Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition

José C. Segura, *Senior Member, IEEE*, Carmen Benítez, *Member, IEEE*, Ángel de la Torre, Antonio J. Rubio, *Senior Member, IEEE*, and Javier Ramírez, *Student Member, IEEE*

***Abstract*—This letter presents a new segmental nonlinear feature normalization algorithm to improve the robustness of speech recognition systems against variations of the acoustic environment. An experimental study of the best delay–performance tradeoff is conducted within the AURORA-2 framework, and a comparison with two commonly used normalization algorithms is presented. Computationally efficient algorithms based on order statistics are also presented. One of them is based on linear interpolation between sampling quantiles, and the other one is based on a point estimation of the probability distribution. The reduction in the computational cost does not degrade the performance significantly.**

***Index Terms*—Histogram equalization, order statistics, robustness, speech recognition.**

I. Introduction

THE ACOUSTIC mismatch between the training and test data [1] degrades the performance of automatic speech recognition (ASR) systems. In this letter, we focus on the so-called robust feature extraction approach, i.e., the extraction of speech features that are minimally affected by the environment.

Most speech recognition systems use parameterizations based on Mel frequency cepstral coefficients (MFCCs). Even for simple models of the acoustic environment (i.e., additive noise and linear channel distortion), the feature space is nonlinearly distorted [2]. As a result, the probability distribution of the features is different for different acoustic environments. This undesired variability is the principal cause of the performance degradation of ASR systems based on probabilistic models (i.e., Gaussian mixture-based recognizers).

Linear methods like cepstral mean subtraction (CMS) [3] or cepstral mean and variance normalization (CMVN) [4] yield significant improvements under noisy conditions. Nevertheless, these methods present important limitations, as they only provide compensation for the first two moments of the probability distributions of speech features [5]. Several histogram equalization (HEQ)-based approaches have been proposed [6]–[9]. The main specificity of our approach [5] is that, instead of trying to invert the nonlinear effects of the acoustic environment, HEQ is used to transform the features into a reference domain less affected by changes in the acoustic environment. This is essentially the same approach as that proposed in [10] and [11] for robust speaker verification.

Cepstral domain HEQ was shown to provide substantial improvements in speech recognition under noisy conditions, either as a standalone technique [5] or in combination with others [12], [13]. However, the original algorithm has been designed to perform the equalization on a sentence-by-sentence basis, and therefore this approach is not suitable for online applications, where a long variable delay is not acceptable. Furthermore, environment variations within a sentence cannot be properly handled with this algorithm. In this letter, we present a segmental implementation of HEQ, where a temporal window around the frame to be equalized is used instead of the whole sentence. We also present an experimental study of the delay–performance tradeoff for the segmental algorithm.

Two computationally efficient algorithms are also proposed. The first one, named quantile-based equalization (QBEQ), uses sampling quantiles to build a piecewise-linear approximation of the nonlinear transformation [8], [9]; the second one, named order statistic equalization (OSEQ), uses order statistics to build a point estimation of the cumulative distribution function (CDF) [14]. These two algorithms are compared in terms of its computational efficiency and performance. Experimental results have been obtained within the AURORA-2 framework [15].

II. HEQ-Based Segmental Feature Normalization

The goal of HEQ is to transform the speech features in such a way that the acoustic environment does not affect its probability distribution. This can be achieved by transforming the distribution of each feature into a fixed reference one. When the target distribution is selected as a Gaussian with zero mean and unity variance, this approach can be seen as an extension of CMVN. HEQ outperforms CMVN because it provides compensation for not only the first two moments affecting the location (mean) and scale (variance) of the distributions, but also for higher order moments affecting the shape of the distributions [5].

For a given random variable y with probability density function $p_y(y)$, a function $x = F(y)$ mapping $p_y(y)$ into a reference distribution $p_x(x)$ can be obtained by equating the CDF of x and y

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

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

Manuscript received April 29, 2003; revised September 17, 2003. This work was supported in part by the Spanish Government under the CICYT Project TIC2001-3323. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

The authors are with the Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada, Granada, 18071, Spain (e-mail: segura@ugr.es; carmen@ugr.es; atv@ugr.es; rubio@ugr.es; javierrp@ugr.es).

