Al-Wakeel *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:15
DOI 10.1186/s13636-015-0059-4

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Stereo-based histogram equalization for robust speech recognition

Randa Al-Wakeel[1*], Mahmoud Shoman[2], Magdy Aboul-Ela[1] and Sherif Abdou[2]

Abstract

Optimal automatic speech recognition (ASR) takes place when the recognition system is tested under circumstances identical to those in which it was trained. However, in the actual real world, there exist many sources of mismatches between the environment of training and the environment of testing. These sources can be due to the sources of noise that exist in real environments. Speech enhancement techniques have been developed to provide ASR systems with the robustness against the sources of noise. In this work, a method based on histogram equalization (HEQ) was proposed to compensate for the nonlinear distortions in speech representation. This approach utilizes stereo simultaneous recordings for clean speech and its corresponding noisy speech to compute stereo Gaussian mixture model (GMM). The stereo GMM is used to compute the cumulative density function (CDF) for both clean speech and noisy speech using a sigmoid function instead of using the order statistics that is used in other HEQ-based methods. In the implementation, we show two choices to apply HEQ, hard decision HEQ and soft decision HEQ. The latter is based on minimum mean square error (MMSE) clean speech estimation. The experimental work shows that the soft HEQ and hard HEQ achieve better recognition results than the other HEQ approaches such as tabular HEQ, quantile HEQ and polynomial fit HEQ. It also shows that soft HEQ achieves notably better recognition results than hard HEQ. The results of the experimental work also show that using HEQ improves the efficiency of other speech enhancement techniques such as stereo piece-wise linear compensation for environment (SPLICE) and vector Taylor series (VTS). The results also show that using HEQ in multi style training (MST) significantly improves the ASR system performance.

Keywords: Robust speech recognition; Speech feature normalization; Histogram equalization; Speech enhancement

1 Introduction

Optimal automatic speech recognition (ASR) takes place when the recognition system is used under circumstances identical to those in which it was trained. However, in the actual real world, there exist many sources of mismatches between the environment of training and the environment of testing. These mismatches can be due to the sources of noise which include additive noise, linear filtering, and nonlinearities in transduction or transmission, as well as impulsive interfering. For this reason, robust speech recognition in noisy environments is one of the focus areas of speech research [1, 2]. The objective of robustness techniques is to provide ASR systems with robustness against the noise sources.

Most ASR systems are trained using training data recorded in typical clean conditions. However, in real environments, the noise introduces a distortion in the feature space and due to its random nature, it causes a loss of information [1]. It usually produces a nonlinear transformation of the feature space depending on the speech representation and the type of noise. For example, in the case of cepstral-based representations, additive noise causes a nonlinear transformation that has no significant effect on high-energy frames but a strong effect on those with energy levels in the same range or below that of the noise. Many speech enhancement techniques assume that the distortion is a function in clean speech and noise sources. Such that the noisy speech can be expressed as:

$$y_t = x_t + f(x_t, n_t, h_t)$$

* Correspondence: ra_roshel@yahoo.com
[1]Sadat Academy for management Science and Information Systems, Cairo, Egypt
Full list of author information is available at the end of the article

where t is the time frame index, x_t, y_t, and n_t are the clean, noisy, and additive noise Mel frequency cepstral coefficients (MFCC) vectors, respectively, and h_t is the corresponding convolutional noise vector. The random nature of the additive and convolutional noises results in one to many mappings between clean and noisy feature spaces and a given clean feature vector can generate different noisy vectors, and vice versa [3].

Figure 1 shows the effect of noise on the probability density function of clean speech when the noise is assumed stationary Gaussian noise [2]. The clean speech has a mean value of 25 and a standard deviation of 10. The noise has a standard deviation of 2. In general, the convolutional noise mainly shifts the mean of the coefficients, whereas additive noise modifies the probability density function (PDF), reducing the variance of the coefficients [3].

Speech enhancement techniques aim to achieve the robustness of the ASR systems in noisy environments. Speech enhancement methods can be classified into two main categories [2]. The first category is feature-domain-based methods which in turn have been classified into three approaches, noise resistant features, feature normalization, and feature compensation. The second main category is the model-domain methods. In noise-resistant feature approach, robust signal processing is employed to reduce the sensitivity of the speech features to the environment conditions that do not match those used to train the acoustic model. Examples of noise-resistant features include PLP [4, 5] and zero-crossing peak amplitude (ZCPA) [6], average localized synchrony detection (ALSD) [7], and perceptual minimum variance distortionless response (PMVDR) [8]. Although these features can usually achieve better performance than MFCC [9] which is popular in most of ASR systems, they have a much more complicated generation process which sometimes prevents them from being widely used together with some noise robustness technologies.

