeline and the adopted mathematical model. Section 3 gives a brief review of the beamforming concept mentioning some of most popular beamformer. Section 4 is devoted to illustrate the multi-channel MMSE and MAP estimators both in frequency and feature domain while section 5 proposes various algorithmic architectures for multi-channel HEQ. Section 6 presents and discuss recognition results in a comparative fashion. Finally, section 7 draws conclusions and proposes future developments.

2. ASR front-end and mathematical background

In the feature-enhancement approach, the features are enhanced before the ASR decoding stage, with the aim of making them as close as possible to the clean-speech environment condition. This means that some extra-cleaning steps are performed into or after the

feature extraction module. As shown in figure 1, the feature extraction pipeline has four possible insertion points, each one being related to different classes of enhancement algorithms. Traditional speech enhancement in the discrete-time Fourier transform (DFT) domain (Ephraim & Malah (1984); Wolfe & Godsill (2003)), is performed at point 1, mel-frequency domain algorithms (Yu, Deng, Droppo, Wu, Gong & Acero (2008); Rotili et al. (2009)), operate at point 2 and log-mel or MFCC (mel frequency cepstral coefficients) domain algorithms (Indrebo et al. (2008); Deng et al. (2004)), are performed at point 3 and 4 respectively. Since the focus of traditional speech enhancement is on the perceptual quality of the enhanced signal, the performance of the former class is typically lower than the other classes. Moreover, the DFT domain has a much higher dimensionality than mel or MFCC domains, which leads to an higher computational cost of the enhancement process. Let us

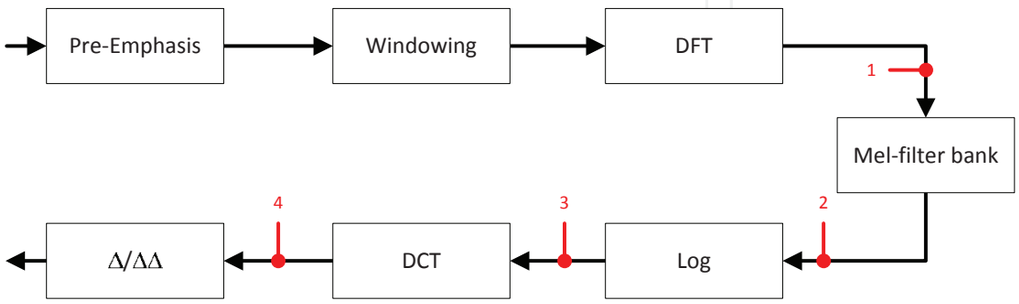


Fig. 1. Feature extraction pipeline.

consider M noisy signals $y_i(t)$, M clean speech signals $x_i(t)$ and M uncorrelated noise signals $n_i(t)$, $i \in \{1, \ldots, M\}$, where t is a discrete-time index. The i-th microphone signal is given by:

$$y_i(t) = x_i(t) + n_i(t). \tag{1}$$

In general, the signal $x_i(t)$ is the convolution between the speech source and the i-th room impulse response. In our case study the far-field model (Lotter et al. (2003)) that assumes equal amplitude and angle-dependent TDOAs (Time Difference Of Arrival) has been considered:

$$x_i(t) = x\left(t - \tau_i(\beta_x)\right), \qquad \tau_i = d \sin\left(\beta_x / c\right) \tag{2}$$

where τ_i is the i-th delay, d is the distance between the source and the microphone array, θ_x is the angle of arrival and c is the speed of sound.
According to figure 1, each input signal $y_i(t)$ is firstly pre-emphasized and windowed with a Hamming window. Then, the fast Fourier transform (FFT) of the signal is computed and the square of the magnitude is filtered with a bank of triangular filters equally spaced in the mel-scale. After that, the energy of each band is computed and transformed with a logarithm operation. Finally, the discrete cosine transform (DCT) stage yields the static MFCC coefficients, and the $\Delta/\Delta\Delta$ stage compute the first and second derivatives.
Given the additive noise assumption, in the DFT domain we have

$$Y_i(k,l) = X_i(k,l) + N_i(k,l) \tag{3}$$

where $X(k,l)$, $Y(k,l)$ and $N(k,l)$ denote the short-time Fourier transforms (STFT) of $x(t)$, $y(t)$ and $n(t)$ respectively, where k is the frequency bin index and l is the time frame index. Equation (3) can be rewritten as follows:

$$Y_i = R_i e^{j\phi_i} = A_i e^{j\alpha_i} + N_i, \qquad 1 \leq i \leq M \tag{4}$$

where R_i, ϕ_i, A_i and α_i are the amplitude and phase terms of Y_i and X_i respectively. For simplicity of notation, the frequency bin and time frame indexes have been omitted.
The mel-frequency filter-bank's output power for noisy speech is

$$m_{y_i}(b,l) = \sum_k w_b(k)|Y_i(k,l)|^2 \tag{5}$$

where $w_b(k)$ is the b-th mel-frequency filter's weight for the frequency bin k. A similar relationship holds for the clean speech and the noise. The j-th dimension of MFCC is calculated as

$$c_{y_i}(j,l) = \sum_b a_{j,b} \log m_{y_i}(b,l) \tag{6}$$

where $a_{j,b} = \cos((\pi b/B)(j-0.5))$ are the DCT coefficients. The output of equation (3) denotes the input of the enhancement algorithms belonging to class 1 (DFT domain) and that of equation (5) the input of class 2 (mel-frequency domain). The logarithm of the output of equation (5) is the input for the class 3 algorithms (log-mel domain) while that of equation (6) the input of class 4 (MFCC domain) algorithms.

3. Beamforming

Beamforming is a method by which signals from several sensors can be combined to emphasize a desired source and to suppress all other noise and interference. Beamforming begins with the assumption that the positions of all sensors are known, and that the positions of the desired sources are known or can be estimated as well.
The simplest of beamforming algorithms, the delay and sum beamformer, uses only this geometrical knowledge to combine the signals from several sensors. The theory of DSB originates from narrowband antenna array processing, where the plane waves at different sensors are delayed appropriately to be added exactly in phase. In this way, the array can be electronically steered towards a specific direction. This principle is also valid for broadband signals, although the directivity will then be frequency dependent.
A DSB aligns the microphone signals to the direction of the speech source by delaying and summing the microphone signals.
Let us define the steering vector of the desired source as

$$\mathbf{v}(\mathbf{k}_d,\omega) = \left[\exp\{j\omega\tau_{d,0}\}, \exp\{j\omega\tau_{d,1}\}, \cdots, \exp\{j\omega\tau_{d,M-1}\}\right]^H, \tag{7}$$

where $\mathbf{k}_d$ is the wave number and $\tau_{d,i}, i \in \{1,,\ldots,M\}$ is the delay relative to the i-th channels. The sensor weights $\mathbf{w}_f(\omega)$ are chosen as the complex conjugate steering vector $\mathbf{v}^*(\mathbf{k}_d,\omega)$, with the amplitude normalized by the number of sensors M:

$$\mathbf{w}_f(\omega) = \frac{1}{M}\mathbf{v}^*(\mathbf{k}_d,\omega). \tag{8}$$

The absolute value of all sensor weights is than equal to $1/M$ (uniform weighting) and the phase is equalized for signals with the steering vector $\mathbf{v}(\mathbf{k}_d,\omega)$ (beamsteering).
The beampattern $B(\omega;\theta,\phi)$ of the DSB with uniform sensor spacing d is obtained as

$$B(\omega;\theta,\phi) = \frac{1}{M}\mathbf{v}^H(\mathbf{k}_d,\omega)\mathbf{v}(\mathbf{k},\omega) = \frac{1}{M}\sum_{m=0}^{M-1} \exp\left\{j\omega\left(\frac{M-1}{2}-m\right)\frac{d}{c}\left(cos\theta_d - cos\theta\right)\right\}. \tag{9}$$

This truncated geometric series may be simplified to a closed form as

$$B(\omega;\theta,\phi) = \frac{1}{M}\frac{sin(\omega M\tau_b/2)}{sin(\omega\tau_b/2)} \tag{10}$$

$$\tau_b = \frac{d}{c}(cos\theta_d - cos\theta). \tag{11}$$

This kind of beamformer is proved to perform well when the number of microphones is relatively high, and when the noise sources are spatially white. On the contrary, performance degrade since noise reduction is strongly dependent on the direction of arrival of the noise signal. As a consequence, DSB performance on reverberant environments is poor.

In order to increase the performance, more sophisticated solution can be adopted. In particular, adaptive beamformers can ideally attain high interference reduction performance with a small number of microphones arranged in a small space. GSC (Griffiths & Jim (1982)) attempt to minimize the total output power of an array of sensor under the constraint that the desired source must be unattenuated.

The main drawback of such beamformer is the target signal cancellation that occurs in the presence of steering vector errors. They are caused by errors in microphone positions, microphone gains, reverberation, and target direction. Therefore, errors in the steering vector are inevitable with actual microphone arrays, and target signal cancellation is a serious problem. Many signal processing techniques have been proposed to avoid signal cancellation. In (Hoshuyama et al. (1999)), a robust GSC (RGSC) able to avoid these difficulties, has been proposed, which uses an adaptive blocking matrix consisting of coefficient-constrained adaptive filters. Such filters exploit the reference signal from the fixed beamformer to adapt themselves and adaptively cancel the undesirable influence caused by steering vector errors. The interference canceller uses norm-constrained adaptive filters (Cox et al. (1987)) to prevent target-signal cancellation when the adaptation of the coefficient-constrained filters is incomplete. In (Herbordt & Kellermann (2001); Herbordt et al. (2007)) a frequency domain implementation of the RGSC has been proposed in conjunction with acoustic echo cancellation.