Digital Object Identifier 10.1109/LSP.2004.826648

where C_x^{-1} denotes the inverse of the reference CDF. The function $F(y)$ is monotonic nondecreasing and nonlinear in the general case.

Under the assumption of statistical independence, HEQ is applied to each cepstral coefficient independently. For each input sentence, the CDF of each coefficient $C_y(y)$ is approximated by its cumulative histogram. Next, the bin centers of this histogram are transformed according to (2) and finally, the transformed features are obtained by linear interpolation between these values.

For stationary noise processes, as more observations are considered, a better estimation of the cumulative histograms is obtained, and therefore, more accurate environment compensation is achieved. However, in the case of nonstationary noises, results can be improved by an adaptive estimation procedure. In the segmental version of HEQ, a temporal window around the frame to be normalized is considered for the estimation of the CDF of the features.

A straightforward extension of HEQ can be considered for a segmental implementation of the nonlinear transformation. At a given time t, a buffer containing $2T+1$ values of a particular cepstral coefficient is considered

$$\mathbf{Y}_t = \{y_{t-T}, \ldots, y_t, \ldots, y_{t+T}\}. \tag{3}$$

The cumulative histogram of these values is used as an estimation of the CDF. Then, a piecewise linear approximation of the transformation function is built, and the transformed value of y_t is obtained from it.

At the beginning of each utterance, once $T+1$ frames have been shifted into the buffer, the upper half of the buffer is replicated into the lower half, and the central frame (i.e., the first frame of the utterance) is equalized. The process then continues by shifting new frames into the buffer and equalizing the central one. When all frames of the utterance have been consumed the buffer remains fixed, and the last T frames of the utterance are equalized using this fixed buffer. Sentences with less than $T+1$ frames are normalized using all their frames.

In this letter, the selected reference distribution C_x was a Gaussian with zero mean and unity variance. The number of bins used in the estimation of the cumulative histograms must be selected taking into account the tradeoff between smoothness and resolution of the cumulative histograms. Several previous experiments have shown that the best performance is obtained with high-resolution cumulative histograms, and therefore, a relative high number of bins are used. In this letter, 100 regularly spaced bins are considered in the interval $[-4\sigma, +4\sigma]$, where σ is the estimated standard deviation of the samples in (3).

The smoothed cumulative histogram $\hat{C}_y$ is obtained by linear interpolation between the raw one $\bar{C}_y$ and that corresponding to a uniform distribution U with an equal number of bins

$$\hat{C}_y = \lambda \bar{C}_y + (1-\lambda)U. \tag{4}$$

The interpolation factor $\lambda = N/(N+10)$ is selected as a function of the number of observations $N = 2T+1$, and therefore, less smoothing is applied when more observations are available.

III. Order-Statistic-Based Transformations

This direct implementation of a segmental version of HEQ is not computationally efficient. For the estimation of the CDF, a whole cumulative histogram is computed every new frame; but according to (2), we only need an estimation of $C_y(y_t)$ to perform the equalization. More efficient algorithms can be formulated by exploiting the relation between order statistics and values of the CDF.

Let us denote the order statistics of (3) by

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(r)} \leq \cdots \leq y_{(2T+1)}. \tag{5}$$

These are the same values in (3) but sorted in ascending order.

A. Quantile-Based Transformation

As a first approach, a reduced number of sampling quantiles can be used to build an interpolated approximation of the nonlinear transformation. With this approach, a more efficient solution is obtained at the cost of reducing the resolution of the estimated transformation.

The algorithm used in this letter is similar to the one used in [8] and [9]. From a Gaussian reference with zero mean and unity variance, N_Q quantiles $Q_x(p_r) = C_x^{-1}(p_r)$ are computed for probability values

$$p_r = \left(\frac{r-0.5}{N_Q}\right), \qquad \forall r = 1, \ldots, N_Q. \tag{6}$$

The corresponding sampling quantiles $Q_y(p_r)$ are estimated from the order statistics (5) as

$$Q_y(p_r) = \begin{cases} (1-f)y_{(k)} + f y_{(k+1)}, & 1 \leq k < 2T \\ y_{(2T+1)}, & k = 2T+1 \end{cases} \tag{7}$$

where k and f are the integer and fractional parts of $1+2Tp_r$, respectively.