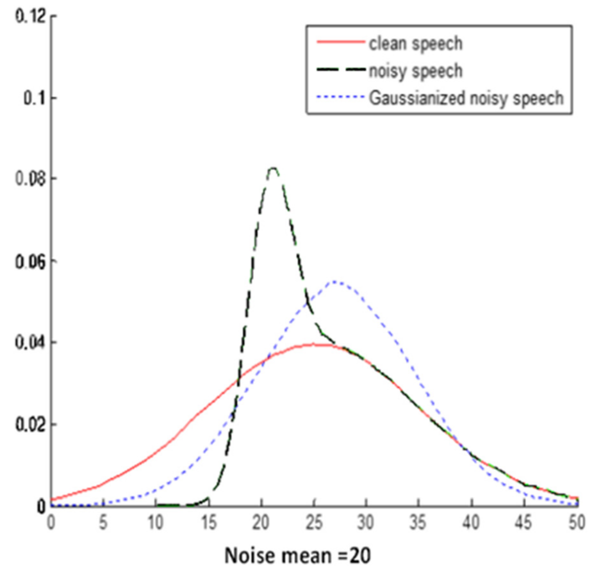

Fig. 1 The effect of noise on the probability density function

Feature moment normalization techniques normalize the statistical moments of speech features. Cepstral mean normalization (CMN) [10] is one of the most popular feature moment normalization techniques. It is involved in most speech recognition systems. It efficiently reduces the effects of unknown linear filtering in the absence of additive noise by subtracting the mean of the cepstral coefficients in order to remove the global shift of the mean in the cepstral vectors [1]. Another technique of moment normalization is cepstral mean and variance normalization (CMVN) [11]. CMVN normalizes the second-order statistical moments. Also, histogram equalization (HEQ) [1] is classified as one of this category of methods. It normalizes the higher statistical moments through the feature histogram.

The third subcategory of feature domain speech enhancement is feature compensation which aims to remove the effect of noise from the observed speech features. So, a clean version of the speech is recognized using the clean models. They include, but are not limited to, spectral subtraction (SS) [12], vector Taylor series (VTS) [13], stereo piece-wise linear compensation for environment (SPLICE) [14], and stereo-based stochastic vector mapping (SSM) [15, 16].

The second main category is the model-space methods [17, 18]. This class of methods attempts to modify the acoustic model parameters to incorporate the effects of noise in order to allow them to represent noisy speech properly. So the noisy speech is recognized using noisy models [19, 20]. Model space methods only adapt the model parameters to fit the distorted speech signal. The model adaptation can operate in either supervised or unsupervised mode. In supervised mode, the correct transcription of the adapting speech utterance is available. It is used to guide model adaptation to obtain adapted model which is used to decode the incoming utterances. In unsupervised model, the correct transcription is not available. Two-pass decoding is usually used. In the first pass, the initial model is used to decode the utterance to generate a hypothesis. And then the updated model is obtained and used to generate the final decoding result [2]. While typically achieving higher accuracy than feature-domain methods, they usually incur significantly higher computational cost. In addition, this class of methods needs a large amount of adaptation data. This may be difficult to be available in many situations.

Speech enhancement techniques may use a single channel recording such as VTS and spectral subtraction or multi-channel recordings such as SSM and SPLICE.

In this work, we are interested in the HEQ method used to compensate for the nonlinear distortions in speech representation. HEQ has a set of advantages. It can be applied in the feature domain, independently from the recognizer back end, without the need for a prior information about typical noise or the signal-to-noise ratio (SNR) to be expected and does not require voice activity detection (VAD) to estimate noise. Also, it is computationally inexpensive. In addition to these advantages, since it can be considered as a feature normalization technique, it can be applied before or after other speech enhancement techniques such as SPLICE and VTS approaches.

In this work, we utilize stereo simultaneous recordings for the clean and the corresponding noisy speech in order to train stereo Gaussian mixture model (GMM). The stereo GMM model is used to compute cumulative density function (CDF) tables for both clean and noisy speech. The CDF values are computed using sigmoid function. We aim to evince the effectiveness of the proposed HEQ method when it is applied either alone or in combination with VTS and SPLICE approaches. In addition, we aim to demonstrate its effectiveness when it is used in a process of multi-style training of the acoustic recognition models and using the resulted model to recognize speech processed by HEQ.

The paper is organized as follows. Section 2 reviews the main theory of HEQ. Three HEQ-based approaches are described. Section 3 introduces a speech enhancement approach based on HEQ. This method depends on the availability of stereo recordings for the training clean speech and its corresponding noisy speech. Two approaches to implement HEQ are presented, the hard decision HEQ and the soft decision HEQ. The experimental work and results are shown in Section 4, and finally, in Section 5, the conclusions and future work are presented.

2 Theory of histogram equalization

The acoustic mismatch between clean reference features and noisy test features caused by the environmental noise produces a statistical difference between their corresponding probability density functions (PDFs) [21]. HEQ attempts to transform the PDF of the original test (noisy) feature into its reference (or training) PDF [22–24] to improve the recognition accuracy. In the implementation of HEQ, reference and test PDFs are replaced by their corresponding reference histograms and the test histogram. The main objective is to find the transformation which achieves the equalization of the CDF of the noise observed coefficient to the CDF of the clean coefficient. HEQ is applied to each feature on a component-by-component basis. The principle of HEQ can be clarified using Fig. 2 [25].