Most of the GSC based beamformers rely on the assumption that the received signals are simple delayed versions of the source signal. The good interference suppression attained under this assumption is severely impaired in complicated acoustic environments, where arbitrary transfer functions (TFs) may be encountered. In (Gannot et al. (2001)), a GSC solution which is adapted to the general TF case (TF-GSC) has been proposed. The TFs are estimated by exploiting the nonstationarity characteristics of the desired signal, as reported in (Shalvi & Weinstein (1996); Cohen (2004)), and then used to calculate the fixed beamformer and the blocking matrix coefficients.

However, in case of incoherent or diffuse noise fields, beamforming alone does not provide sufficient noise reduction, and postfiltering is normally required. Postfiltering includes signal detection, noise estimation, and spectral enhancement.

Recently, a multi-channel postfilter was incorporated into the TF-GSC beamformer (Cohen et al. (2003); Gannot & Cohen (2004)). The use of both the beamformer primary output and the reference noise signals (resulting from the blocking branch of the GSC) for distinguish between desired speech transient and interfering transient, enables the algorithm to work in nonstationary noise environments. The multi-channel postfilter, combined with the TF-GSC, proved the best for handling abrupt noise spectral variations. Moreover, in this algorithm, the decisions made by the postfilter, distinguishing between speech, stationary noise, and

transient noise, might be fed back to the beamformer to enable the use of the method in real-time applications. Exploiting this information will also enable the tracking of the acoustical transfer functions, caused by the talker movements.
A perceptually based variant of the previous architecture have been presented in (Hussain, Cifani, Squartini, Piazza & Durrani (2007); Cifani et al. (2008)) where a perceptually-based multi-channel signal detection algorithm and a perceptually-optimal spectral amplitude (PO-SA) estimator presented in (Wolfe & Godsill (2000)) have been combined to form a perceptually-based postfilter to be incorporated into the TF-GSC beamformer
Basically, all the presented beamforming techniques outperform the DSB. Recalling the assumption of far-field model (equation (2)) where no reverberation is considered and the observed signals are a simple delayed version of the speech source, the DSB is well suited for our purpose and it is not required to take into account more sophisticated beamformers.

4. Multi-channel bayesian estimators

The estimation of a clean speech signal x given its noisy observation y is often performed under the Bayesian framework. Because of the generality of this framework, x and y may represent DFT coefficients, mel-frequency filter-bank outputs or MFCCs. Applying the standard assumption that clean speech and noise are statistically independent across time and frequency as well as from each other, leads to estimators that are independent of time and frequency.
Let $\epsilon = x - \hat{x}$ denote the error of the estimate and let $C(\epsilon) \triangleq C(x, \hat{x})$ denote a non-negative function of ϵ. The average cost, i.e. $E[C(x, \hat{x})]$, is known as Bayes risk $\mathcal{R}$ (Trees (2001)), and it is given by

$$\mathcal{R} \triangleq E[C(x, \hat{x})] = \int\int C(x, \hat{x}) p(x, y) dx dy \tag{12}$$

$$= \int p(y) dy \int C(x, \hat{x}) p(x|y) dx, \tag{13}$$

in which Bayes rule has been used to separate the role of the observation y and the a priori knowledge.
Minimizing $\mathcal{R}$ with respect to $\hat{x}$ for a given cost function results in a variety of estimators. The traditional mean square error (MSE) cost function,

$$C^{MSE}(x, \hat{x}) = |x - \hat{x}|^2, \tag{14}$$

gives the following expression:

$$\mathcal{R}^{MSE} = \int p(y) dy \int |x - \hat{x}|^2 p(x|y) dx. \tag{15}$$

$\mathcal{R}^{MSE}$ can be minimized by minimizing the inner integral, yielding the MMSE estimate:

$$\hat{x}^{MMSE} = \int x p(x|y) dx = E[x|y]. \tag{16}$$

The log-MMSE estimator can be obtained by means of the cost function

$$C^{log-MSE}(x, \hat{x}) = (\log x - \log \hat{x})^2 \tag{17}$$

thus yielding to:

$$\hat{x}^{log-MMSE} = \exp\left\{E[\ln x|y]\right\}. \tag{18}$$

By using the uniform cost function,

$$C^{MAP}(x,\hat{x}) = \begin{cases} 0, & |x-\hat{x}| \leq \Delta/2 \\ 1, & |x-\hat{x}| > \Delta/2 \end{cases} \tag{19}$$

we get the maximum a posteriori (MAP) estimate:

$$\hat{x}^{MAP} = \underset{x}{\operatorname{argmax}}\, p(x|y). \tag{20}$$

In the following several multi-channel bayesian estimators are addressed. First the multi-channel MMSE and MAP estimators in frequency domain, presented in (Lotter et al. (2003)), are briefly reviewed. Afterwards, the feature domain counterpart of the MMSE and MAP estimators respectively is proposed. It is important to remark that feature domain algorithms are able to exploit the peculiarities of the feature space and produce more effective and computationally more efficient solutions.

4.1 Speech feature statistical analysis

The statistical modeling of the process under consideration is a fundamental aspect of the Bayesian framework. Considering DFT domain estimators, huge efforts have been spent in order to find adequate signal models. Earlier works (Ephraim & Malah (1984); McAulay & Malpass (1980)), assumed a Gaussian model from a theoretical point of view, by invoking the central limit theorem, stating that the distribution of the DFT coefficients will converge towards a Gaussian probability density function (PDF) regardless of the PDF of the time samples, if successive samples are statistically independent or the correlation is short compared to the analysis frame size. Although this assumption holds for many relevant acoustic noises, it may fail for speech where the span of correlation is comparable to the typical frame sizes (10-30 ms). Spurred by this issue, several researchers investigated the speech probability distribution in the DFT domain (Gazor & Zhang (2003); Jensen et al. (2005)), and proposed new estimators leaning on different models, i.e., Laplacian, Gamma and Chi (Lotter & Vary (2005); Hendriks & Martin (2007); Chen & Loizou (2007)).

In this section the study of the speech probability distribution in the mel-frequency and MFCC domains is reported, so as to open the way to the development of estimators leaning on different models in these domains as well.

The analysis has been performed either on the TiDigits (Leonard (1984)) and on the Wall Street Journal (Garofalo et al. (1993)) database using one hour clean speech segments built by concatenation of random utterances. DFT coefficients have been extracted using a 32 ms Hamming window with 50% overlap. The aforementioned Gaussian assumption models the real and imaginary part of the clean speech DFT coefficient by means of a Gaussian PDF. However, the relative importance of short-time spectral amplitude (STSA) rather than phase has led researchers to re-cast the spectral estimation problem in terms of the former quantity. Moreover, amplitude and phase are statistically less dependent than real and imaginary parts, resulting in a more tractable problem. Furthermore, it can be shown that phase is well modeled by means of a uniform distribution $p(\alpha) = 1/2\pi$ for $\alpha \in [-\pi, \pi)$. This has lead the authors to investigate the probability distribution of the STSA coefficients.

For each DFT channel, the histogram of the corresponding spectral amplitude was computed and then fitted by means of a nonlinear least-squares (NLLS) technique to six different PDFs:

$$
\begin{aligned}
\textit{Rayleigh:}\quad & p = \frac{x}{\sigma}\exp\left(\frac{-x^2}{2\sigma}\right)\\
\textit{Laplace:}\quad & p = \frac{1}{2\sigma}\exp\left(\frac{-|x-a|}{\sigma}\right)\\
\textit{Gamma:}\quad & p = \frac{1}{\theta^k\Gamma(k)}|x|^{k-1}\exp\left(\frac{-|x|}{\theta}\right)\\
\textit{Chi:}\quad & p = \frac{2}{\theta^k\Gamma(k/2)}|x|^{k-1}\exp\left(\left(\frac{-|x|}{\theta}\right)^2\right)\\
\textit{Approximated Laplace:}\quad & p = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}|x|^{\nu}\exp\left(\frac{-\mu|x|}{\sigma}\right),\ \mu = 2.5 \text{ and } \nu = 1\\
\textit{Approximated Gamma:}\quad & p = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)}|x|^{\nu}\exp\left(\frac{-\mu|x|}{\sigma}\right),\ \mu = 1.5 \text{ and } \nu = 0.01
\end{aligned}
$$

The goodness-of-fit has been evaluated by means of the Kullback-Leibler (KL) divergence, which is a measure that quantifies how close a probability distribution is to a model (or candidate) distribution. Choosing p as the N bins histogram and q as the analytic function that approximates the real PDF, the KL divergence is given by:

$$D_{KL} = \sum_{n=1}^{N}(p(n) - q(n))\log\frac{p(n)}{q(n)}. \tag{21}$$

D_{KL} is non-negative (≥ 0), not symmetric in p and q, zero if the distributions match exactly and can potentially equal infinity. Table 1 shows the KL divergence between measured data and model functions. The divergences have been normalized to that of the Rayleigh PDF, that is, the Gaussian model. The curves in figure 2 represent the fitting results, while

STSA Model	TiDigits	WSJ
Laplace	0.15	0.17
Gamma	0.04	0.04
Chi	0.23	0.02
Approximated Laplace	0.34	0.24
Approximated Gamma	0.31	0.20

Table 1. Kullback-Leibler divergence between STSA coefficients and model functions.

the gray area represents the STSA histogram averaged over the DFT channels. As the KL divergence highlights, the Gamma PDF provides the best model, being capable of adequately fit the histogram tail as well. The modeling of mel-frequency coefficients has been carried out using the same technique employed in the DFT domain. The coefficients have been extracted by applying a 23-channel mel-frequency filter-bank to the squared STSA coefficients. The divergences, normalized to that of the Rayleigh PDF, have been reported in table 2. Again,

Mel-frequency Model	TiDigits	WSJ
Laplace	0.21	0.29
Gamma	0.08	0.07
Chi	0.16	0.16
Approximated Laplace	0.21	0.22
Approximated Gamma	0.12	0.12

Table 2. Kullback-Leibler divergence between mel-frequency coefficients and model functions.