As each pair of quantiles $(Q_y(p_r), Q_x(p_r))$ represents a point of the nonlinear transformation, the transformed value of the central frame y_t is obtained by linear interpolation between the tabulated points. Linear extrapolation is used whenever y_t is less than the first sampling quantile or greater than the last one. This way, the nonlinear transformation is approximated with $N_Q - 1$ linear segments. In the following, we will refer to this algorithm as QBEQ (quantile-based equalization).

Obtaining the sorted dataset (5) requires $(2T+1)\log_2(2T+1)$ comparisons on average. The computation of N_Q quantiles requires $2N_Q$ products and N_Q additions (unless probability values are selected to match the corresponding quantiles to order statistics), and the interpolation process requires two products and two additions. For a reduced number of quantiles, this computational cost is lower than the corresponding segmental version of HEQ.

B. Direct Estimation

An even more efficient algorithm is formulated from a direct estimation of $C_y(y_t)$. An asymptotically unbiased point estimation of the CDF can be defined [14] as

$$\hat{C}_y(y_{(r)}) = \frac{r-0.5}{2T+1} \quad \forall r = 1, \ldots, 2T+1. \tag{8}$$

Using (8) and (2), an estimation of the transformed value of y_t can be obtained as

$$\hat{x}_t = C_x^{-1}(\hat{C}_y(y_t)) = C_x^{-1}\left(\frac{r(y_t)-0.5}{2T+1}\right) \tag{9}$$

where $r(y_t)$ denotes the rank of y_t (i.e., the index r of the order statistics that corresponds to the value y_t) that is obtained by counting the number of values less or equal than y_t in the temporal buffer Y_t. Note that as C_x and T are fixed, if the values

$$G[r] = C_x^{-1}\left(\frac{r-0.5}{2T+1}\right) \quad \forall r = 1,\ldots,2T+1 \tag{10}$$

are tabulated in advance, the transformed value (9) can be obtained by simply indexing the table G. As for HEQ and QBEQ, the selected reference distribution is a Gaussian with zero mean and unity variance.

The computational cost of this algorithm is much less than the corresponding one for QBEQ as only $2T$ comparisons are needed to obtain the transformed value of a given feature. In the following, we will refer to this algorithm as OSEQ (order statistics-based equalization).

IV. Experimental Results

The segmental version of HEQ has been evaluated within the AURORA-2 experimental framework [15]. A re-endpointed[1] version of the database is used as suggested for the last AURORA special session at ICSLP 2002. The working database is a subset of TI-DIGITS, and contains connected digits recorded in a clean environment. Utterances have been contaminated by the addition of several noise types at different SNR levels. Three test sets are defined. Two of them contain only additive noise, and the last one includes also a simulated channel mismatch. The task consists of two kinds of recognition experiments: one using a recognizer trained with clean speech [clean condition (CC)] and the other one using a recognizer trained with sentences contaminated by different kinds and levels of noise [multicondition (MC)].

Continuous density left-to-right HMMs are used for the acoustic models. Digits are modeled with 16 emitting states and a three Gaussian mixture per state. Additionally, two pause models are defined. The first one consists of three states with a six Gaussian mixture per state, and models beginning and end pauses. The second one models interdigit pauses and has only one state tied with the central one of the previous model. The recognizer is based on HTK and uses a 39-component feature vector: 12 MFCC plus the logarithmic energy and the corresponding delta and acceleration coefficients (see [15] for details). Features are extracted at a frame-rate of 100 Hz.

For comparison purposes, segmental versions of CMVN and CMS have also been evaluated within the same framework. A common buffer of $2T+1$ frames is used for CDF, mean, and variance estimations. In the CMS experiments, the mean is subtracted from the static features before regression coefficients (delta and acceleration coefficients) are computed. In HEQ and CMVN experiments, the regression coefficients are computed first, and then all the 39 components of the feature vector are normalized independently.

[1]The database has been accurately endpointed leaving a 200-ms silence period at the beginning and at the end of each utterance.

