# Probabilistic Class Histogram Equalization Based on Posterior Mean Estimation for Robust Speech Recognition

Youngjoo Suh, *Member, IEEE*, and Hoirin Kim, *Member, IEEE*

***Abstract*—In this letter, we propose a new probabilistic class histogram equalization technique for noise robust speech recognition. To cope with the sparse data problem which is common in the case of short test data, the proposed histogram equalization technique employs the posterior mean estimator, a kind of the Bayesian estimator, for test CDF. Experiments on the Aurora-4 framework showed that the proposed method produces performance improvement over the conventional maximum likelihood estimation-based approach.**

***Index Terms*—CDF estimation, feature normalization, histogram equalization, posterior mean, robust speech recognition.**

## I. Introduction

ROBUST speech recognition aims at providing speech recognition systems with robustness against the acoustic mismatch caused by additive noise and channel distortion, etc. It has been a long-standing and on-going research issue in the areas of automatic speech recognition (ASR). An easiest approach to robust speech recognition is feature normalization which normalizes the statistical moments of speech features which are corrupted by background noise and channel distortion [1]. Acoustic conditions corrupted by additive noise and channel distortion cause a nonlinear transform in logarithm-based feature spaces such as cepstrum and log filter-bank energy [2]. For this reason, the conventional linear transform-based feature normalization approaches such as cepstral mean normalization (CMN) [3] or cepstral mean and variance normalization (CMVN) [4] have fundamental limitations, even though they provides noticeable performance gains in noisy conditions. The histogram equalization (HEQ) technique [5]–[13] is an efficient nonlinear transformation-based feature normalization or model adaptation approach due to its algorithmic simplicity. In addition, it does not require any prior assumptions about noise process or the way the noise affects the speech model [6]. The basic idea of HEQ is to normalize the probability density functions (PDFs) between the training and test data. This normalization can be achieved by converting PDF of the test features into that of the training. By this approach, HEQ can compensate for the acoustic mismatch between training and test data. However, HEQ has some fundamental limitations when employed to the real-world ASR applications. To remedy these limitations, the probabilistic class HEQ (CHEQ) approach has been proposed [7], which equalizes different acoustic classes separately according to their corresponding class-specific distribution. Another issue in HEQ is to reliably estimate cumulative distribution function (CDF). In ASR, training CDF can be accurately approximated by its cumulative histogram. However, such approximations tend to be unreliable for short test utterances which are common in real-world environments. When the amount of data is insufficient, the order statistics based method can produce more accurate and reliable CDF [5]. A quantile-based HEQ was also proposed to provide robust estimation of test CDF by adjusting test CDF to training CDF with a set of quantiles [12]. Even though these methods can estimate test CDF more accurately, the resulting test CDF still suffers from poor estimation accuracy due to overfitting when the data are sparse. This overfitting problem gets worse in the case of CHEQ, where the amount of test data per class becomes smaller according to the number of total classes. The overfitting problem in the estimation of test CDF can be alleviated by using the Bayesian priors such as maximum *a posteriori* (MAP) or posterior mean (PM) estimation approaches [14]. Of the two methods, the PM approach can provide a more straightforward solution in coping with the sparse data problem in the case of histogram-based estimation where the Dirichlet-multinomial conjugate can be a mathematically tractable approach to modeling the histogram parameters.

In this paper, we propose probabilistic class HEQ employing the PM based test CDF estimation technique for feature normalization in ASR.

## II. Histogram Equalization

### A. HEQ

For training and test feature components $x$ and $y$, respectively, both of which are assumed to be random variables, let $p_X(x)$ and $p_Y(y)$ denote their corresponding PDFs. Histogram equalization aims to normalize test feature data by transforming PDF of test feature $y$ into that of training feature $x$ and is given in [5] as

$$\hat{x} = F(y) = C_X^{-1}[C_Y(y)] \tag{1}$$

Manuscript received May 18, 2015; revised August 10, 2015; accepted October 04, 2015. Date of publication October 13, 2015; date of current version October 19, 2015. This work was supported by Basic Science Research Program through the National Research Foundation of Korea under Grant NRF-2014R1A1A2055896. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaodong He.