figure 3 represents the fitting results and the mel-frequency coefficient histogram averaged

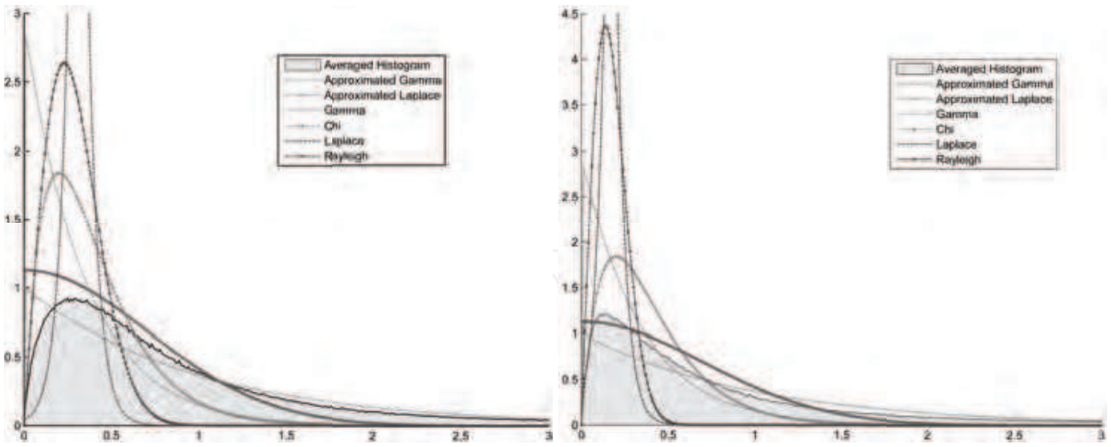


Fig. 2. Averaged Histogram and NLLS fits of STSA coefficients for the TiDigits (left) and WSJ database (right).

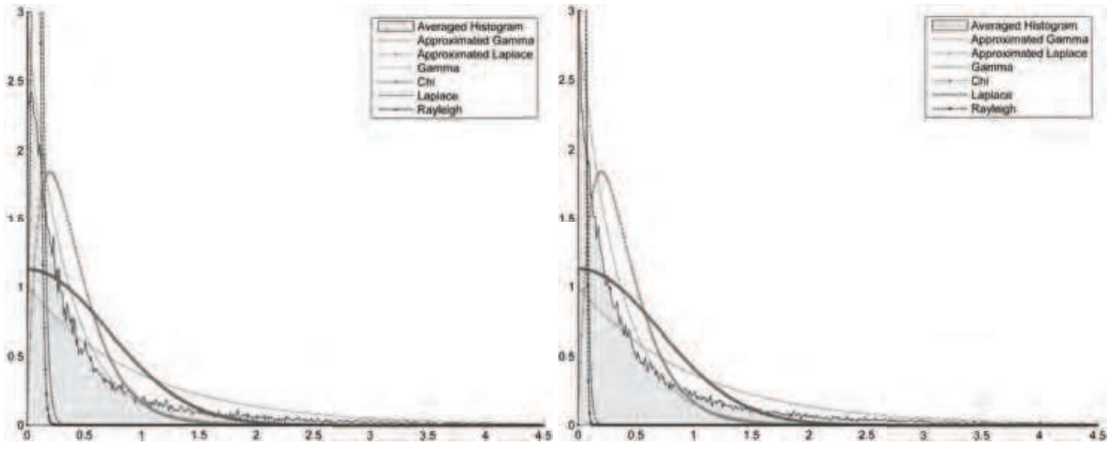


Fig. 3. Averaged Histogram and NLLS fits of mel-Frequency coefficients for the TiDigits (left) and WSJ database (right).

over the filter-bank channels. The Gamma PDF still provides the best model, even if the difference with other PDFs are more modest.

The modeling of log-mel coefficients and MFCCs cannot be performed using the same technique employed above. In fact, the histograms of these coefficients, depicted in figure 4 and 5, reveal that their distributions are multimodal and cannot be modeled by means of unimodal distributions. Therefore, multimodal models, such as Gaussian mixture models (GMM) (Redner & Walker (1984)) are more appropriate in this task: finite mixture models and their typical parameter estimation methods can approximate a wide variety of PDFs and are thus attractive solutions for cases where single function forms fail. The GMM probability density function can be designed as a weighted sum of Gaussians:

$$p(x) = \sum_{c=1}^{C} \alpha_c \mathcal{N}(x; \mu_c, \Sigma_c), \quad \text{with } \alpha_c \in [0,1], \quad \sum_{c=1}^{C} \alpha_c = 1 \tag{22}$$

where α_c is the weight of the c-th component. The weight can be interpreted as a priori probability that a value of the random variable is generated by the c-th source. Hence, a GMM PDF is completely defined by a parameter list $\rho = \{\alpha_1, \mu_1, \Sigma_1, \ldots, \alpha_C, \mu_C, \Sigma_C\}$.
A vital question with GMM PDF's is how to estimate the model parameters ρ. In literature exists two principal approaches: maximum-likelihood estimation and Bayesian estimation. While the latter has strong theoretical basis, the former is simpler and widely used in practice. Expectation-maximization (EM) algorithm is an iterative technique for calculating maximum-likelihood distribution parameter estimates from incomplete data. The Figuredo-Jain (FJ) algorithm (Figueiredo & Jain (2002)) represents an extension of the EM which allows not to specify the number of components C and for this reason it has been adopted in this work. GMM obtained after FJ parameter estimation are shown in figure 4 and 5.

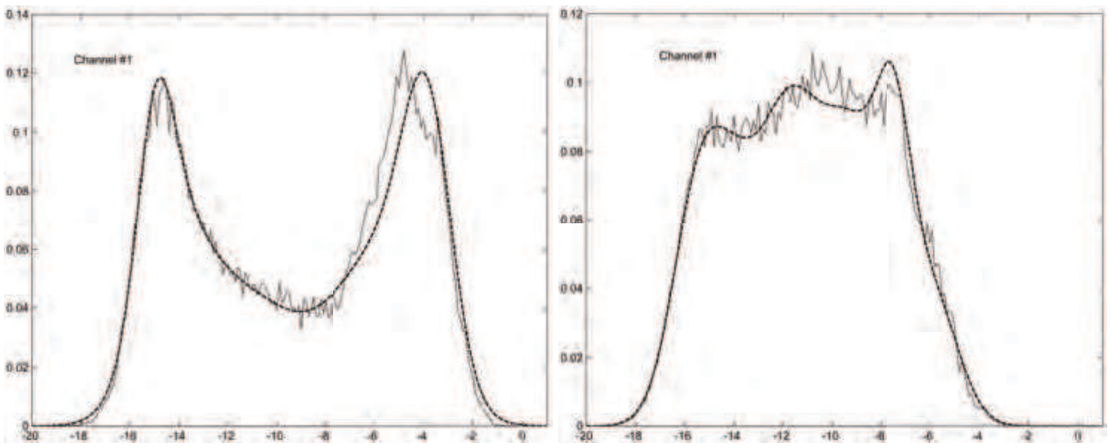


Fig. 4. Histogram (solid) and GMM fit (dashed) of the first channel of LogMel coefficients for TiDigits (left) and WSJ database (right).

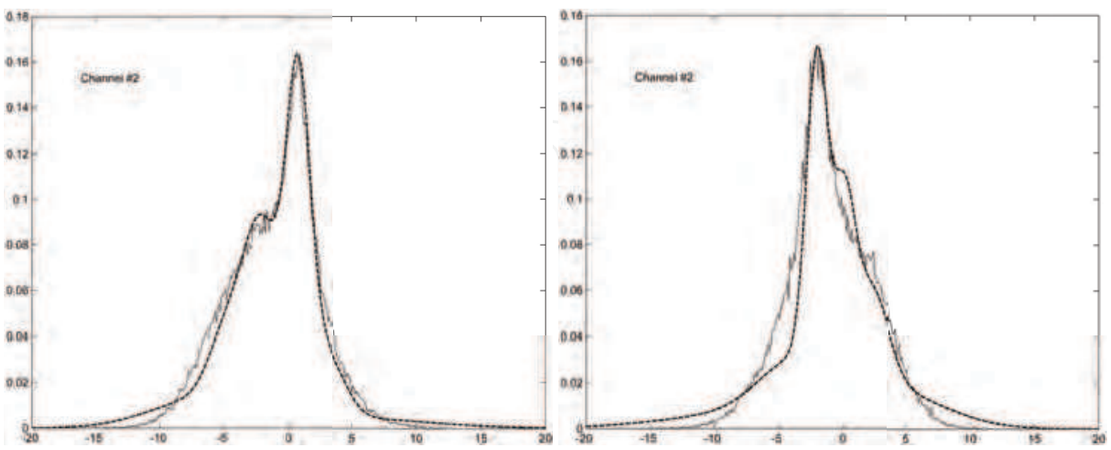


Fig. 5. Histogram (solid) and GMM fit (dashed) of the second channel of MFCC coefficients for TiDigits (left) and WSJ database (right).