Let the random variable y with pdf $p(y)$ and CDF $C_y(y)$ can be transformed to a random variable $\hat{x} = T_x(y)$ with a probability density function $\phi_x(\hat{x})$ and CDF $C_x(\hat{x})$ which is identical to that of the reference CDF, such as:

$$C_y(y) = C_x(\hat{x}) \tag{1}$$

So the estimated clean speech coefficient can be obtained by applying the inverse of the reference cumulative density function on the noisy CDF:

$$\hat{x} = C_x^{-1}\left(C_y(y)\right) \tag{2}$$

This process is assumed to transform the test data distribution into the training data distribution.

Clearly, the effectiveness of HEQ is directly related to the reliable estimation of both reference and test CDFs [26]. CDF can be estimated efficiently by using a large amount of sample data. This amount of data can be obtained at the training phase. So, the reference CDF can be obtained quite reliably. On the other hand, at the test phase, when short utterances are to be recognized, the amount of sample data may be insufficient for the reliable estimation of the test CDF. In this case, many researches [27, 24, 25] used the order statistics to compute the test CDF.

2.1 Computing CDF using order statistics

Let W be the test utterance to be recognized. W consists of N frames. Since HEQ works in a component-by-component basis, the kth component in all the N frames

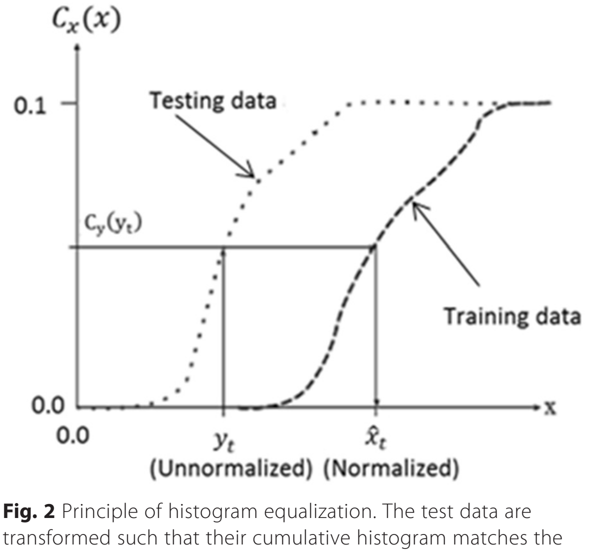

Fig. 2 Principle of histogram equalization. The test data are transformed such that their cumulative histogram matches the cumulative histogram of the training data distribution

is considered. This sequence of values of the *k*th component over all the N frames can be denoted as:

$$Y_k = \left[y_k(1), y_k(2), \ldots\ldots\ldots, y_k(N)\right] \tag{3}$$

The first step is to sort this sequence of values in ascending order such that [24]:

$$y_k([1]) \leq \ldots \leq y_k([r_k]) \leq \ldots \leq y_k([N]) \tag{4}$$

where $[r_k]$ and r_k denote the original frame index of feature component *y([r_k])*, and its rank (respectively) when the elements of the sequence Y_k are sorted in ascending order. The order statistics-based estimate of the test CDF is as follows:

$$C_{Y_k}\left(y_k[r_k]\right) = \frac{r_k - 0.5}{N} \tag{5}$$

This is the principle of HEQ speech enhancement approach. Speech enhancement methods based on HEQ have implemented HEQ in a variety of ways.

2.2 Variations of HEQ

We will briefly describe three effective HEQ-based approaches in this section.

2.2.1 Table-based histogram equalization (THEQ)

THEQ [28] is a non-parametric method to make the distributions of the test speech match those of the training speech. The cumulative histogram is used to estimate the corresponding CDF value of each feature vector component y. It has two phases. The first phase is the training phase, in which the cumulative histogram of each feature vector component of the training data is computed resulting in a table of the CDF values for a number of K bins between the maximum and minimum values of each feature component.

The second phase is the table look-up process in which the restored value of each feature vector component x of the test utterance is obtained by using its estimated CDF value as the key to finding the corresponding transformed (restored) value in the CDF table [29].

2.2.2 Quantile-based histogram equalization (QHEQ)

Instead of the full match of the cumulative histogram that is implemented by THEQ, QHEQ [30, 3] uses a quantile-corrective manner to calibrate the CDF of each feature vector component of the test speech to that of the training speech [29]. In QHEQ, a transformation function is estimated by minimizing the mismatch between the quantiles of the test utterance and those of the training data. The transformation function is then applied to each feature component x to make the CDF of the equalized component match that observed in training. The estimation of the transformation function can be implemented using a single test utterance (or extremely, a very short utterance), without the need of an additional set of adaptation data [3]. On the other hand, in order to find the optimum transformation parameters for each feature vector component, an exhaustive online grid search is needed. This search process is very time-consuming [29].

2.2.3 Polynomial-fit histogram equalization (PHEQ)

PHEQ uses the data fitting scheme to efficiently approximate the inverse functions of the CDFs of the training speech for HEQ [29]. Data fitting is a method for mathematical optimization. This method takes a series of data points (u_i,v_i) with i = 1,..., N and attempts to find a function $G(u_i)$ whose output $\tilde{v}_i$ closely approximates v_i by minimizing the sum of the squares error between the points $(u_i,\ \tilde{v}_i)$ and their corresponding points (u_i,v_i) in the data. N is the number of points. The function $G(u_i)$ to be estimated can be either linear or nonlinear in its coefficients.