The authors are with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea (e-mail: yjsuh@kaist.ac.kr; hoirkim@kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2015.2490202

where $C_X^{-1}(x)$ is the inverse of training CDF $C_X(x)$, and $C_Y(y)$ is the test CDF of feature $y$.

In practice, training CDF, $C_X(x)$, can be approximated by corresponding cumulative histograms with finite numbers of bins or a set of quantiles with the Gaussian distribution of a zero mean and unit variance. Test CDF, $C_Y(y)$, is approximated by the corresponding cumulative histogram or the order statistics-based CDF.

### B. Class HEQ

Although HEQ achieves the goal of feature normalization quite effectively for noise robust ASR, it still has some limitations in practice [7]. One major limitation of HEQ is that it needs phonetic or acoustic class distributions of training and test to be identical or similar to each other. The other is that the acoustic mismatch between training and test needs to be monotonic transformation. The former condition is not well met when test utterances are short. The latter tends to be violated by the corruption of random noise under noisy environments. An effective approach to overcoming these limitations of the original HEQ is the CHEQ technique [7], where test data are equalized by employing class-specific training and test CDFs as

$$\hat{x} = \sum\nolimits_{j=1}^{J} P(j|y) C_{X(j)}^{-1}[C_{Y(j)}(y)] \tag{2}$$

where $J$ denotes the number of classes, $C_{X(j)}^{-1}$ represents the inverse of the $j$th training CDF, $C_{Y(j)}$ stands for the $j$th test CDF, and $P(j|y)$ is the posterior probability of the $j$th class given test feature $y$, which is given by

$$P(j|y) = \frac{p(y|j)P(j)}{\sum_{j'=1}^{J} p(y|j')P(j')} \tag{3}$$

where $P(j)$ represents the prior probability of the $j$th class and $p(y|j)$ is the likelihood of $y$ for the $j$th class and can be given by the Gaussian model [7].

### C. ML-based CDF Estimation

The classical approach to estimating histogram is the bin-counting method, which can be derived by the maximum likelihood (ML) with a multinomial distribution as follows.

Suppose that $\hbar$ is the histogram bin into which feature component $y$ falls. Then, for a multinomial distribution with $\hbar$ being a categorical random variable with $K$ categories, i.e., histogram bins $\{\mathcal{H}_1, \cdots, \mathcal{H}_K\}$, let $P(\hbar = \mathcal{H}_k) = \theta_k$ be the probability of $\hbar$ being the $k$th histogram bin. The likelihood of $\hbar$ given the parameter set of histogram bins is represented by

$$P(\hbar|\theta) = \prod\nolimits_{k=1}^{K} \theta_k^{I(\hbar=\mathcal{H}_k)} \tag{4}$$

where the indicator function $I(\hbar = \mathcal{H}_k)$ has 1 if $\hbar = \mathcal{H}_k$ and 0 for otherwise. Then, when test feature sequence $S = \{y_1, \cdots, y_N\}$ is given, the likelihood of the corresponding bin set $\Xi = \{\hbar_1, \cdots, \hbar_N\}$ is given by

$$P(\Xi|\theta) = \prod\nolimits_{n=1}^{N} \prod\nolimits_{k=1}^{K} \theta_k^{I(\hbar_n=\mathcal{H}_k)} = \prod\nolimits_{k=1}^{K} \theta_k^{N_k} \tag{5}$$

with $N_k = \sum\limits_{n} I(\hbar_n = \mathcal{H}_k)$ denoting the number of times that the elements of $\Xi$ are the $k$th histogram bin with the condition of $N = \sum\limits_{k} N_k$.