4.2 Frequency domain multi-channel estimators

Let us consider a model of equation (4). It is assumed that the real and imaginary parts of both the speech and noise DFT coefficients have zero mean Gaussian distribution with equal variance. This results in a Rayleigh distribution for speech amplitudes A_i, and in Gaussian and Ricians distributions for $p(Y_i|A_i,\alpha_i)$ and $p(R_i|A_i)$ respectively. Such single-channel distributions are extended to the multi-channel ones by supposing that the correlation between the noise signals of different microphones is zero. This leads to

$$p(R_1,\ldots,R_M|A_n) = \prod_{i=1}^{M} p(R_i|A_n), \tag{23}$$

$$p(Y_1,\ldots,Y_M|A_n,\alpha_n) = \prod_{i=1}^{M} p(Y_i|A_n,\alpha_n), \tag{24}$$

$\forall n \in \{1,\ldots,M\}$. The model assumes also that the time delay between the microphones is small compared to the short-time stationarity of the speech. Thus, $A_i = c_i A_r$ and $\sigma^2_{X_i} = E[|X_i|^2] = c_i\sigma^2_X$ where c_i is a constant channel dependent factor. In addition

$$E[N_i N_j^*] = \begin{cases} \sigma^2_{N_i}, & i = j, \\ 0, & i \neq j. \end{cases}$$

These assumptions give the following probability density functions:

$$p(A_i,\alpha_i) = \frac{A_i}{\pi\sigma^2_{X_i}} \exp\left(-\frac{A_i^2}{\sigma^2_{X_i}}\right), \tag{25}$$

$$p(Y_1,\ldots,Y_M|A_n,\alpha_n) = \prod_{i=1}^{M} \frac{1}{\pi\sigma^2_{N_i}} \exp\left(-\sum_{i=1}^{M} \frac{|Y_i - (c_i/c_n)A_i e^{j\alpha_i}|^2}{\sigma^2_{N_i}}\right) \tag{26}$$

$$p(R_1,\ldots,R_M|A_n) = \exp\left(-\sum_{i=1}^{M} \frac{R_i^2 + (c_i/c_n)^2 A_n^2}{\sigma^2_{N_i}}\right) \prod_{i=1}^{M} \frac{2R_i}{\sigma^2_{N_i}} I_0\left(\frac{2(c_i/c_n)A_n R_i}{\sigma^2_{N_i}}\right). \tag{27}$$

where $\sigma^2_{X_i}$ and $\sigma^2_{N_i}$ are the variance of the clean speech and noise signals in channel i, and I_0 denotes the modified Bessel function of the first kind and zero-th order. As in (Ephraim & Malah (1984)), the a priori SNR $\xi_i = \sigma^2_{X_i}/\sigma^2_{N_i}$ and a posteriori SNR $\gamma_i = R_i^2/\sigma^2_{N_i}$ are used in the final estimators, and ξ_i is estimated using the decision directed approach.

4.2.1 Frequency domain multi-channel MMSE estimator (F-M-MMSE)

The multi-channel MMSE estimate of the speech spectral amplitude is obtained by evaluating the expression:

$$\hat{A}_i = E[A_i|Y_1,\ldots,Y_M] \qquad \forall i \in \{1,\ldots,M\}. \tag{28}$$

By mean of Bayes rule, and supposing that $\alpha_i = \alpha\ \forall i$, it can be shown (Lotter et al. (2003)) that the gain factor for channel i is given by:

$$G_i = \Gamma(1.5)\sqrt{\frac{\xi_i}{\gamma_i(1+\sum_{r=1}^{M}\xi_r)}} F_1\left(-0.5, 1, \frac{|\sum_{r=1}^{M}\sqrt{\gamma_r\xi_r}e^{j\phi_r}|^2}{1+\sum_{r=1}^{M}\xi_r}\right), \tag{29}$$

where F_1 denotes the confluent hypergeometric series, and Γ is the Gamma function.

4.2.2 Frequency domain multi-channel MAP estimator (F-M-MAP)

In (Lotter et al. (2003)), in order to remove the dependency from the direction of arrival (DOA) and obtain a closed-form solution, MAP estimator has been used. The assumption $\alpha_i = \alpha$ $\forall i \in \{1, \ldots, M\}$ is in fact only valid if $\beta_x = 0°$, or after perfect DOA correction. Supposing that the time delay of the desired signal is small respect to the short-time stationarity of speech, the noisy amplitudes R_i are independent from β_x.

MAP estimate was obtained extending the approach described in (Wolfe & Godsill (2003)). The estimate $\hat{A}_i$ of the spectral amplitude of the clean speech signal is given by

$$\hat{A}_i = \arg\max_{A_i} p(A_i | R_1, \ldots, R_M) \tag{30}$$

The gain factor for channel i is given by (Lotter et al. (2003)):

$$G_i = \frac{\sqrt{\xi_i / \gamma_i}}{2 + 2\sum_{r=1}^{M} \xi_r} \mathrm{Re}\left[\sum_{r=1}^{M} \sqrt{\gamma_r \xi_r} + + \sqrt{\left(\sum_{i=r}^{M} \sqrt{\gamma_r \xi_r} \right)^2 + (2 - M)\left(1 + \sum_{r=1}^{M} \xi_r \right)} \right]. \tag{31}$$

4.3 Feature domain multi-channel bayesian estimators

In this section the MMSE and the MAP estimators in the feature domain, recently proposed in (Principi, Rotili, Cifani, Marinelli, Squartini & Piazza (2010)), are presented. They extend the frequency domain multi-channel algorithms in (Lotter et al. (2003)) and the single-channel feature domain algorithm in (Yu, Deng, Droppo, Wu, Gong & Acero (2008)). Let assume again the model of section 2. As in (Yu, Deng, Droppo, Wu, Gong & Acero (2008)), for each channel i it is useful to define three artificial complex variables M_{x_i}, M_{y_i} and M_{n_i} that have the same modulus of m_{x_i}, m_{y_i} and m_{n_i} and phases θ_{x_i}, θ_{y_i} and θ_{n_i}. Assuming that the artificial phases are uniformly distributed random variables leads to consider M_{x_i} and $M_{y_i} - M_{n_i}$ as random variables following zero mean complex Gaussian distribution. High correlation between m_{x_i} of each channel is also supposed in analogy with the frequency domain model (Lotter et al. (2003)). This, again, results in $m_{x_i} = \lambda_i m_x$, with λ_i a constant channel dependent factor.

These statistical assumptions result in probability distributions similar to the frequency domain ones (Lotter et al. (2003)):

$$p\left(m_{x_i}, \theta_{x_i}\right) = \frac{m_{x_i}}{\pi \sigma_{x_i}^2} \exp\left[-\left(m_{x_i}\right)^2 / \sigma_{x_i}^2 \right], \tag{32}$$

$$p\left(M_{y_r} | m_{x_i}, \theta_{x_i}\right) = \frac{1}{\pi \sigma_{d_r}^2} \exp\left[-|\Psi|^2 / \sigma_{d_r}^2 \right], \tag{33}$$

where $\sigma_{x_i}^2 = E\left[|M_{x_i}|^2\right]$, $\sigma_{d_r}^2 = E\left[|M_{y_r} - M_{x_r}|^2\right]$, $\Psi = M_{y_r} - \Lambda_{ri} m_{x_i} e^{j\theta_{x_i}}$ and $\Lambda_{ri} = (\lambda_r / \lambda_i)^2$.

In order to simplify the notation, the following vectors can be defined:

$$\begin{aligned} \mathbf{c}_y(p) &= \left[c_{y_1}(p), \ldots, c_{y_M}(p)\right], \\ \mathbf{m}_y(b) &= \left[m_{y_1}(b), \ldots, m_{y_M}(b)\right], \\ \mathbf{M}_y(b) &= \left[M_{y_1}(b), \ldots, M_{y_M}(b)\right]. \end{aligned} \tag{34}$$

Each vector contains respectively the MFCCs, mel-frequency filter-bank outputs and artificial complex variables of all channels of the noisy signal $y(t)$. Similar relationships hold for the speech and noise signals.

4.3.1 Feature domain multi-channel MMSE estimator (C-M-MMSE)

The multi-channel MMSE estimator can be found by evaluating the conditioned expectation $\hat{c}_{x_i} = E\left[c_{x_i}|\mathbf{c}_y\right]$. As in the single-channel case, this is equivalent to (Yu, Deng, Droppo, Wu, Gong & Acero (2008)):

$$\widehat{m}_{x_i} = \exp\left\{E\left[\log m_{x_i}|\mathbf{m}_y\right]\right\} = \exp\left\{E\left[\log m_{x_i}|\mathbf{M}_y\right]\right\}. \tag{35}$$

Equation (35) can be solved using the moment generating function (MGF) for channel i:

$$\widehat{m}_{x_i} = \exp\left(\frac{d}{d\mu}\Phi_i(\mu)\Big|_{\mu=0}\right), \tag{36}$$

where $\Phi_i(\mu) = E\left[(m_{x_i})^{\mu}\,|\,\mathbf{M}_y\right]$ is the MGF for channel i. After applying Bayes rule, $\Phi_i(\mu)$ becomes:

$$\Phi_i(\mu) = \frac{\int_0^{+\infty}\int_0^{2\pi}(m_{x_i})^{\mu}\,p\left(\mathbf{M}_y\middle|m_{x_i},\theta_x\right)p\left(m_{x_i}|\theta_x\right)d\theta_x dm_{x_i}}{\int_0^{+\infty}\int_0^{2\pi}p\left(\mathbf{M}_y\middle|m_{x_i},\theta_x\right)d\theta_x dm_{x_i}}. \tag{37}$$

Supposing the conditional independence of each component of the $\mathbf{M}_y$ vector, we can write

$$p\left(\mathbf{M}_y\middle|m_{x_i},\theta_x\right) = \prod_{r=1}^{M} p\left(M_{y_r}\middle|m_{x_i},\theta_x\right), \tag{38}$$

where it was supposed that $\theta_{x_i} = \theta_x$, i.e. perfect DOA correction. The final expression of the MGF can be found by inserting (32), (33) and (38) in (37).
The integral over θ_x has been solved applying equation (3.338.4) in (Gradshteyn & Ryzhik (2007)), while the integral over m_{x_i} has been solved using (6.631.1). Applying (36), the final gain function $G_i(\xi_i,\gamma_i) = G_i$, for channel i is obtained:

$$G_i = \frac{\left|\sum_{r=1}^{M}\sqrt{\xi_r\gamma_r}e^{j\theta_{y_r}}\right|}{1+\sum_{r=1}^{M}\xi_r}\sqrt{\frac{\xi_i}{\gamma_i}}\exp\left(\frac{1}{2}\int_{v_i}^{+\infty}\frac{e^{-t}}{t}\,dt\right), \tag{39}$$

where

$$v_i = \frac{\left|\sum_{r=1}^{M}\sqrt{\xi_r\gamma_r}e^{j\theta_{y_r}}\right|^2}{1+\sum_{r=1}^{M}\xi_r}, \tag{40}$$

and $\xi_i = \sigma_{x_i}^2/\sigma_{n_i}^2$ is the a priori SNR and $\gamma_i = m_{y_i}^2/\sigma_{n_i}^2$ is the a posteriori SNR of channel i.
The gain expression is a generalization of the single-channel cepstral domain approach shown in (Yu, Deng, Droppo, Wu, Gong & Acero (2008)). In fact, setting $M = 1$ yields the single-channel gain function. In addition, equation (39) depends on the fictitious phase terms introduced to obtain the estimator. Uniformly distributed random values will be used during computer simulations.