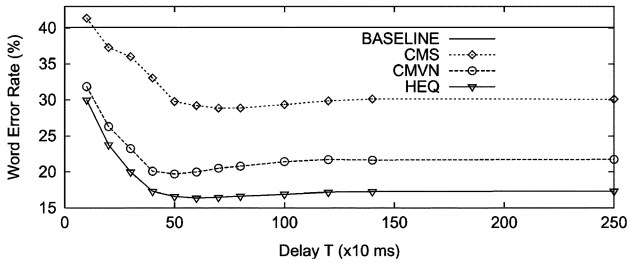

Fig. 1. CC results (averaged for SNR levels between 0–20 dB) as a function of the delay for the segmental versions of HEQ, CMS, and CMVN. AURORA-2 baseline results are also shown for reference.

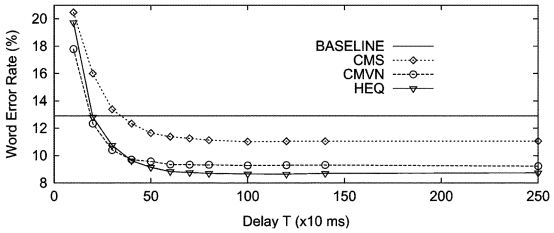

Fig. 2. MC results (averaged for SNR levels between 0–20 dB) as a function of the delay for the segmental versions of HEQ, CMS, and CMVN. AURORA-2 baseline results are also shown for reference.

A first set of experiments has been conducted using the CC recognizer. To evaluate the performance of the algorithms as a function of the delay, the front-end of the system has been modified to perform feature normalization based on a temporal buffer of $2T+1$ frames. Features have been normalized for both training and test data; and for each normalization algorithm (CMS, CMVN, and HEQ), the recognizer has been trained and evaluated for delay values from 100–1400 ms. A delay greater than half the maximum duration of the sentences (2500 ms) has been used to obtain the asymptotic performance values corresponding to the nonsegmental versions. Fig. 1 shows the word error rates obtained for the segmental versions of HEQ, CMVN, and CMS as a function of the delay. These results are averaged values for all the noise types and for SNR levels between 0–20 dB.

First of all, the asymptotic values of the word error rate show how the progressive compensation of higher order moments of the feature distributions results in better recognition performance; CMVN (21.74%) performs better than CMS (30.11%), and HEQ (17.23%) has the best performance.

Second, the plots show how HEQ performance is improved as the delay is increased, obtaining the best result (16.35%) for a delay value of 600 ms. From this point, no further improvement is obtained by increasing the delay. This behavior shows the successfulness of the segmental version of HEQ. A similar behavior (consistent with Viikki results [4]) is observed for CMVN, with the lower error rate (19.70%) obtained for a delay 500 ms; and for CMS (28.87%) for a delay of 700 ms.

The previously described set of experiments has been carried out using the MC recognizer. In Fig. 2, it can be observed that the behavior of the segmental algorithms is now different. The word error rate decreases almost monotonically with the delay, although small reduction is obtained for delays greater than 600 ms.

TABLE I
AVERAGED WORD ERROR RATES AND RELATIVE IMPROVEMENTS (AS DEFINED FOR THE AURORA SPECIAL SESSIONS) FOR A DELAY OF 600 ms

	Word error rate			Relative
	MC	CC	Average	improvement
BASELINE	12.97%	41.94%	27.46%	—
HEQ	8.83%	16.35%	12.59%	44.51%
OSEQ	8.92%	16.12%	12.52%	44.59%
QBEQ N_Q=30	8.89%	16.71%	12.80%	44.11%
QBEQ N_Q=15	8.97%	16.93%	12.95%	43.21%
QBEQ N_Q=5	8.99%	16.92%	12.96%	40.56%
QBEQ N_Q=3	9.23%	17.39%	13.31%	36.94%
QBEQ N_Q=2	9.79%	21.86%	15.82%	24.87%

Asymptotic word error rates are consistent with those obtained for the CC recognizer: HEQ (8.74%) has the lowest error rate, and CMVN (9.23%) outperforms CMS (11.06%). However, the differences between algorithms are now smaller because the mismatch between training and test data is greatly reduced by the multistyle training used for this recognizer. Notice that the word error rate for HEQ with a delay of 600 ms (8.83%) is not significantly higher than the asymptotic value (8.74%).