Such data fitting (or so-called least squares regression) is used in PHEQ to estimate the inverse functions of the CDFs of the training speech. For each speech feature vector component of the training data, given the pair of the CDF value $C_{\mathrm{Train}}(y_i)$ of the vector component y_i and y_i itself, the linear polynomial function $G(C_{\mathrm{Train}}(y_i))$ with output $\tilde{y}_i$ can be expressed as:

$$G(C_{\mathrm{Train}}(y_i)) = \tilde{y}_i = \sum_{s=0}^{S} a_s (C_{\mathrm{Train}}(y_i))^s \tag{6}$$

where the coefficients a_s can be estimated by minimizing the squares error expressed in the following equation:

$$\begin{aligned} E^2 &= \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2 \\ &= \sum_{i=1}^{N} \left(y_i - \sum_{s=0}^{S} a_s (C_{\mathrm{Train}}(y_i))^s \right)^2 \end{aligned} \tag{7}$$

Here, N denotes the total number of training speech feature vectors, and S is the order of the polynomial function.

However, the polynomial function is efficient in constructing the transformation function, it has some limitations. High-order polynomial functions might cause over-fitting of the training data. In addition, the polynomial function provides good fits for input data points which exist within the range of values of the training data, but it would also probably have rapid deterioration when the input data points exist outside that range [29].

3 Histogram equalization using stereo databases

In this work, stereo training database (i.e., data that consists of simultaneous recordings of both the clean and noisy speech) is used to build stereo Gaussian mixture model (GMM) which is used to estimate both the reference CDF and the test CDF of each feature vector component. The proposed HEQ method uses CDF tables which are computed using stereo GMM which models the joint PDF of clean-noise speech. In addition, the method estimates the clean speech feature using a minimum mean square error (MMSE) criterion. Two approaches to apply HEQ, hard decision HEQ and soft decision HEQ, are investigated.

Stereo databases were used with HEQ in earlier work in [31]. In [31], the stereo database was used to train two separated GMM models, one to model the test environment and the other to model the clean speech. The two models are used to estimate the clean speech in a MMSE framework.

In this work, the stereo database is used to train a stereo GMM by concatenating each clean speech frame together with the corresponding noisy speech feature vector. Another difference is that cumulative density function tables for both clean and noisy speech are computed using the sigmoid function that utilizes the stereo GMM, so the order statistics is not used to compute the test CDF.

The following sections describe the details of implementing the proposed approach.

3.1 The joint probability distribution

The joint distribution is built using the stereo database. Let $x = (x_1, x_2,\ldots\ldots, x_K)$ be the clean feature vector where K is the dimension of the vector. The corresponding simultaneously recorded noisy representation is $y = (y_1, y_2,\ldots\ldots, y_K)$

Define $z \equiv (x, y)$ as the concatenation of the two channels. So the dimension of z is $2^* K$.

Gaussian mixtures are used to model the joint probability $p\ (z)$. So, $p\ (z)$ can be estimated as:

$$p(z) = \sum_{m=1}^{M} c_m N\left(z; \mu_{z,m}, \Sigma_{zz,m}\right) \tag{8}$$

where M is the number of mixture components, c_m, $\mu_{z,m}$, and $\Sigma_{zz,m}$ are the mixture weight, mean, and covariance of the mth component, respectively. Also, both the mean and covariance can be partitioned as

$$\mu_{z,m} = \begin{pmatrix} \mu_{x,m} \\ \mu_{y,m} \end{pmatrix} \tag{9}$$

$$\Sigma_{zz,m} = \begin{pmatrix} \Sigma_{xx,m} & \Sigma_{xy,m} \\ \Sigma_{yx,m} & \Sigma_{yy,m} \end{pmatrix} \tag{10}$$

The GMM model can be trained using the expectation maximization (EM) algorithm [32].

3.2 The use of sigmoid function to approximate CDF values

As mentioned above, the efficiency of HEQ approach is directly related to the reliable estimation of the CDF values. So, our objective now is to find a reliable method to estimate the CDF values given the stereo GMM.

In probability theory and statistics, the logistic distribution [33] is a continuous probability distribution resembles in its shape the Gaussian distribution and its cumulative distribution function has a similar shape as the CDF of the Gaussian distribution but it has heavier tails (higher kurtosis). The CDF of logistic distribution at the point x is computed using the following formula:

$$C_l(x) = \frac{1}{1 + \exp\left[\frac{-(x-\mu_l)}{\sigma_l}\right]} \tag{11}$$

where μ_l and σ_l are the mean and the standard deviation of the logistic distribution.

In this work, a formula similar to (11) is used to compute the CDF values using the mean and standard deviation of the stereo GMM.

3.3 HEQ using stereo GMM and sigmoid function

In HEQ, it is assumed that speech features are statistically independent, so it works in a component-by-component basis. The proposed HEQ approach is implemented in two phases. In the first phase, tables for the CDF values are computed using the stereo GMM both for the clean speech and noisy speech. In the second phase, the CDF tables are used to transform the observed noisy coefficient to its clean estimate.

1) Calculating the CDF tables: using each Gaussian mixture m, $1 \le m \le M$, and for each component k, $1 \le k \le K$, the cumulative histogram is estimated by considering a large number of bins V (e.g., 100) uniform intervals calculated between $\mu_m - 4\sigma_m$ and $\mu_m + 4\sigma_m$ where μ_m is the mean and σ_m is the standard deviation of the component at mixture m.

For each bin v, the CDF is calculated using the sigmoid function

$$Q_{v} = \frac{1}{1+\gamma \exp\left[\frac{v-\mu_m}{\sigma_m}\right]} \tag{12}$$

The parameter γ is a constant, and its value is chosen experimentally so the approximated HEQ transformation is smooth.

This process is implemented for both clean part and noisy part of the stereo GMM mixture. In the following, the notation Q_{v^c} is used to denote the CDF for the clean part and Q_{v^n} is used to denote the CDF of the noisy part. At the end of this phase, the CDF tables for clean and noisy parts of the GMM are obtained. The tables contain pairs of the bins and their corresponding CDF values. Each (bin, CDF) pair is denoted by (v, Qv).

4 Applying HEQ to the test speech

For each component y_k in the observed speech feature vector y, the clean speech estimate can be calculated using Gaussian mixture m as follows. The noise CDF table is used to detect the interval to which the component y_k belongs. So, we want to find the nearest two bins v_i^n and v_{i+1}^n to the component y_k, so $y_k \in \left[v_i^n,\ v_{i+1}^n\right]$.

The value of the cumulative density function at y_k $C(y_k)$ is computed as a linear interpolation of their CDF values as:

$$C(y_k) = \frac{Q_{v_i^n} + Q_{v_{i+1}^n}}{2} \tag{13}$$

where $1 \leq i \leq V$.

The computed CDF $C(y_k)$ is used to estimate the clean speech feature $\hat{x}$ using the clean CDF table, such that

$$\hat{x} = \frac{v_j^c + v_{j+1}^c}{2} \tag{14}$$

where v_j^c and v_{j+1}^c are the values of the bins whose CDF values are the nearest to $C(y_k)$.

However, in practice, there is no way to decide which mixture the observed speech vector y belongs to. Two choices were investigated:

1. Hard decision-based HEQ: in this approach, the mixture m whose $p(y|m)$ is maximum, is selected:

$$\hat{m} = \arg\max_{1\leq m\leq M} p(Y|m) \tag{15}$$

So, the HEQ is implemented using the CDF tables which are related to mixture $\hat{m}$.

2. Soft decision-based HEQ: in this approach, the HEQ algorithm is applied using all the mixtures. The clean speech estimate is computed in a minimum mean square error (MMSE) basis using

$$\hat{x} = E[x|y] \tag{16}$$

The main assumption is that the speech features are statistically independent. So, in the following, we will denote the speech feature as x for clean speech feature and y for noisy speech feature to indicate they are scalars. Considering the GMM structure of the joint distribution, Equation (16) can be further decomposed to:

$$\begin{aligned}
\hat{x} &= \int_x p(x|y)x dx = \sum_m \int_x p(x,m|y)x dx \\
\hat{x} &= \sum_m p(m|y)\int_x p(x,m|y)x dx \\
&= \sum_m p(m|y)\int_x p(x|m,y)x dx \\
&= \sum_m p(m|y)E[x|m,y]
\end{aligned} \tag{17}$$

Considering that $E[x|m,y]$ is equal to the clean speech estimate computed in (14), and setting it as,

$$\tilde{x}_m = E[x|m,y] \tag{18}$$

so we reach the following equation,

$$\tilde{x}_m = E[x|m,y] \tag{19}$$

where $p(m|y)$ is the posterior probability of the mixture m given the observed vector y. It can be computed by:

$$p(m|y) = \frac{p(y|m)p(m)}{\sum_{m=1}^{M} p(y|m)p(m)} \tag{20}$$

Soft-HEQ approach requires computing (14) M times, where M is the number of mixtures, while for hard HEQ, Equation (14) is computed only once (for the mixture with $p(Y|m)$ is maximum). So, for each component, soft HEQ requires $M(M + 1)$ multiplications and M^2 additions more than hard HEQ. This computational cost can be neglected when the number of mixtures is not large.

5 Experimental work and results

The experiments presented in this paper have been implemented using CARVUI database recorded inside a moving car. The data was collected in Bell Labs area, under various driving conditions (highway/city roads) and noise environments (with and without radio/music in the background). About two thirds of the recordings contain music or bubble noise in the background. A total of 56 speakers participated in the data collection. The speech material from 50 speakers is used for training, and the data from the 6 remaining speakers is used for test. Simultaneous recordings were made using a close-taking microphone and a 16-channel array of first-order hypercardiod microphones mounted on the visor. Data from two channels only are used. The first one is the close-talking microphone (CT) channel. The second one is a single channel from the microphone array, referred to as hands-free data (HF) henceforward. The average SNR is about 21 db for the CT channel and 8 db for the HF channel. The experiments were implemented using the part of the database that contains only the digit utterances.