4.3.2 Feature domain multi-channel MAP estimator (C-M-MAP)

In this section, a feature domain multi-channel MAP estimator is derived. The followed approach is similar to (Lotter et al. (2003)) in extending the frequency MAP estimator to the multi-channel scenario. The use of the MAP estimator is useful because the computational complexity can be reduced respect to the MMSE estimator and DOA independence can be achieved.

A MAP estimate of the MFCC coefficients of channel i can be found by solving the following expression:

$$\widehat{c}_{x_i} = \arg\max_{c_{x_i}} p(c_{x_i}|\mathbf{c}_y). \tag{41}$$

As in Section 4.3.1, MAP estimate on MFCC coefficients is equivalent to an estimate on mel-frequency filter-bank's output power. By means of Bayes rule, the estimate problem becomes

$$\widehat{m}_{x_i} = \arg\max_{m_{x_i}} p(\mathbf{m}_y|m_{x_i})p(m_{x_i}). \tag{42}$$

Maximization can be performed using (32) and knowing that

$$p(\mathbf{m}_y|m_{x_i}) = \exp\left\{-\sum_{i=1}^{M}\frac{m_{y_i}+(\lambda_i/\lambda_r)^2\,(m_{x_i})^2}{\sigma_{n_i}^2}\right\}\prod_{i=1}^{M}\left[\frac{2m_{y_i}}{\sigma_{n_i}^2}I_0\left(\frac{2(\lambda_i/\lambda_r)m_{y_i}m_{x_i}}{\sigma_{n_i}^2}\right)\right], \tag{43}$$

where conditional independence of m_{y_i} was supposed.
A closed form solution can be found if the modified Bessel function I_0 is approximated as $I_0(x) = (1/\sqrt{2\pi x})e^x$. The final gain expression is:

$$G_i = \frac{\sqrt{\xi_i/\gamma_i}}{2+2\sum_{r=1}^{M}\xi_r}\cdot \mathrm{Re}\left[\sum_{r=1}^{M}\sqrt{\xi_r\gamma_r}+\sqrt{\left(\sum_{r=1}^{M}\sqrt{\xi_r\gamma_r}\right)^2+(2-M)\left(1+\sum_{r=1}^{M}\xi_r\right)}\right]. \tag{44}$$

5. Multi-channel histogram equalization

As shown in the previous sections, feature enhancement approaches improve the test signals quality to produce features closer to the clean training ones. Another important class of feature enhancement algorithms is represented by statistical matching methods, according to which feature are normalized through suitable transformations with the objective of making the noisy speech statistics as much close as possible to the clean speech one. The first attempt in this sense has been made with CMN and cepstral mean and variance nomalization (CMVN) (Viikki et al. (2002)). They employ linear transformations that modify the first two moments of noisy observations statistics. Since noise induces a nonlinear distortion on signal feature representation, other approaches oriented to normalize higher-order statistical moments have been proposed (Hsu & Lee (2009); Peinado & Segura (2006)).
In this section the focus is on those methods based on histogram equalization (Garcia et al. (2009); Molau et al. (2003); Peinado & Segura (2006)): it consists in applying a nonlinear transformation based on the clean speech cumulative density function (CDF) to the noisy statistics. As recognition results confirm, the approach is extremely effective but suffers of some drawbacks, which motivated the proposal of some different variants in the literature. One important issue to consider is that the estimation of noisy speech statistics cannot usually rely on sufficient amount of data.
Up to the author's knowledge, no efforts have been put to employ the availability of multichannel acoustic information, coming from a microphone array acquisition, to augment the amount of useful data for statistics modeling and therefore improve the HEQ performances. Such a lack motivated the present work, where original solutions to combine multichannel audio processing and HEQ at a feature-domain level are advanced and experimentally tested.

5.1 Histogram equalization

Histogram equalization is the natural extension of CMN and CVN. Instead of normalizing only a few moments of the MFCCs probability distributions, histogram equalization normalizes all the moments to the ones of a chosen reference distribution. A popular choice for the reference distribution is the normal distribution.

The problem of finding a transformation that maps a given distribution in a reference one is difficult to handle and it does not have a unique solution in the multidimensional scenario. For the mono-dimensional case an unique solution exists and it is obtained by coupling the original and transformed CDFs of the reference and observed feature vectors.

Let y be a random variable with probability distribution $p_y(y)$. Let also x be a random variable with probability distribution $p_x(x)$ such that $x = T_y(y)$, where $T_y(\cdot)$ is a given transformation. If $T_y(\cdot)$ is invertible, it can be shown that the CDFs $C_y(y)$ and $C_x(x)$ of y and x respectively coincide:

$$C_y(y) = \int_{-\infty}^{y} p_y(v)\partial v = \int_{-\infty}^{x=T_y(y)} p_x(v)\partial v = C_x(x). \tag{45}$$

From equation (45), it is easy to obtain the expression of $x = T_y(y)$ from the CDFs of observed and transformed data:

$$C_y(y) = C_x(x) = C_x(T_y(y)), \tag{46}$$

$$x = T_y(y) = C_x^{-1}(C_y(y)). \tag{47}$$

Finally, the relationship between the probability distributions can be obtained from equation (47):

$$\begin{aligned} p_y(y) &= \frac{\partial C_y(y)}{\partial y} = \frac{\partial C_x(T_y(y))}{\partial y} = \\ &= p_x(T_y(y))\frac{\partial T_y(y)}{\partial y} = p_x(x)\frac{\partial T_y(y)}{\partial y}. \end{aligned} \tag{48}$$

Since $C_y(y)$ and $C_x(x)$ are both non-decreasing monotonic functions, the resulting transformation will be a non-linear monotonic increasing function (Segura et al. (2004)).

The CDF $C_x(x)$ can be obtained from the histograms of the observed data. The histogram of every MFCC coefficient is created partitioning the interval $[\mu - 4\sigma, \mu + 4\sigma]$ into 100 uniformly distributed bins B_i, $i = 1, 2, \ldots, 100$, where μ and σ are respectively the mean and standard deviation of the MFCC coefficient to equalize (Segura et al. (2004)). Denoting with Q the number of observations, the PDF can be approximated by its histogram as:

$$p_y(y \in B_i) = \frac{q_i}{Qh} \tag{49}$$

and the CDF as:

$$C_y(y_i) = C_y(y \in B_i) = \sum_{j=1}^{i} \frac{q_j}{Q}, \tag{50}$$

where q_i is the number of observations in the bin B_i and $h = 2\sigma/25$ is the bin width. The center y_i of every bin is then transformed using the inverse of the reference CDF function, i.e. $x = C_x^{-1}(y_i)$. The set of values (y_i, x_i) defines a piecewise linear approximation of the desired transformation. Transformed values are finally obtained by linear interpolation of such tabulated values.

5.2 Multi-channel histogram equalization

One of the well-known problems in histogram equalization is represented by the fact that there is a minimum amount of data per sentence necessary to correctly calculate the needed cumulative densities. Such a problem exists both for reference and noisy CDFs and it is obviously related to the available amount of speech to process. In the former case, we can use the dataset for acoustic model training: several results in literature (De La Torre et al. (2005); Peinado & Segura (2006)) have shown that Gaussian distribution represents a good compromise, specially if the dataset does not provide enough data to suitably represent the speech statistics (as it occurs for Aurora 2 database employed in our simulations). In the latter, the limitation resides in the possibility of using only the utterance to be recognized (like in command recognition task), thus introducing relevant biases in the estimation process. In conversational speech scenarios, is possible to consider a longer observation period, but this inevitably would have a significant impact not only from the perspective of computational burden but also and specially in terms of processing latency, not always acceptable in real-time applications. Of course, the amount of noise presence makes the estimation problem more critical, likely reducing the recognition performances.

The presence of multiple audio channels can be used to alleviate the problem: indeed occurrence of different MFCC sequences, extrapolated by the ASR front-end pipelines fed by the microphone signals, can be exploited to improve the HEQ estimation capabilities. Two different ideas have been investigated on purpose:

- MFCC averaging over all channels;
- alternative CDF computation based on multi-channel audio.

Starting from the former, it is basically assumed that the noise captured by microphones is highly incoherent and far-field model with DOA equal to 0° applied to speech signal (see section 6); therefore it is reasonable to suppose of reducing its variance by simply averaging over the channels.

Consider the noisy MFCC signal model (Moreno (1996)) for the i-th channel

$$\mathbf{y}_i = \mathbf{x} + \mathbf{D}\log(1 + \exp(\mathbf{D}^{-1}(\mathbf{n}_i - \mathbf{x})), \tag{51}$$

where $\mathbf{D}$ is the discrete cosine transform matrix and $\mathbf{D}^{-1}$ its inverse: it can be easily shown that the averaging operation reduces the noise variance w.r.t the speech one, thus resulting in an SNR increment. This allows the subsequent HEQ processing, depicted in figure 6, to improve its efficiency.