Both QBEQ and OSEQ algorithms have been evaluated using the same experimental setup described above. The resulting delay behavior of OSEQ was essentially the same observed for HEQ, with the maximum performance obtained for a delay of 600 ms. QBEQ has been evaluated for this same delay with different number of quantiles N_Q. Table I shows the averaged word error rates and relative improvements for the baseline and the different algorithms.

From these results, it can be concluded that the performance of OSEQ is almost the same as that obtained for HEQ. Considering the averaged relative improvements, the performances of OSEQ and HEQ are virtually equal. Although HEQ performs slightly better in MC and OSEQ performs slightly better in CC, the difference between relative improvements is less than 0.75%.

For QBEQ, a consistent improvement is obtained by increasing the number of quantiles. Results with two and three quantiles are of special interest. The first situation is similar to CMVN, and the second one is similar to the modification of CMVN proposed in [16]. Although the results for 30 quantiles are close to those for OSEQ, the lower complexity and computational cost of this last method makes it the best selection. Note that the quantiles have been uniformly selected, and an open question is if an optimized selection can result in a better performance. This subject is under investigation.

V. CONCLUSION

In this letter, we have studied several feature normalization algorithms working in the cepstral domain. We have found that a temporal context of about 1.2 s is enough for the proper estimation of the nonlinear transformation for a connected digit recognition task. This result is consistent with those previously reported for CMVN.

For the clean-condition recognizer, the segmental version of HEQ can perform better than the nonsegmental one. This can be explained by the ability of the segmental algorithm to adapt the normalizing transformation to changes in the acoustic environment within a sentence. This result is not obtained in the case of the multicondition recognizer because of the multistyle training used in this case.

The segmental version of HEQ has been compared with segmental implementations of two other feature normalization algorithms: CMS and CMVN. Experimental results have shown that the compensation of higher order moments provided by HEQ gives the best recognition performance.

We have also presented a computationally efficient implementation of the HEQ technique based on order statistics. Two alternative algorithms have been considered; one of them based on the estimation of a reduced number of quantiles and the other based on a direct estimation of the CDF. Experimental results have shown that OSEQ performance is comparable to that achieved with the segmental version of HEQ. Although QBEQ can reach almost the same performance, OSEQ is simpler and its computational cost is lower.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, 1995.
[2] R. Stern, B. Raj, and P. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. ESCA-NATO Tutorial Research Workshop Robust Speech Recognition for Unknown Communication Channels*, 1997, pp. 33–42.
[3] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proc. ARPA Speech Natural Language Workshop*, Princenton, NJ, Mar. 1993, pp. 69–74.
[4] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, pp. 133–147, 1998.
[5] A. de la Torre, J. Segura, C. Benitez, A. Peinado, and A. Rubio, "Nonlinear transformations of the feature space for robust speech recognition," in *Proc. ICASSP*, Orlando, FL, May 2002, pp. 401–404.
[6] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 556–559.
[7] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *Proc. ASRU*, Trento, Italy, Dec. 2001, pp. 21–24.
[8] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sept. 2001, pp. 1135–1138.
[9] F. Hilger, S. Molau, and H. Ney, "Quantile based histogram equalization for online applications," in *Proc. ICSLP*, Denver, CO, Sept. 2002, pp. 1135–1138.
[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Oddyssey Conf.*, June 2001.
[11] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time gausssianization for robust speaker verification," in *Proc. ICASSP*, Orlando, FL, May 2002, pp. 681–684.
[12] J. Segura, C. Benitez, A. de la Torre, S. Dupont, and A. Rubio, "VTS residual noise compensation," in *Proc. ICASSP*, Orlando, FL, May 2002, pp. 409–412.
[13] J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR," in *Proc. ICSLP*, Denver, CO, Sept. 2002, pp. 225–228.
[14] R. Suoranta, K.-P. Estola, S. Rantala, and H. Vaataja, "PDF estimation using order statistic filter bank," in *Proc. ICASSP*, vol. 3, Apr. 1994, pp. 625–628.
[15] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Proc. ISCA ITRW ASR2000 Conf. "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, Sept. 2000.
[16] X. Menéndez-Pidal, R. Chan, D. Wu, and M. Tanaka, "Compensation of channel and noise distortions combining normalization and speech enhacement techniques," *Speech Commun.*, vol. 34, pp. 115–126, 2001.