Under this assumption, the ML estimate of $\theta_k$ is obtained from the derivation of the log-likelihood as

$$l(\theta) = \log P(\Xi|\theta) = \sum\nolimits_{k} N_k \log \theta_k \tag{6}$$

With the probability constraint $\sum\limits_{k} \theta_k = 1$, the Lagrange function can be derived by using a Lagrange multiplier

$$\tilde{l}(\theta) = \sum\nolimits_{k} N_k \log \theta_k + \lambda(1 - \sum\nolimits_{k} \theta_k) \tag{7}$$

By taking derivative with respect to $\theta_k$ and $\lambda$, separately, and using the constraint of $\sum\limits_{k} \theta_k = 1$, we have a result $\lambda = N$ which gives the ML estimate of $\theta_k$ as

$$\hat{\theta}_k^{ML} = \frac{N_k}{N} \tag{8}$$

Then, test CDF approximated by the ML-based cumulative histogram estimation is given by summing the estimates cumulatively as

$$\hat{C}_Y^{ML}(y) = \frac{1}{N} \sum\nolimits_{k'=1}^{k} N_{k'} \tag{9}$$

where $k$ satisfies the constraint of $y \in \mathcal{H}_k$, i.e., $\hbar = \mathcal{H}_k$. When assuming a large amount of training data are available in developing speech recognition systems, training CDF can be accurately obtained by estimating the cumulative histogram or directly utilizing a Gaussian with zero mean and unity variance only once in the training phase. On the contrary, test CDF needs to be estimated utterance-by-utterance or segment-by-segment in the test phase. Moreover, the CDF estimation suffers from the sparse data problem more severely when the test speech utterance gets shorter or the number of classes in CHEQ becomes larger. To cope with this sparse data problem, more efficient and reliable methods for the test CDF estimation need to be employed.

### D. Order Statistics-based CDF Estimation

An efficient approach to dealing with the sparse data problem in CDF estimation is the order statistics-based method [5], given as follows.

For random variable sequence $S$, its order statistic is defined as

$$y_{T(1)} \leq y_{T(2)} \leq \cdots \leq y_{T(r)} \leq \cdots \leq y_{T(N)} \tag{10}$$

where $T(r)$ denotes the original time frame index of $y$ in which its rank is $r$ when the elements of $S$ are sorted in ascending order. The order statistics-based asymptotically unbiased estimate of CDF is obtained by

$$\hat{C}_Y^{OS}(y_n) = \frac{r(y_n) - 0.5}{N} \tag{11}$$

where $r(y)$ stands for the rank of $y$ which ranges from 1 to $N$.

## III. CHEQ with Posterior Mean-based CDF Estimation

When the amount of data gets much smaller, the estimation accuracy of the two aforementioned CDF estimation methods also deteriorates due to overfitting of the test CDF mainly caused by the sparse test data. One remedy for this overfitting in CDF estimation is using the Bayesian priors such as the MAP or PM estimation methods. With the Bayesian priors, the histogram can be modeled by the Dirichlet-multinomial conjugate distributions as follows.

This paper previously published in IEEE Signal Processing Letters

When the histogram bin $\hbar$ of the corresponding feature component $y$ has a multinomial distribution in (4), its conjugate prior is the Dirichlet distribution defined by

$$p(\theta) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{12}$$

where $\theta_k$ satisfies the constraint of $\sum_k \theta_k = 1$, the hyper parameter $\alpha$ is a set of pseudo counts $\{\alpha_k\}$ and the normalizing constant $B(\alpha)$ is the multinomial Beta function given by

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{13}$$

with the gamma function defined as $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$. The posterior probability is given by

$$\begin{aligned} p(\theta|\Xi) &\propto p(\Xi|\theta) p(\theta|\alpha) \\ &= \frac{N!}{\prod_{k=1}^{K} N_k!} \prod_{k=1}^{K} \theta_k^{N_k} \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \end{aligned} \tag{14}$$

For the Dirichlet prior with parameter $\alpha$, the PM estimate of $\theta_k$ is given [14] as

$$\begin{aligned} \hat{\theta}_k^{PM} &= \int \theta_k p(\theta|\Xi) d\theta = \frac{N_k + \alpha_k}{N + \sum_{k'=1}^{K} \alpha_{k'}} \\ &= (1 - \eta) \hat{\theta}_k^{ML} + \eta \alpha'_k \end{aligned} \tag{15}$$

with $\eta \stackrel{\text{def}}{=} \sum_k \alpha_k / (N + \sum_k \alpha_k)$ and $\alpha'_k \stackrel{\text{def}}{=} \alpha_k / N'$ with its range of $0 \le \alpha'_k \le 1$ using the prior strength $N' \stackrel{\text{def}}{=} \sum_k \alpha_k$.