Coming now to the alternative options for CDF computation, the multi-channel audio information availability cab be exploited as follows (figure 7):

1. histograms are obtained independently for each channel and then all results averaged (CDF Mean);
2. histograms are calculated on the vector obtained concatenating the MFCC vectors of each channel (CDF Conc).

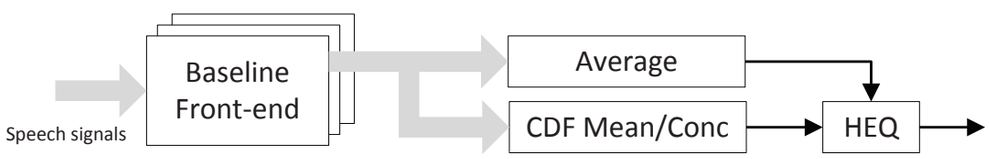


Fig. 7. HEQ MFCCmean CDF mean/conc: HEQ based on averaged MFCCs and mean of CDFs or concatenated signals.

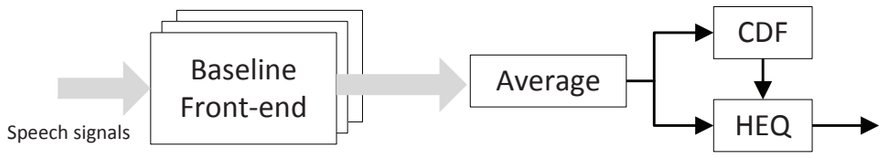
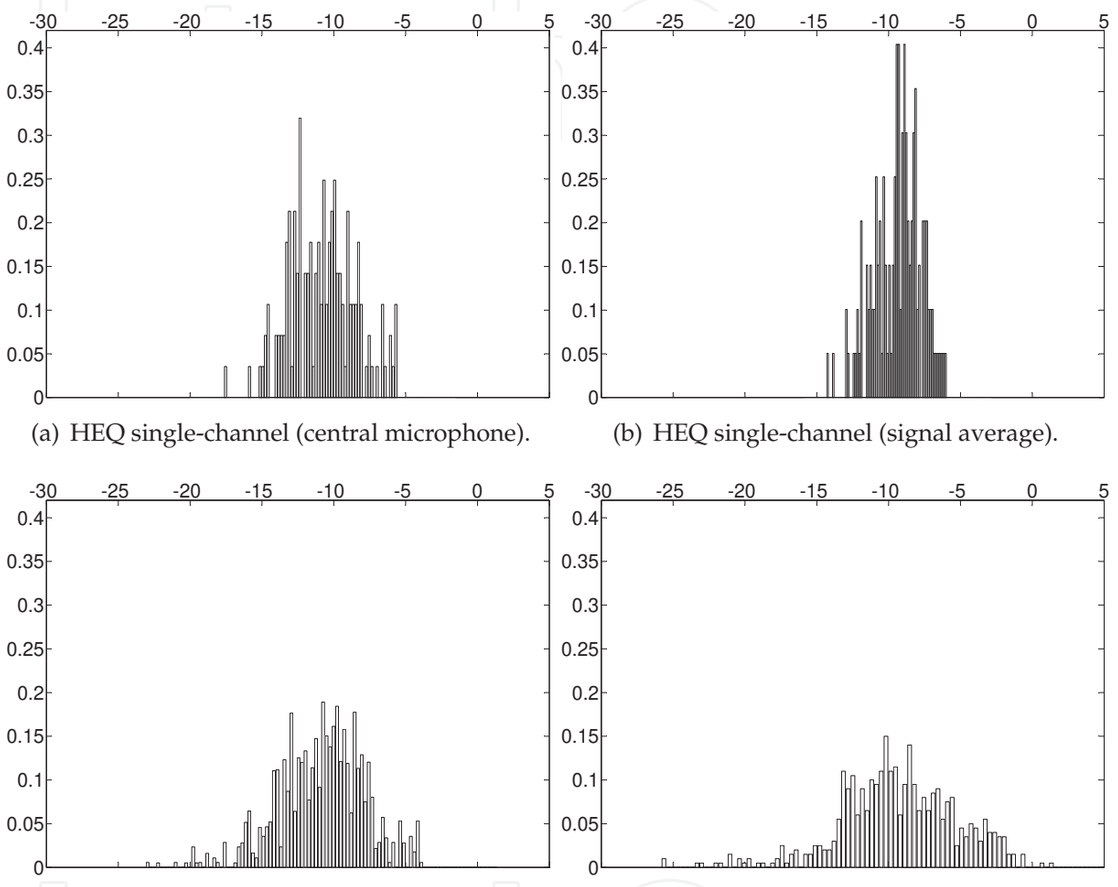


Fig. 6. HEQ MFCCmean: HEQ based on averaged MFCCs.



(a) HEQ single-channel (central microphone).

(b) HEQ single-channel (signal average).

(c) CDF mean approach.

(d) CDF conc approach.

Fig. 8. Histograms of cepstral coefficient c_1 related utterance FAK_5A corrupted with car noise at SNR 0 dB. CDF mean and CDF con histograms are estimated using four channels.

The two approaches are equivalent if the bins used to build the histogram coincide. However, in the CDF Mean approach, taking the average of the bin centers as well, gives slightly smoother histograms which helps the equalization process. Whatever the estimation algorithm, equalization has to be accomplished taking into account that the MFCC sequence used as input in the HEQ transformation must fit the more accurate statistical estimation performed, otherwise outliers occurrence due to noise contribution could degrade the performance: this explains the usage of the aforementioned MFCC averaging.
Figure 8 shows histograms of single-channel and multi-channel approaches of the first cepstral coefficient using four microphones in far-field model. Bins are calculated as described in section 5.1. A short utterance of length 1.16 s has been chosen to emphasize the difference

in histogram estimation in single and multi-channel approaches. Indeed, histograms of multi-channel configurations depicted in figure 7 better represent the underlying distribution (figure 8(c)-(d)) specially looking at the distribution tails, not properly rendered by the other approaches. This is due to availability of multiple signals corrupted by incoherent noise, which augments the observations available for the estimation of noisy feature distributions. Such a behavior is particularly effective at low SNRs, as recognition results in section 6 will demonstrate.
Note that operations described above are done independently for each cepstral coefficient: such an assumption is widely accepted and used among scientist working with statistics normalization for robust ASR.

6. Computer Simulations

In this section the computer simulations carried out to evaluate the performance of the algorithms previously described are reported. The work done in (Lotter et al. (2003)) has been taken as reference: simulations have been conducted considering the source signal in far-field model (see equation (2)) with respect to an array of $M = 4$ microphones with distance $d = 12$ cm. The source is located at 25 cm from the microphone array. The near-field and reverberant case studies will be considered in future works.
Three values of θ_x have been tested: 0°, 10° and 60°. Delayed signals have been obtained by suitably filtering the clean utterances of tests A, B and C of the Aurora 2 database (Hirsch & Pearce (2000)). Subsequently, noisy utterances in test A, B and C were obtained from the delayed signals by adding the same noises of Aurora 2 test A, B and C respectively. For each noise, signals with SNR in the range of 0-20 dB have been generated using tools (Hirsch & Pearce (2000)) provided with Aurora 2.
Automatic speech recognition has been performed using the Hidden Markov Model Toolkit (HTK) (Young et al. (1999)). Acoustic models structure and recognition parameters are the same as in (Hirsch & Pearce (2000)). The feature vectors are composed of 13 MFCCs (with C0 and without energy) and their first and second derivatives. Acoustic model training has been performed in a single-channel scenario and applying each algorithm in its insertion point of the ASR front-end pipeline as described in section 2. "Clean" and "Multicondition" acoustic models have been created using the provided training sets.
For the sake of comparison, in table 3 are reported the recognition results using the baseline feature extraction pipeline and the DSB. In using DSB the exact knowledge of the DOAs which leads to a perfect signal alignment is assumed. Recalling the model assumption made in section 2, since the DSB performs the mean over all the channels it reduces the variance of the noise providing higher performance than the baseline case. The obtained results can be employed to better evaluate the improvement arising from the insertion of the feature enhancement algorithms presented in this chapter.

	Test A		Test B		Test C		A-B-C AVG	
	C	M	C	M	C	M	C	M
baseline ($\beta_x = 0^\circ$)	63.56	83.18	65.87	84.91	67.93	86.27	65.79	84.78
DSB	76.50	93.12	79.86	94.13	81.47	94.96	79.27	94.07

Table 3. Results for both baseline feature-extraction pipeline and DSB

6.1 Multi-channel bayesian estimator

Tests have been conducted on algorithms described in Sections 4.2 and 4.3, as well as on their single-channel counterpart. The results obtained with the log-MMSE estimator (LSA) and its cepstral extension (C-LSA), and those obtained with frequency and feature domain MAP single-channel estimators are also reported for comparison purpose.

Frequency domain results in table 4 show as expected that the multi-channel MMSE algorithm gives the best performance when $\beta_x = 0°$, while accuracy degrades as β_x increases. Results in table 5 confirm the DOA independence of multi-channel MAP: averaging on β_x and acoustic models, recognition accuracy is increased of 11.32% compared to the baseline feature extraction pipeline. Good performance of multi-channel frequency domain algorithms confirm the segmental SNR results in (Lotter et al. (2003)).

On clean acoustic model, feature domain multi-channel MMSE algorithm gives a recognition accuracy around 73% regardless of the value of β_x (table 6). Accuracy is below the single-channel MMSE algorithm, and differently from its frequency domain counterpart it is DOA independent. This behaviour is probably due to the presence of artificial phases in the gain expression. The multi-channel MAP algorithm is, as expected, independent of the value of β_x, and while it gives lower accuracies respect to F-M-MMSE and F-M-MAP algorithms, it outperforms both the frequency and feature domain single-channel approaches (table 7).