The data are recorded at 24-kHz sampling rate and are down sampled to 8 kHz. They are then windowed with 12.5 ms frame rate and 25 ms frame size. Filter bank analysis is then applied. The filter bank has 26 channels. MFCCs are calculated by applying DCT to log filter bank amplitudes to produce 12 cepstral coefficients. The energy coefficients are also computed and appended to the cepstral vectors. Therefore, each cepstral vector contains 13 coefficients. During recognition, delta and delta-delta coefficients are computed on the fly resulting in vectors of 39 coefficients.

We have implemented speech enhancement using Gaussian mixtures models with sizes 16, 64, and 256 mixtures.

The recognition models for digits are trained using about 6500 training clean speech files collected from the CT microphone and tested using about 800 utterances. For each digit from 0 to 9, there is a hidden Markov model (HMM). The digit 0 has an additional model for the utterance “oh.” A 12th model is considered to model silence. Each of these models consists of 6 states. Each state contains 8 mixtures. The clean speech files are also used to build the clean speech Gaussian mixture models that were used in the experiments. Training and recognition is done using HTK [34].

In the HEQ experiments, constant γ used in the computation of the CDF in (12) takes the value −1.7.

A baseline set of results for this task is given in Table 1.

This table shows the evaluation of the recognition system in the different test/train conditions. In terms of sentence error rate (SER), Clean refers to the CT channel and noisy refers to the HF channel.

The first result shows the SER when clean speech is recognized using systems trained using clean database. The second result shows the result when the recognition system is trained and tested using noisy speech. The error rate is low in the two cases. This is because the environment of testing is similar to the environment of training. However, when the system is trained using clean speech and tested using noisy speech the recognition performance degrades severely. This is clear in the third experiment of Table 1 which shows how the system would perform in practical situations when the system is trained using clean speech and tested using observed speech. We can see dramatic degradation in performance due to the noise, 31.72 SER in noisy environment vs. 12.94 SER in clean environment.

Applying compensation techniques to the noisy speech improves the recognition results.

The SPLICE, VTS, and HEQ are applied to the MFCC coefficients before CMN. After applying the compensation, CMN was performed then the delta and delta-delta coefficients were computed.

Table 2 shows the results obtained by applying SPLICE and first-order vector Taylor series (VTS) technique.

We see that the best result obtained with the number of mixtures is 256. In this case, SPLICE represents about 13 % relative improvement to the baseline, and VTS represents about 9 % relative improvement.

The next set of experiments test the performance of the proposed hard and soft HEQ approaches in comparison with THEQ, QHEQ, and PHEQ approaches. Table 3 shows the results of these experiments.

From results in Table 3, we can see that the proposed approach outperformed the other HEQ approaches which did not provide significant improvement compared with the baseline system. Also, we can see the superiority of soft HEQ over the hard HEQ approach which achieved a relative 25.5 % reduction in SER over the baseline using 64 mixtures while the hard decision HEQ achieved only 9 % relative improvement in the SER using 256 mixtures.

The next set of experiments tested the effect of combining soft HEQ with other speech enhancement methods. Table 4 shows the results of applying SPLICE and VTS speech enhancement methods to the soft HEQ processed noisy speech.

From results shown in Table 4, we can see that applying SPLICE and VTS after HEQ achieves better results than the case when the SPLICE and VTS are applied directly on the noisy speech. Applying HEQ to the noisy speech before SPLICE introduces improvement to the performance of SPLICE by about 16 %. And in case of

Table 1 The base line recognition results

Condition	SER
Clean/Clean	12.94
Noisy/Noisy	16.79
Clean/Noisy	31.72

Table 2 SER after SPLICE and VTS speech enhancement

	16	64	256
SPLICE	29.85	28.48	27.49
VTS-first order	31.84	30.47	28.86

Table 3 SER after histogram equalization

	16	64	256
THEQ	31.43		
QHEQ	31.68		
PHEQ	30.39		
Hard HEQ	29.85	29.85	28.86
Soft HEQ	26.87	23.63	24.38

VTS, the improvement was about 13.6 % of the SER obtained by VTS alone.

In the last set of experiments, the soft HEQ was evaluated with the multi style trained (MST) recognition models. This is done in two steps:

a) Applying soft HEQ to the noisy training data to yield HEQ enhanced speech.
b) Constructing the training database by merging the clean speech database with the SSM-enhanced speech database. The new database is used to train the recognition models.

The resulted models are used to recognize HEQ processed noisy speech and noisy speech processed by THEQ, QHEQ, PHEQ, soft HEQ followed by VTS, and soft HEQ followed by SPLICE. Table 5 shows the results of these experiments.

The first row of Table 5 shows the SER obtained when the noisy unprocessed speech is recognized using multi-style trained recognition models.

The next three rows show the SER when the noisy speech is processed by THEQ, QHEQ, and PHEQ and then recognized by the multi-style trained HMM model. The last three rows show the SER when the soft HEQ is applied alone and with SPLICE and VTS speech enhancement methods.

THEQ, QHEQ, and PHEQ approaches achieve only 13.55, 13.46, and 15.5 % (respectively) improvements in the SER relative to the baseline. However, in case of 64 mixtures, the improvement in the SER achieved in case of MST when the noisy speech is processed by only soft HEQ was about 50 % relative to the baseline. When the noisy speech is processed by HEQ then by SPLICE the improvement was 48.6 % in case of 64 mixtures and this percentage is decreased when the number of mixtures increased to 256. When the noisy speech is processed by HEQ then by VTS, the improvement was about 39 %.