For comparison, the MAP estimate of $\theta_k$ can be given by

$$\begin{aligned} \hat{\theta}_k^{MAP} &= \arg\max_{\theta_k} p(\theta|\Xi) = \frac{N_k + \alpha_k - 1}{N + \sum_{k'=1}^{K} (\alpha_{k'} - 1)} \\ &= (1 - \zeta) \hat{\theta}_k^{ML} + \frac{\zeta (\alpha'_k - \frac{1}{N'})}{\sum_{k'=1}^{K} (\alpha'_{k'} - \frac{1}{N'})} \end{aligned} \tag{16}$$

with $\zeta \stackrel{\text{def}}{=} \sum_k (\alpha_k - 1) / (N + \sum_k (\alpha_k - 1))$.

In (16), it is noted that the MAP estimate can still suffer from the sparse data problem in case of $N_k + \alpha_k = 1$. Therefore, the PM-based CDF estimation can be more straightforward in resolving the sparse data problem and approaches the ML estimation as $N \to \infty$.

Test CDF with the PM-based cumulative histogram estimation can be obtained by summing the PM estimates in (15) cumulatively as

$$\hat{C}_Y^{PM}(y) = (1 - \eta) \hat{C}_Y^{ML}(y) + \eta \sum_{k'=1}^{k} \alpha'_{k'} \tag{17}$$

A critical issue in (17) is to find the actual information for this prior. In HEQ, one strong candidate for the prior is training CDF which is approximated by a cumulative histogram or a Gaussian pdf. With cumulative histogram-based training CDF, the PM estimate of test CDF is then given by

$$\hat{C}_Y^{PM}(y) = (1 - \eta) \hat{C}_Y^{ML}(y) + \eta \hat{C}_X^{ML}(y) \tag{18}$$

where $\hat{C}_X^{ML}(\cdot)$ is the ML estimate version of training CDF.

To take advantage of the order statistic-based CDF estimation in (11) in the case of short test data, the PM-based test CDF estimate can be incorporated with the order statistic-based CDF, which is given by

$$\hat{C}_Y^{PM-OS}(y_n) = (1 - \eta) \frac{r(y_n) - 0.5}{N} + \eta \hat{C}_X^{ML}(y_n) \tag{19}$$

Finally, CHEQ with the PM-based test CDF estimation incorporating the order statistic-based sample CDF is given by replacing test CDF in (2) with the PM-OS estimate as

$$\hat{x}_n = \sum_{j=1}^{J} P(j|y_n) C_{X(j)}^{-1} \left[ (1 - \eta) \frac{R_j(y_n)}{\mathcal{N}(j)} + \eta \hat{C}_{X(j)}^{ML}(y_n) \right] \tag{20}$$

with the soft counting statistics of the $j$th class defined by

$$\mathcal{N}(j) = \sum_{n=1}^{N} P(j|y_n) \tag{21}$$

$$R_j(y_n) = \sum_{l=1}^{r_j(y_n)} P(j|y_{T_j(l)}) \tag{22}$$

where $r_j(y_n)$ is the rank of $y_n$ in the data set of the $j$th class.

## IV. Experimental Results

Experiments to evaluate the effectiveness of the proposed CHEQ technique with the PM-based test CDF estimation defined in (20) are carried out in the Aurora-4 framework [15]. The Aurora-4 database has two training sets, clean and multi-condition sets. The clean training set consists of 7138 utterances of the WSJ0 SI84 corpus. The multi-condition training set contains 7138 utterances consisting of one clean and six noisy subsets for testing microphone and noise conditions. Noisy subsets were built by artificially adding six types of noise at the randomly selected SNR range between 10 dB and 20 dB with two microphone conditions. The test set consists of 14 subsets. Two of them consist of clean speech dataset of the Nov'92 5000 words evaluation set, each of which was collected from different microphones. The remaining 12 test subsets were built by adding six types of noise at the randomly selected SNRs between 5 dB and 15 dB with two microphones. Our experiments were focused on the 16 kHz sampling rate.

The 39-dimensional mel-frequency cepstral coefficient (MFCC)-based feature vectors, each of which consists of log energy and 12 static MFCCs and their first and second derivatives, are extracted with a frame length of 25 ms and an interval of 10 ms in the experiments. Hidden Markov model (HMM)-based speech recognition systems were used where each triphone-based HMM consists of 3 states and each state has 16 mixture components. Diagonal covariance matrices are used in the HMM. The state-tying technique was employed which yields about 2000 tied states through the experiments. The bigram language model was used. The number of histogram bins in the training CDFs was empirically set to 64. Histogram equalization was applied on all 39-dimensional MFCCs for both training and test feature data in the segment level to deal with temporal noise variability and real-time requirement. The parameter $\eta$ was chosen empirically from the experiments, which mostly ranges between 0.4 and 0.8.

Fig. 1 and 2 represent recognition results by CHEQ with ML and PM based test CDF estimation along the number of classes in the Aurora-4 clean and multi-condition training tasks, respectively. The conditions of matched and mismatched

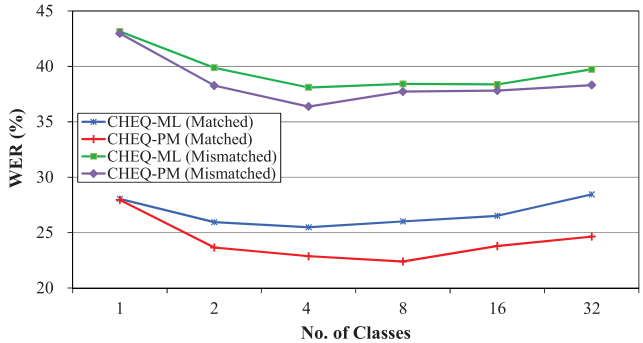


Fig. 1. Recognition results of CHEQ using ML and PM based test CDF estimation in the Aurora-4 clean-condition training task (averaged over one clean and six noisy evaluation subsets, segmental CHEQ with segment size of 2 sec).

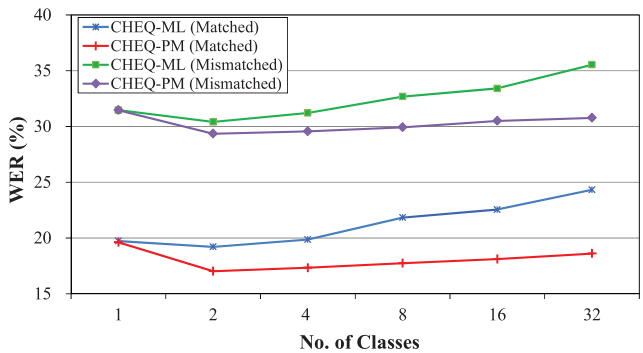


Fig. 2. Recognition results of CHEQ using ML and PM based test CDF estimation in the Aurora-4 multi-condition training task (averaged over one clean and six noisy evaluation subsets, segmental CHEQ with segment size of 2 sec).

TABLE I
RECOGNITION RESULTS OF VARIOUS FEATURE NORMALIZATION AND COMPENSATION TECHNIQUES IN THE AURORA-4 TASK (AVERAGED WER (%) OVER ONE CLEAN AND SIX NOISY EVALUATION SUBSETS, SEGMENT SIZE OF 2 SEC)

| Feature normalization techniques | Clean-condition | | Multi-condition | |
|---|---|---|---|---|
| | Matched | Mismatched | Matched | Mismatched |
| MFCC | 43.19 | 61.55 | 19.21 | 38.09 |
| CMN | 36.88 | 53.83 | 18.84 | 31.32 |
| CMVN | 34.72 | 51.65 | 19.97 | 34.05 |
| SCMN | 31.89 | 47.56 | 18.97 | 31.77 |
| SCMVN | 33.75 | 50.04 | 21.55 | 33.94 |
| ETSI-AFE | 29.38 | 53.04 | 18.14 | 33.34 |
| HEQ-ML | 28.04 | 43.15 | 19.73 | 31.47 |
| HEQ-PM | 27.95 | 42.97 | 19.63 | 31.40 |
| CHEQ-ML | 25.48 | 38.10 | 19.21 | 30.41 |
| CHEQ-PM | 22.40 | 37.73 | 17.02 | 29.35 |

refer to test sets of 1-7 and 8-14 of the Aurora-4 database, respectively [15]. The segment size is set to 2 sec with the same interval of 10 ms [5]. In the figures, it is observed that CHEQ is clearly superior to the original HEQ. For both matched and mismatched conditions, the PM-based method mostly provides significant performance improvements over the ML-based method. Its performance gain is generally more prominent in the multi-condition training task, which may be resulted from the fact that the prior contains multi-condition information. Table I shows performance comparison with other well-known feature normalization and compensation techniques including the utterance versions of CMN and CMVN, segmental versions of CMN (SCMN) and CMVN (SCMVN), and the ETSI advanced front-end (ETSI-AFE) feature extraction algorithm [16]. In the table, we see that CHEQ-PM produces consistent performance gains over the other feature normalization and compensation techniques by providing relative error reductions of 48%, 12% and 11%, 11% over MFCC and CHEQ-ML in the matched condition of both clean and multi-condition training tasks, respectively. The performance gains over MFCC are also significant in the mismatched conditions of both clean and multi-condition tasks with relative error reductions of 38% and 22%, respectively. However, performance improvements over CHEQ-ML seem marginal in the mismatched conditions of both training tasks. These results may be largely due to the mismatch between training and test data sets, which results in discrepancy between prior and sample distributions.

## V. CONCLUSION

We proposed a CHEQ approach employing the PM method, a kind of the Bayesian estimator, in estimating test CDF to improve the performance of HEQ by reducing the overfitting of HEQ when test data are sparse. The proposed method provides substantial performance gain over other feature normalization techniques including the ML-based CHEQ method.

## REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. H.-U. , "An overview of noise-robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, 2014.
[2] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 64, pp. 1304–1312, 1974.
[4] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, no. 1–3, pp. 133–147, 1998.
[5] J. C. Segura, C. Benítez, Á. de la Torre, A. J. Rubio, and J. Ramírez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 11, no. 5, pp. 517–520, 2004.
[6] Á. de la Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 3, pp. 355–366, 2005.
[7] Y. Suh, M. Ji, and H. Kim, "Probabilistic class histogram equalization for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 4, pp. 287–290, 2007.
[8] L. García, C. B. Ortúzar, A. De la Torre, and J. C. Segura, "Class-based parametric approximation to histogram equalization for ASR," *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 415–418, 2012.
[9] H.-J. Hsieh, B. Chen, and J.-W. Hung, "Histogram equalization of real and imaginary modulation spectra for noise-robust speech recognition," in *Proc. Interspeech*, 2013.
[10] X. Xiao, E. Chng, and H. Li, "Attribute-based histogram equalization (HEQ) and its adaptation for robust speech recognition," in *Proc. Interspeech*, 2013.
[11] V. Joshi, R. Bilgi, S. Umesh, L. Garcia, and C. Benítez, "Sub-band based histogram equalization in cepstral domain for speech recognition," *Speech Commun.*, vol. 69, pp. 46–65, 2015.
[12] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 845–854, 2006.
[13] Y. Suh and H. Kim, "Environment model adaptation based on histogram equalization," *IEEE Signal Process. Lett.*, vol. 16, no. 4, pp. 264–267, 2009.
[14] K. P. Murphy, *Binomial and multinomial distributions*, 2006 [Online]. Available: http://www.cs.ubc.ac./~murphyk/Teaching/CS340-Fall06/reading/bernoulli.pdf
[15] N. Parihar and J. Picone, "DSR front end LVCSR evaluation," Dec. 2002, AU/384/02, Aurora Working Group.
[16] ETSI standard doc, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI ES 202 050 v1.1.3, 2003.