	Test A		Test B		Test C		A-B-C AVG	
	C	M	C	M	C	M	C	M
F-M-MMSE ($\beta_x = 0°$)	84.23	93.89	83.73	92.19	87.10	94.71	85.02	93.60
F-M-MMSE ($\beta_x = 10°$)	80.91	92.61	81.10	91.19	84.78	93.78	82.26	92.53
F-M-MMSE ($\beta_x = 60°$)	70.83	88.29	71.84	86.50	76.68	91.67	73.12	88.82
LSA	76.83	87.02	77.06	85.24	78.97	88.48	77.62	86.91

Table 4. Results of frequency domain MMSE-based algorithms

	Test A		Test B		Test C		A-B-C AVG	
	C	M	C	M	C	M	C	M
F-M-MAP ($\beta_x = 0°$)	82.52	89.62	82.13	88.29	86.11	91.30	83.59	89.73
F-M-MAP ($\beta_x = 10°$)	82.20	89.46	81.93	88.00	85.84	90.39	83.32	89.28
F-M-MAP ($\beta_x = 60°$)	82.39	89.36	82.07	88.05	86.13	90.38	83.53	89.26
MAP	75.95	84.97	76.29	82.81	77.75	85.72	76.66	84.44

Table 5. Results of frequency domain MAP-based algorithms

	Test A		Test B		Test C		A-B-C AVG	
	C	M	C	M	C	M	C	M
C-M-MMSE ($\beta_x = 0°$)	70.80	89.94	73.00	88.75	75.37	92.02	72.96	90.23
C-M-MMSE ($\beta_x = 10°$)	70.40	89.68	72.88	88.72	75.21	91.89	72.83	90.10
C-M-MMSE ($\beta_x = 60°$)	70.72	89.69	72.77	88.80	75.19	91.93	72.89	90.14
C-LSA	75.68	87.81	77.06	86.85	76.94	89.25	76.56	87.97

Table 6. Results of feature domain MMSE-based algorithms

	Test A		Test B		Test C		A-B-C AVG	
	C	M	C	M	C	M	C	M
C-M-MAP ($\beta_x = 0°$)	78.52	91.51	79.28	89.99	81.63	93.04	79.81	91.51
C-M-MAP ($\beta_x = 10°$)	78.13	91.22	79.04	89.94	81.59	92.68	79.52	91.28
C-M-MAP ($\beta_x = 60°$)	78.23	91.22	79.04	90.07	81.38	92.68	79.55	91.32
C-MAP	74.62	88.44	76.84	87.67	75.61	89.58	75.69	88.56

Table 7. Results of feature domain MAP-based algorithms

To summarize, computer simulations conducted on a modified Aurora 2 speech database showed the DOA independence of the C-M-MMSE algorithm, differently from its frequency domain counterpart, and poor recognition accuracy probably due to the presence of random phases in the gain expression. On the contrary, results of the C-M-MAP algorithm confirm, as expected, its DOA independence and show that it outperforms single-channel algorithms in both frequency and feature domain.

6.2 Multi-channel histogram equalization

Experimental results for all tested algorithmic configurations are reported in tables 8 and 9 in terms of recognition accuracy. Table 10 shows results for different values of β_x and number of channels for the MFCC CDF Mean algorithm: since the other configurations behave similarly, results are not reported. Focusing on "clean" acoustic model results, the following conclusions can be drawn:

- No significant variability with DOA is registered (table 8): this represents a remarkable result, specially if compared with the MMSE approach in (Lotter et al. (2003)) where such a dependence is much more evident. This means that no delay compensation procedure have to be accomplished at ASR front-end input level. A similar behaviour can be observed both in the multi-channel mel domain approach of (Principi, Rotili, Cifani, Marinelli, Squartini & Piazza (2010)), and in the frequency domain MAP approach of (Lotter et al. (2003)), where phase information is not exploited.
- Recognition rate improvements are concentrated at low SNRs (table 9): this can be explained by observing that the MFCC averaging operation significantly reduces the feature variability leading to computational problems in correspondence of CDF extrema values when nonlinear transformation (47) is applied.
- As shown in table 10, the average of MFCCs over different channels is beneficial when applied with HEQ: in this case we can also take advantage of the CDF averaging process or of the CDF calculation based on MFCC channel vectors concatenation. Note that the improvement is proportional to the number of audio channels employed (up to 10% of accuracy improvement w.r.t. the HEQ single-channel approach).

In the "Multicondition" case study, the MFCCmean approach is the best performing and improvements are less consistent than the "Clean" case but still significative (up to 3% of accuracy improvement w.r.t. the HEQ single-channel approach). For the sake of completeness, it must be said that similar simulations have been performed using the average on the mel coefficients, so before the log operation (see figure 1): the same conclusions as above can be drawn, even though performances are approximatively and on the average 2% less than those obtained with MFCC based configurations.
In both "Clean" and "Multicondition" case the usage of the DSB as pre-processing stage for the HEQ algorithm leads to a sensible performance improvement with regard to the only

single-channel HEQ. The configuration with the DSB and the single channel HEQ have been tested in order to compare the effect of averaging the channels in the time domain or in the MFCC domain. As shown in table 8, the DSB + HEQ outperform the HEQ MFCCmean CDFMean/CDFconc algorithms but it must be pointed out that in using the DSB a perfect DOAs estimation is assumed. In this sense the obtained results can be seen as reference for future implementations, where a DOA estimation algorithm is employed with the DSB.

(a) Clean acoustic model

	$\beta_x = 0^\circ$	$\beta_x = 10^\circ$	$\beta_x = 60^\circ$
HEQ MFCCmean	85.75	85.71	85.57
HEQ MFCCmean CDFMean	90.68	90.43	90.47
HEQ MFCCmean CDFconc	90.58	90.33	91.36
HEQ Single-channel	81.07		
DSB + HEQ Single-channel	92.74		
Clean signals	99.01		

(b) Multicondition acoustic model

	$\beta_x = 0^\circ$	$\beta_x = 10^\circ$	$\beta_x = 60^\circ$
HEQ MFCCmean	94.56	94.45	94.32
HEQ MFCCmean CDFMean	93.60	93.54	93.44
HEQ MFCCmean CDFconc	92.51	92.48	92.32
HEQ Single-channel	90.65		
DSB + HEQ Single-channel	96.89		
Clean signals	97.94		

Table 8. Results for HEQ algorithms: accuracy is averaged across Test A, B and C.

	0 dB	5 dB	10 dB	15 dB	20 dB	AVG
HEQ MFCCmean	66.47	82.63	89.96	93.72	95.96	85.74
HEQ MFCC CDFmean	73.62	89.54	95.09	97.02	98.18	90.69
HEQ MFCCmean CDFconc	72.98	89.42	95.23	97.16	98.14	90.58
HEQ Single-channel	47.31	76.16	89.93	94.90	97.10	81.78

Table 9. Recognition results for Clean acoustic model and $\beta_x = 0^\circ$: accuracy is averaged across Test A, B and C.

	2 Channels		4 Channels		8 Channels	
	C	M	C	M	C	M
0°	88.27	93.32	90.68	93.60	91.44	93.64
10°	87.97	93.18	90.39	93.44	91.19	93.46
60°	87.81	92.95	90.43	93.43	91.32	93.52

Table 10. Results for different values of β_x and number of channels for the HEQ MFCC CDFmean configuration. "C" denotes clean whereas "M" multi-condition acoustic models. Accuracy is averaged across Test A, B and C.

7. Conclusions

In this chapter, different multi-channel feature enhancement algorithms for robust speech recognition were presented and their performances have been tested by means of the Aurora 2 speech database suitably modified to deal with the multi-channel case study in a far-field acoustic scenario. Three are the approaches here addressed, each one operating at a different

level of the common speech feature extraction front-end, and comparatively analyzed: beamforming, bayesian estimators and histogram equalization.
Due to the far-field assumption, the only beamforming technique here addressed is the delay and sum beamformer. Supposing that the DOA is ideally estimated, DSB improves recognition performances both alone as well as coupled with single-channel HEQ. Future works will investigate DSB performances when DOA estimation is carried out by a suitable algorithm.
Considering bayesian estimators, the multi-channel feature-domain MMSE and MAP estimators extend the frequency domain multi-channel approaches in (Lotter et al. (2003)) and generalize the feature-domain single-channel MMSE algorithm in (Yu, Deng, Droppo, Wu, Gong & Acero (2008)). Computer simulations showed the DOA independence of the C-M-MMSE algorithm, differently from its frequency domain counterpart, and poor recognition accuracy probably due to the presence of random phases in the gain expression. On the contrary, results of the C-M-MAP algorithm confirm, as expected, its DOA independence and show that it outperforms single-channel algorithms both in frequency and feature-domain.
Moving towards the statistical matching methods, the impact of multi-channel occurrences of same speech source in histogram equalization has been also addressed. It has been shown that averaging both the cepstral coefficients related to different audio channels and the cumulative density functions of the noisy observations allow augmenting the equalization capabilities in terms of recognition performances (up to 10% of word accuracy improvement using clean acoustic model), with no need of worrying about the speech signal direction of arrival.
Further works are also intended to establish what happens in near-field and reverberant conditions. Moreover, the promising HEQ based approach could be extended to other histogram equalization variants, like segmental HEQ (SHEQ) (Segura et al. (2004)), kernel-based methods (Suh et al. (2008)) and parametric equalization (PEQ) (Garcia et al. (2006)), which the proposed idea can be effectively applied.
Finally, due to the fact of operating in different domains, it is possible to envisage of suitably merge the three approaches here addressed in a unique performing noise robust speech feature extractor.

8. References

Atal, B. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *the Journal of the Acoustical Society of America* 55: 1304.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech and Signal Processing* 27(2): 113–120.

Chen, B. & Loizou, P. (2007). A Laplacian-based MMSE estimator for speech enhancement, *Speech communication* 49(2): 134–143.