Table 4 SER SPLICE and VTS compensation applied on soft histogram equalized speech

	16	64	256
SPLICE	23.88	23.88	23.89
VTS-first order	29.44	26.34	27.11

Table 5 SER after HEQ evaluated by multi-style trained recognition models

	16	64	256
Noisy unprocessed	27.49		
THEQ	27.42		
QHEQ	27.45		
PHEQ	26.88		
Soft HEQ	16.42	15.8	17.41
Soft HEQ then VTS	20.27	19.47	19.4
Soft HEQ then SPLICE	16.29	16.29	25.88

The results of the experiments also show that the improvement achieved when the soft HEQ is implemented alone is better than the case when it is combined with SPLICE and VTS which provide a computationally an effective noise reduction approach.

6 Conclusions

In this paper, we proposed a speech enhancement-method based on HEQ. HEQ attempts to eliminate the nonlinear distortions of noise by transforming the PDF of the original noisy feature into its reference training PDF to improve the recognition performance. In this work, we used stereo speech recordings to build stereo GMM to model the joint probability of clean and noisy speech. The stereo GMM is used to compute the CDF tables using the sigmoid function. Two approaches to implement the HEQ method were investigated. The first is hard decision HEQ and the second is soft decision HEQ. The experimental work shows that soft decision HEQ notably achieves better speech recognition results than hard decision HEQ. Both soft HEQ and hard HEQ provided better performance than the other HEQ approaches such as THEQ, QHEQ, and PHEQ using clean speech trained and multi-style trained recognition models.

Also, the HEQ approach achieves better performance than SPLICE and VTS methods and applying HEQ to the noisy speech before SPLICE and VTS speech enhancement methods improves the performance of such enhancement methods but did not achieve better results than the results of applying HEQ alone.

The best recognition results are obtained when using number of mixtures in the GMM equal to 16. When used larger number of mixtures, the percentage of achieved performance improvement decreases. May using larger training data set provide better performance for larger number of mixtures?

Finally, more improvements to the recognition performance of ASR are obtained when the HEQ is used to enhance a set of noisy speech and incorporating this new speech data to the training speech corpus in a multi style training framework.

It is obvious that histogram normalization is computationally attractive as it does not require a process of noise estimation before applying the enhancement approach such as VTS. It also does not need the computation of correction vectors like SPLICE. In addition, since HEQ is implemented in a component-by-component manner, it does not require matrix operations in its implementation which is very time consuming. One additional advantage of HEQ is that it can be easily incorporated with most feature representations and other enhancement techniques without the need of any prior knowledge about the actual distortions caused by various kinds of noises.

On the other hand, HEQ has some limitations. The proposed HEQ approach makes use of large CDF tables which must be available during the testing of the recognition systems. This requires a large storage space.

Another limitation is that the HEQ operates in a component-by-component basis as it assumes that the MFCC features are independent. However, this assumption is used only for the simplicity and ease of implementation and in fact the MFCC features are correlated. Hence, there is a need to find a way to consider the correlation between the speech features in HEQ.

In this work, we have used sigmoid function to estimate CDF. Other methods of computing CDF can be developed and tested with HEQ approach.

Recently, deep neural network (DNN) [35, 36] has provided superior performance than GMM for speech recognition systems. So we suggest for future work to implement the proposed approach using DNN-HMM instead of GMM-HMM.

We also suggest for future work to estimate the clean speech using a maximum a posteriori (MAP) approach which has shown to outperform the MMSE approach.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Sadat Academy for management Science and Information Systems, Cairo, Egypt. [2]Faculty of Computers and Information, Information Technology Department, Cairo University, Cairo, Egypt.

Received: 30 August 2014 Accepted: 15 May 2015
Published online: 09 June 2015

References

1. A de la Torre, AM Peinado, JC Segura, JL Perez-Cordoba, MC Benitez, AJ Rubio, Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Speech Audio Process. **13**(3), 355–366 (2005)
2. J Li, L Deng, Y Gong, R Haeb-Umbach, An overview of noise-robust automatic speech recognition. T- IEEE T-ASLP **22**(4), 745–777 (2014)
3. F Hilger, H Ney, Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE T-ASLP **13**(3), 845–854 (2006)
4. H Hermansky, BA Hanson, H Wakita, Perceptually based linear predictive analysis of speech, in *Proc. ICASSP*, vol. Ith edn., 1985, pp. 509–512
5. H Hermansky, Perceptual linear predictive (PLP) analysis of speech. JASA **87**(4), 1738–1752 (1990)
6. DS Kim, SY Lee, RM Kil, Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE T-SAP **7**(1), 55–69 (1999)
7. AMA Ali, JV der Spiegel, P Mueller, Robust auditory based speech processing using the average localized synchrony detection. IEEE T-SAP **10**(5), 279–292 (2002)
8. UH Yapanel, JHL Hansen, A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition. Speech Commun. **50**(2), 142–152 (2008)
9. CK On, PM Pandiyan, Mel-frequency cepstral coefficient analysis in speech recognition, in *Computing & Informatics 2006, ICOCI'06, no. 2*, 2006, pp. 2–6
10. F Liu, RM Stern, XH Huang, A Acero, Efficient Cepstral Normalization for Robust speech Recognition, in *Proc. of ARPA Workshop on Human Language Technology*, 1993, pp. 69–74
11. O Viikki, D Bye, K Laurila, A recursive feature vector normalization approach for robust speech recognition in noise, in *Proc. ICASSP*, 1998, pp. 733–736
12. SF Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE T-ASSP **27**(2), 113–120 (1979)
13. PJ Moreno, *Speech recognition in noisy environments, PhD thesis* (Carnegie Mellon University, Pittsburgh, Pensilvania, 1996)
14. J Droppo, L Deng, A Acero, *Evaluation of the SPLICE Algorithm on the AURORA 2 Database* (Proc. Eurospeech, Denmark, 2001), pp. 217–220
15. Q Huo, D Zhu, A maximum likelihood training approach to irrelevant variability compensation based on piecewise linear transformations, in *Proc. Interspeech'06, Pittsburgh, Pennsylvania*, 2006, pp. 1129–1132
16. M Afify, X Cui, Y Gao, Stereo-based stochastic mapping for robust speech recognition, in *Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 7*, 2009, pp. 1325–1334
17. HY Jung, BO Kang, B Kangj, Y Lee, Model adaptation using discriminative noise adaptive training approach for new environments. ETRI J. **30**(6), 865–867 (2008)
18. X Wang, D O'Shaughnessy, Environmental independent ASR model adaptation/compensation by Bayesian parametric representation. IEEE Trans. Audio Speech Lang. Proc. **15**(4), 1204–1217 (2007)
19. J Wu, Q Huo, Supervised adaptation of MCE-trained CDHMMs using minimum classification error linear regression, in *Proc. ICASSP*, vol. Ith edn., 2002, pp. 605–608
20. X He, W Chou, Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs, in *Proc. ICASSP*, Ith edn., 2003, pp. 556–559
21. F Mihelič, J Zbert, Histogram equalization for robust speech recognition, in *Speech Recognition, Technologies and Applications* (I-Tech, Vienna, Austria, 2008), p. 550. ISBN 978-953-7619-29-9
22. L Buera, E Lleida, A Miguel, A Ortega, O Saz, Cepstral vector normalization based on stereo data for robust speech recognition, in *Audio, Speech, and Language Proc., IEEE Transactions on, vol. 15, no. 3*, 2007, pp. 1098–1113
23. Y Suh, SB Kwon, H Kim, Feature compensation with class-based histogram equalization for robust speech recognition. Paper presented at the 9th WESPAC 2006 conference, Korea, 26–28 June 2006
24. Y Suh, H Kim, Class-based histogram equalization for robust speech recognition. ETRI J. **28**, 502–505 (2006)
25. S Molau, M Pitz, H Ney, *Histogram-Based Normalization in the Acoustic feature space* (Proc ASRU, Trento, Italy, 2001), pp. 21–24
26. Y Suh, H Kim, Environmental model adaptation based on histogram equalization. Signal Process. Lett. IEEE **16**(4), 264–267 (2009)
27. X Xiao, J Li, E Chng, H Li, Maximum likelihood adaptation of histogram equalization with constraint for robust speech recognition. Paper presented at the 2011 ICASSP, Prague, 22–27 May 2011
28. S Dharanipragada, M Padmanabhan, A nonlinear unsupervised adaptation technique for speech recognition. Paper presented at the 6th proceedings of the international conference on spoken language processing (ICSLP 2000), Beijing, China, 2000
29. S-H Lin et al., A comparative study of histogram equalization (HEQ) for robust speech recognition. Comput. Linguist. Chin. Lang. Proc. **12**(2), 217–238 (2007)
30. F Hilger, H Ney, *Quantile based histogram equalization for noise robust speech recognition* (Proc EUROSPEECH, Aalborg, Denmark, 2001), pp. 1135–1138
31. L Buera, E Lleida, A Miguel, A Ortega, Multi-environment models based linear normalization for robust speech recognition, in *SPECOM '2004: 9th Conference Speech and Computer St. Petersburg, Russia*, 2004, pp. 20–22

32. J Bilmes, A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, in *Technical Report ICSI-TR-97-021, University of Berkeley*, 1998
33. SR Bowling, MT Khasawneh, S Kaewkuekool, BR Cho, A logistic approximation to the cumulative normal distribution. J. Indus. Eng. Manag. **2**(1), 114–127 (2009)
34. S Young, G Evermann, T Hain, D Kershaw, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book. Revised for HTK Version 3.4*, 2006. http://htk.eng.cam.ac.uk/
35. G Hinton, L Deng, D Yu, G Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, T Sainath, B Kingsbury, Deep neural networks for acoustic modeling in speech recognition, in *Signal Processing Magazine, IEEE, vol. 29, no. 6*, 2012, pp. 82–97
36. W Gevaert, G Tsenov, V Mladenov, Neural networks used for speech recognition. J. Automatic Control Univ. Belgrade **20**, 1–7 (2010)