Cifani, S., Principi, E., Rocchi, C., Squartini, S. & Piazza, F. (2008). A multichannel noise reduction front-end based on psychoacoustics for robust speech recognition in highly noisy environments, *Proc. of IEEE Hands-Free Speech Communication and Microphone Arrays*, pp. 172–175.

Cohen, I. (2004). Relative transfer function identification using speech signals, *Speech and Audio Processing, IEEE Transactions on* 12(5): 451–459.

Cohen, I., Gannot, S. & Berdugo, B. (2003). An integrated real-time beamforming and postfiltering system for nonstationary noise environments, *EURASIP Journal on Applied Signal Processing* 11: 1064?1073.

Cox, H., Zeskind, R. & Owen, M. (1987). Robust adaptive beamforming, *Acoustics, Speech, and Signal Processing, IEEE Transactions on* 35: 1365–1376.

De La Torre, A., Peinado, A., Segura, J., Perez-Cordoba, J., Benítez, M. & Rubio, A. (2005). Histogram equalization of speech representation for robust speech recognition, *Speech and Audio Processing, IEEE Transactions on* 13(3): 355–366.

Deng, L., Droppo, J. & Acero, A. (2004). Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features, *IEEE Transactions on Speech and Audio Processing* 12(3): 218–233.
URL: *http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1288150*

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, *Acoustics, Speech and Signal Processing, IEEE Transactions on* 32(6): 1109–1121.

Ephraim, Y. & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33(2): 443–445.

Figueiredo, M. & Jain, A. (2002). Unsupervised learning of finite mixture models, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24(3): 381 –396.

Gales, M. & Young, S. (2002). An improved approach to the hidden Markov model decomposition of speech and noise, *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Vol. 1, IEEE, pp. 233–236.

Gannot, S., Burshtein, D. & Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech, *Signal Processing, IEEE Transactions on* 49(8): 1614–1626.

Gannot, S. & Cohen, I. (2004). Speech enhancement based on the general transfer function gsc and postfiltering, *Speech and Audio Processing, IEEE Transactions on* 12(6): 561–571.

Garcia, L., Gemello, R., Mana, F. & Segura, J. (2009). Progressive memory-based parametric non-linear feature equalization, *INTERSPEECH*, pp. 40–43.

Garcia, L., Segura, J., Ramirez, J., De La Torre, A. & Benitez, C. (2006). Parametric nonlinear feature equalization for robust speech recognition, *Proc. of ICASSP 2006*, Vol. 1, pp. I –I.

Garofalo, J., Graff, D., Paul, D. & Pallett, D. (1993). CSR-I (WSJ0) Complete, *Linguistic Data Consortium* .

Gazor, S. & Zhang, W. (2003). Speech probability distribution, *Signal Processing Letters, IEEE* 10(7): 204 – 207.

Gong, Y. (1995). Speech recognition in noisy environments: A survey, *Speech communication* 16(3): 261–291.

Gradshteyn, I. & Ryzhik, I. (2007). *Table of Integrals, Series, and Products, Seventh ed.*, Alan Jeffrey and Daniel Zwillinger (Editors) - Elsevier Academic Press.

Griffiths, L. & Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming, *Antennas Propagation, IEEE Transactions on* 30(1): 27–34.

Hendriks, R. & Martin, R. (2007). MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions, *Audio, Speech, and Language Processing, IEEE Transactions on* 15(3): 918–927.

Herbordt, W., Buchner, H., Nakamura, S. & Kellermann, W. (2007). Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming, *Audio, Speech and Language Processing, IEEE Transactions on* 15(4): 1340–1351.

Herbordt, W. & Kellermann, W. (2001). Computationally efficient frequency-domain combination of acoustic echo cancellation and robust adaptive beamforming, *Proc. of EUROSPEECH*.

Hirsch, H. & Pearce, D. (2000). The aurora experimental framework for the performance speech recognition systems under noise conditions, *Proc. of ISCA ITRW ASR, Paris, France*.

Hoshuyama, O., Sugiyama, A. & Hirano, A. (1999). A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, *Signal Processing, IEEE Transactions on* 47(10): 2677–2684.

Hsu, C.-W. & Lee, L.-S. (2009). Higher order cepstral moment normalization for improved robust speech recognition, *Audio, Speech, and Language Processing, IEEE Transactions on* 17(2): 205 –220.

Hussain, A., Chetouani, M., Squartini, S., Bastari, A. & Piazza, F. (2007). Nonlinear Speech Enhancement: An Overview, *in* Y. Stylianou, M. Faundez-Zanuy & A. Esposito (eds), *Progress in Nonlinear Speech Processing*, Vol. 4391 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 217–248.

Hussain, A., Cifani, S., Squartini, S., Piazza, F. & Durrani, T. (2007). A novel psychoacoustically motivated multichannel speech enhancement system, *Verbal and Nonverbal Communication Behaviours, A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro (Eds.), Lecture Notes in Computer Science Series, Springer Verlag* 4775: 190–199.

Indrebo, K., Povinelli, R. & Johnson, M. (2008). Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Using a Novel Distortion Model, *Audio, Speech, and Language Processing, IEEE Transactions on* 16(8): 1654–1661.

Jensen, J., Batina, I., Hendriks, R. & Heusdens, R. (2005). A study of the distribution of time-domain speech samples and discrete fourier coefficients, *Proceedings of SPS-DARTS 2005 (The first annual IEEE BENELUX/DSP Valley Signal Processing Symposium)*, pp. 155–158.

Leonard, R. (1984). A database for speaker-independent digit recognition, *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, Vol. 9, pp. 328 – 331.

Li, J., Deng, L., Yu, D., Gong, Y. & Acero, A. (2009). A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions, *Computer Speech & Language* 23(3): 389–405.

Lippmann, R., Martin, E. & Paul, D. (2003). Multi-style training for robust isolated-word speech recognition, *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, Vol. 12, IEEE, pp. 705–708.

Lotter, T., Benien, C. & Vary, P. (2003). Multichannel direction-independent speech enhancement using spectral amplitude estimation, *EURASIP Journal on Applied Signal Processing* pp. 1147–1156.

Lotter, T. & Vary, P. (2005). Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model, *EURASIP Journal on Applied Signal Processing* 2005: 1110–1126.

McAulay, R. & Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter, *Acoustics, Speech and Signal Processing, IEEE Transactions on* 28(2): 137 – 145.

Molau, S., Hilger, F. & Ney, H. (2003). Feature space normalization in adverse acoustic conditions, *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, Vol. 1, IEEE.

Moreno, P. (1996). *Speech recognition in noisy environments*, PhD thesis, Carnegie Mellon University.

Omologo, M., Matassoni, M., Svaizer, P. & Giuliani, D. (1997). Microphone array based speech recognition with different talker-array positions, *Proc. of ICASSP*, pp. 227–230.

Peinado, A. & Segura, J. (2006). Speech recognition with hmms, *Speech Recognition Over Digital Channels*, pp. 7–14.

Principi, E., Cifani, S., Rotili, R., Squartini, S. & Piazza, F. (2010). Comparative evaluation of single-channel mmse-based noise reduction schemes for speech recognition, *Journal of Electrical and Computer Engineering* 2010: 1–7.
URL: *http://www.hindawi.com/journals/jece/2010/962103.html*

Principi, E., Rotili, R., Cifani, S., Marinelli, L., Squartini, S. & Piazza, F. (2010). Robust speech recognition using feature-domain multi-channel bayesian estimators, *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 2670 –2673.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm, *SIAM Review* 26(2): 195–239.

Rotili, R., Principi, E., Cifani, S., Squartini, S. & Piazza, F. (2009). Robust speech recognition using MAP based noise suppression rules in the feature domain, *Proc. of 19th Czech & German Workshop on Speech Processing*, Prague, pp. 35–41.

Segura, J., Benitez, C., De La Torre, A., Rubio, A. & Ramirez, J. (2004). Cepstral domain segmental nonlinear feature transformations for robust speech recognition, *IEEE Signal Process. Lett.* 11(5).

Seltzer, M. (2003). *Microphone array processing for robust speech recognition*, PhD thesis, Carnegie Mellon University.

Shalvi, O. & Weinstein, E. (1996). System identification using nonstationary signals, *Signal Processing, IEEE Transactions on* 44(8): 2055–2063.

Squartini, S., Fagiani, M., Principi, E. & Piazza, F. (2010). Multichannel Cepstral Domain Feature Warping for Robust Speech Recognition, *Proceedings of WIRN 2010, 19th Italian Workshop on Neural Networks May 28-30, Vietri sul Mare, Salerno, Italy*.

Stouten, V. (2006). Robust automatic speech recognition in time-varying environments, *KU Leuven, Diss* .

Suh, Y., Kim, H. & Kim, M. (2008). Histogram equalization utilizing window-based smoothed CDF estimation for feature compensation, *IEICE - Trans. Inf. Syst.* E91-D(8): 2199–2202.

Trees, H. L. V. (2001). *Detection, Estimation, and Modulation Theory, Part I*, Wiley-Interscience.

Viikki, O., Bye, D. & Laurila, K. (2002). A recursive feature vector normalization approach for robust speech recognition in noise, *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Vol. 2, IEEE, pp. 733–736.

Wolfe, P. & Godsill, S. (2000). Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement, *Proc. of IEEE ICASSP*, Vol. 2, pp. 821–824.

Wolfe, P. & Godsill, S. (2003). Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement, *EURASIP Journal Applied Signal Processing* 2003: 1043–1051.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. & Woodland, P. (1999). *The HTK Book. V2.2*, Cambridge University.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y. & Acero, A. (2008). Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor, *Audio, Speech, and Language Processing, IEEE Transactions on* 16(5): 1061–1070.

Yu, D., Deng, L., Wu, J., Gong, Y. & Acero, A. (2008). Improvements on Mel-frequency cepstrum minimum-mean-square-error noise suppressor for robust speech recognition, *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*, IEEE, pp. 1–4.

Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen