PARAMETRIC NONLINEAR FEATURE EQUALIZATION FOR ROBUST SPEECH RECOGNITION

Luz García, José C. Segura, Javier Ramírez, Angel de la Torre, Carmen Benítez

Dpto. Teoría de la Señal, Telemática y Comunicaciones (TSTC)
Universidad de Granada

ABSTRACT

A new front-end normalization algorithm that uses a parametric nonlinear transformation is proposed in this paper. The method improves histogram equalization based nonlinear transformations by finding a simple and computationally inexpensive parametric expression of the nonlinear transformation. The new parametric approach relies on a two Gaussian model for the probability distribution of the features, and on a simple Gaussian classifier to label the input frames as belonging to the speech or non-speech classes. The result is a more robust equalization, less dependent on the percentage of speech and non-speech frames. Recognition experiments on the AURORA 4 database have been performed and the effectiveness of the algorithm is analyzed in comparison with other linear and nonlinear feature equalization techniques.

1. INTRODUCTION

On the framework of Robust Speech Recognition, Histogram Equalization (HEQ) based feature transformations have been successfully applied [1, 2, 3, 4] to deal with the nonlinear effect of the acoustic environment in the feature domain. These techniques differ in the domain where the normalization is performed [3, 5] and in the way the probability distributions are modeled [6, 7].

The main goal of this approach is to normalize the probability distributions of the features in such a way that the acoustic environment effects are (partially) removed. HEQ based transformations can be seen as an extension of linear transformation techniques like cepstral mean subtraction (CMS) and cepstral mean and variance normalization [8] that only deal with the normalization of the two first moments of the probability distributions of the features.

Although HEQ based techniques have been proved to significantly improve the robustness of speech recognition systems against acoustic environment degradations [9, 10, 11], they suffer from several limitations. In most cases, HEQ based techniques rely on a local estimation of the probability distributions of the features based on a reduced number of observations belonging to a single utterance to be equalized. Using histograms leads to rather noisy estimations of the cumulative distribution functions (CDF). To overcome this drawback, sampling quantiles [6, 7] have been proposed as an alternative to characterize the CDF of the features. In this approach, the transformation function is defined as a piecewise linear mapping between estimated quantiles and a set of quantiles describing the reference CDF.

A second drawback of HEQ based techniques is that the nonlinear transformation is based on mapping the global CDF of each feature into a reference one. When the estimation of the CDF is built using a reduced number of observations from a single utterance, variations in the amount of non-speech frames in the utterance introduce unwanted variability in the estimated CDF [12] and therefore in the corresponding transformation function that may degrade the performance of the technique. This problem can be alleviated using a voice activity detector (VAD) to drop non-speech frames prior to the equalization [10]. Finally, HEQ transformations are usually based on a component-by-component equalization of the feature vector, thus discarding any cross-information between features in the equalization process.

To overcome these limitations, we propose here a parametric nonlinear equalization technique that is based on a simple two-class Gaussian mixture model of the feature probability distributions. The main advantage of the proposed approach is that using parametric models with little free parameters allows smoother estimations of the probability distributions, and can also take into account both the cross-feature correlations and the variable amount of non-speech frames in each utterance. Experiments on AURORA4 database have shown the benefits of this new approach in comparison with the usual non-parametric component-by-component approaches to HEQ.

The rest of the paper is organized as follows. In section 2, the traditional HEQ approach is briefly revised and the new approach is presented. Section 3 shows comparative results on the AURORA4 task and finally, the conclusions of this work are presented in section 4 along with future work.

2. PARAMETRIC EQUALIZATION

2.1. Histogram-based and quantile-based equalization

HEQ techniques use the following property of the random variables: 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:

$$C_y(y) = C_x(x) = C_x(F(y)) \quad (1)$$

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

where C_x^{-1} denotes the inverse of the reference CDF. The function $F(y)$ is monotonic non-decreasing 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.

Instead of using histograms, the transformation can also be defined as a piecewise mapping between a predefined set of quantiles of the reference probability distribution and those estimated from observations of a given utterance to be equalized [6, 9, 7]. From a

reference distribution, N_Q quantiles $Q_x(p_r) = C_x^{-1}(p_r)$ are computed for probability values:

$$p_r = \left(\frac{r - 0.5}{N_Q}\right) \quad \forall r = 1, \ldots, N_Q \tag{3}$$

The corresponding sampling quantiles $Q_y(p_r)$ are estimated from the order statistics [7]. 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.

2.2. Dependence on the silence content

As said before, the relative content of non-speech frames is a cause of variability in the HEQ transformation. This is because an estimation of the global probability distribution is used, that takes into account both speech and non-speech frames. As an example, figure 1 shows this situation. Figure 1 (a) shows the temporal values of the first cepstral coefficient of a typical utterance and figure 1 (b) shows the same coefficient after part of the initial silence has been removed. The estimated CDF's for both utterances are depicted in figure 1 (c). It is clearly shown that, although both utterances have the same values for the speech-frames, the different amount of non-speech frames alters the global CDF. This differences on the estimated CDF's induces an unwanted variability on the estimated transformation, as illustrated in figure 1 (d).

The reason of this variability is evident if we express the CDF as a mixture of two CDF's corresponding to the speech and non-speech frames:

$$C_x(x) = \alpha C_{nx}(x) + (1 - \alpha) C_{sx}(x) \tag{4}$$

where α is the fraction of non-speech frames and $C_{nx}(x)$ and $C_{sx}(x)$ are the CDF's of non-speech and speech frames respectively. Even if the probability distributions of the features remain unaltered, different values of α result in different $C_x(x)$ distributions.

2.3. Two-class parametric equalization

This unwanted variability of the transformation induced by the variable proportion of non-speech frames of each utterance can be reduced by removing non-speech frames before the estimation of the transformation. Another possibility is to use different transformations for speech and non-speech frames [12]. That is, instead of using a transformation to map the global CDF's of the features, we can build separate mappings for speech and non-speech frames. Although theoretically attractive, this approach can hardly be implemented in a sentence-by-sentence equalization strategy because of the lack of sufficient data to obtain an accurate enough estimation of the required CDF's.

As an alternative, we propose the use of a parametric form of the equalization transform based on a two Gaussian mixture model. The first Gaussian is used to represent non-speech frames, while the second one represents speech frames. For each class, a parametric linear transformation is defined to map the clean and noisy representation spaces:

$$\hat{x} = \mu_{n,x} + (y - \mu_{n,y}) \left(\frac{\Sigma_{n,x}}{\Sigma_{n,y}}\right)^{1/2} \quad \text{if } y \text{ is non-speech} \tag{5}$$

$$\hat{x} = \mu_{s,x} + (y - \mu_{s,y}) \left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}}\right)^{1/2} \quad \text{if } y \text{ is speech} \tag{6}$$

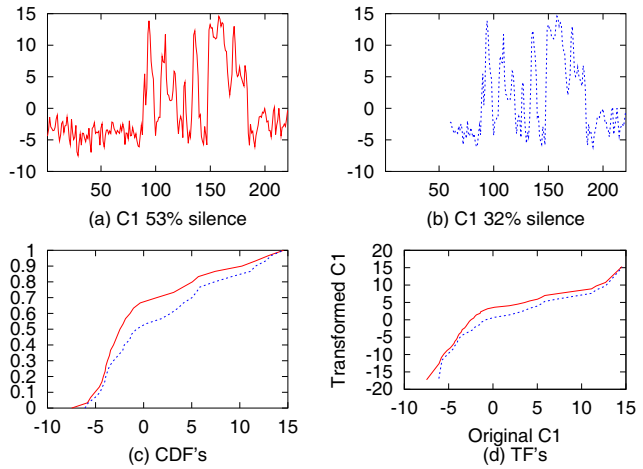

Fig. 1. CDF dependence on the non-speech content of an utterance. (a) First cepstral coefficient of a typical utterance. (b) The same coefficient after partially removal of the initial non-speech frames. (c) Estimated CDF's for the cepstral coefficient in (a) (solid line) and in (b) (dashed line). (d) Transformations obtained from the CDF's for the original coefficient (solid line) and after initial non-speech frames removal (dashed line).

where $\mu_{n,x}$, $\Sigma_{n,x}$, $\mu_{s,x}$ and $\Sigma_{s,x}$ correspond to the Gaussians modeling clean non-speech and speech frames, respectively, and $\mu_{n,y}$, $\Sigma_{n,y}$, $\mu_{s,y}$ and $\Sigma_{s,y}$ correspond to the Gaussians modeling noisy non-speech and speech frames. With these definitions of the linear transformations, the noisy means $\mu_{n,y}$ and $\mu_{s,y}$ are transformed into the clean means $\mu_{n,x}$ and $\mu_{s,x}$, and the noisy covariance matrices $\Sigma_{n,y}$ and $\Sigma_{s,y}$ are transformed into the clean covariance matrices $\Sigma_{n,x}$ and $\Sigma_{s,x}$ (for both, the non-speech and speech models). The clean Gaussians for speech and non-speech frames can be estimated from the training database, while the noisy Gaussians should be estimated from the utterance to be equalized.

In order to select whether the current frame y is speech or non-speech, a voice activity detector could be used. However, this implies a hard decision between both linear transformations that could create discontinuities in the limit of the non-speech/speech decision. Instead, a soft decision can be used:

$$\hat{x} = P(n|y) \left(\mu_{n,x} + (y - \mu_{n,y}) \left(\frac{\Sigma_{n,x}}{\Sigma_{n,y}}\right)^{1/2}\right) + P(s|y) \left(\mu_{s,x} + (y - \mu_{s,y}) \left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}}\right)^{1/2}\right) \tag{7}$$

by including the conditional probabilities of frame y being non-speech or speech. The posterior probabilities $P(n|y)$ and $P(s|y)$ are obtained using a simple two-class Gaussian classifier on the log-energy term (the C_0 cepstral coefficient). Initially, those frames with C_0 below the mean value are assigned to the non-speech class and those with C_0 above the mean are assigned to the speech class. The EM algorithm is then iterated until convergence (usually, 10 iterations are enough) to obtain the final classifier. This classifier is used to obtain the class probabilities $P(n|y)$ and $P(s|y)$ and also to obtain the mean and covariance matrices $\mu_{n,y}$, $\Sigma_{n,y}$, $\mu_{s,y}$ and $\Sigma_{s,y}$ for the non-speech and speech classes for the given noisy input utterance. Then, the input utterance can be equalized using equation

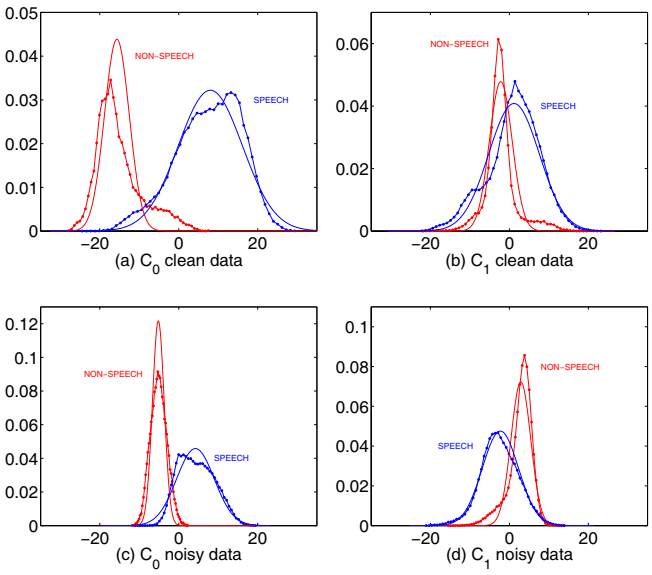

Fig. 2. Two Gaussian model and histograms of two first cepstral coefficients of clean (a) and (b) and noisy data (c) and (d).

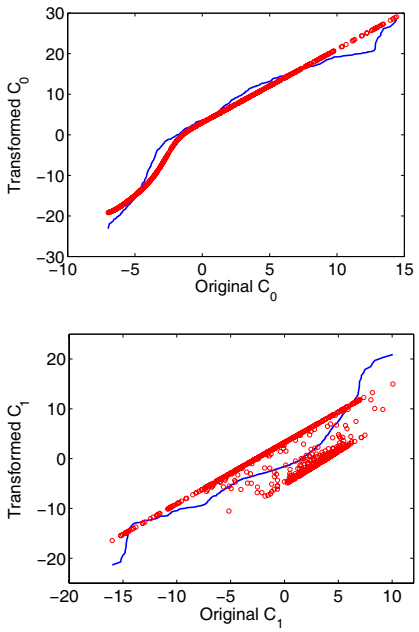

Fig. 3. Transformations provided by the proposed parametric equalization technique (circles) along with the transformation provided by histogram equalization (solid line) for the two first cepstral coefficients C_0 (top) and C_1 (bottom) of a typical noisy utterance.

(7). This equation leads to a non-linear interpolation of two class-dependent linear transformations.

Figures 2 (a) and 2 (b) show the two Gaussian model for the C_0 and C_1 cepstral coefficients (used as reference model) along with the histograms of the speech and non-speech frames for a set of clean utterances. Figures 2 (c) and 2 (d) show the same data for a set of noisy utterances.

Figure 3 represents the resulting transformation for a typical noisy utterance according to the proposed parametric procedure. The transformation provided by histogram equalization is also represented for comparison. The transformations are represented for the C_0 (top) and C_1 (bottom) cepstral coefficients. Since the parametric equalization relies on the class probabilities $P(n|y)$ and $P(s|y)$, and they depend on the level of the C_0 cepstral coefficient, equation (7) providing $\hat{x}$ as a function of y is a non-linear transformation which tends to the linear mapping given by equation (6) when $P(s|y) \gg P(n|y)$ and to the linear mapping of equation (5) when $P(n|y) \gg P(s|y)$. In the case of C_1 cepstral coefficient, since the probabilities $P(n|y)$ and $P(s|y)$ depend on C_0, the relationship between the noisy and equalized values is not a monotonic function, and a value of noisy C_1 could provide different values of equalized C_1, depending on the value of C_0 for this frame. This behavior is consistent with the probability distributions of C_1 cepstral coefficient for non-speech and speech frames in clean and noisy conditions, observed in figure 2 (b) and 2 (d). In this case, a non-linear trend can also be observed, and histogram equalization-based transformation tries to provide the best non-linear monotonic function matching the parametric equalization of the proposed method.

3. EXPERIMENTAL RESULTS

The proposed parametric equalization algorithm has been tested on the AURORA4 (WSJ0) database, following the standard clean training test. All the procedures for recognition and training are identical to the reference experiments with the exception of the front-end that includes the normalization procedure described in this paper. The recognition system used in all cases is based on continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state. Training and recognition are performed using the HMM Tool Kit (HTK) software. The language model is the standard bigram for the WSJ0 task. A feature vector of 13 cepstral coefficients is used as the basic parameterization of the speech signal using C_0 instead of the logarithmic energy. This basic feature vector is augmented with first and second order regressions yielding a final 39 components feature vector. The baseline reference system (BASE) uses sentence-by-sentence subtraction of the mean values of each cepstral coefficient (CMS).

For the proposed parametric equalization technique (PEQ), the parameters of the reference distribution have been obtained by averaging over the whole clean training set of utterances. Both training and test utterances have been then equalized to this reference distribution using equation (7). Cepstral coefficients are equalized before the computation of the regressions.

For comparison purposes, two additional experiments have been conduced. The first one (HEQ) is similar to PEQ with the difference that in this case, the cepstral coefficients are equalized using a quantile based approach as described in section 2.1. For each cepstral coefficient, 31 quantiles are estimated for each cepstral coefficient and each utterance. The reference CDF of each cepstral coefficient is obtained by averaging the quantiles of each of the clean training data utterances. Training and test data are then equalized to this reference distribution by estimating the quantiles of each input utterance and using a piecewise linear transformation between these values and the corresponding ones of the reference distribution. The last experiment (AFE) uses the ETSI standard advanced front-end parameterization algorithm [13].

Table 1 shows the obtained word error rates for the 14 test sets of AURORA4. First row (BASE) corresponds to the baseline system

Table 1. Word error rates for the 14 test of AURORA4 clean training experiment (8KHz / 166 small tests). Results for the baseline system (BASE), histogram based equalization (HEQ), the proposed parametric equalization technique (PEQ) and the ETSI advanced front-end (AFE).

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	Avg
BASE	13,2	24,7	46,0	47,6	52,7	44,8	54,7	22,6	36,2	55,4	58,3	65,3	54,1	62,3	45,6
HEQ	11,2	23,2	37,7	38,3	37,5	37,7	39,7	20,9	33,4	46,0	49,8	52,2	45,9	51,0	37,5
PEQ	13,4	17,6	32,5	35,4	35,1	30,0	37,5	16,8	22,6	37,0	41,2	44,4	36,0	41,9	31,5
AFE	12,7	17,8	30,4	34,8	30,6	34,9	31,7	18,8	25,1	38,0	44,9	40,4	39,3	38,4	31,3

which is based on a simple CMS linear normalization technique. The second row (HEQ) shows the word error rates when using a standard quantile-based implementation of HEQ. A relative word error rate reduction of 17.8% is obtained in this case. The performance of HEQ is clearly improved by PEQ as shown in the third row, with a relative word error reduction of 30.8%. This result is very close to the one obtained for the AFE, which yields a 31.4% reduction of the word error rate. Moreover, PEQ outperforms AFE in half of the tests (i.e. 02, 06, 08, 09, 10, 11 and 13).

4. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a simple parametric feature transformation technique for nonlinear feature equalization and its application for robust speech recognition.

The transformation is based on a nonlinear interpolation of two independent linear transformations. The linear transformations are obtained using a simple Gaussian model for the classes of speech and non-speech features.

The technique has been evaluated on a complex continuous speech recognition task showing its competitive performance against linear and nonlinear feature equalization techniques like CMS and HEQ. Moreover, the recognition accuracy is comparable to that obtained with standard front-end algorithms like the ETSI advanced front-end.

In the implementation presented in this paper, diagonal covariance matrices have been assumed in the multivariate Gaussian model for the features in each class. A study of influence of within class cross-correlations is currently under development.

5. ACKNOWLEDGEMENTS

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SR3-VoIP project (TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

6. REFERENCES

[1] R. Balchandran, *Non-parametric estimation and correction of non-linear distortion in speech systems*, Ph.D. thesis, Rutgers University, 1997.

[2] S. Dharanipragada and M. Padmanabhan, "A non-linear unsupervised adaptation technique for speech recognition," in *Proc. of ICSLP'00*, Peking, China, October 2000, pp. 556–559.

[3] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *Proc. of ASRU'01*, Trento, Italy, December 2001, pp. 21–24.

[4] A. de la Torre, J. C. Segura, M. C. Benítez, A. M. Peinado, and A. J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proc. of ICASSP'02*, Orlando, Florida, September 2002, pp. 401–404.

[5] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez, M. C. Benítez, and A. Rubio, "Histogram equalization of the speech representation for robust speech recognition," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[6] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition," in *Proc. of EUROSPEECH'01*, Aalborg, Denmark, September 2001, pp. 1135–1138.

[7] J. C. Segura, M. C. Benítez, A. de la Torre, and A. J. Rubio, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal PProcessing Letters*, vol. 11, no. 5, pp. 517–520, May 2004.

[8] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[9] F. Hilger, S. Molau, and H. Ney, "Quantile based histogram equalization for online applications," in *Proc. of ICSLP'02*, Denver, Colorado, September 2002, pp. 237–240.

[10] J. C. Segura, M. C. Benítez, A. de la Torre, and A. J. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR," in *Proc. of ICSLP'02*, Denver, Colorado, September 2002, pp. 225–228.

[11] J. C. Segura, M. C. Benítez, A. de la Torre, S. Dupont, and A. J. Rubio, "VTS residual noise compensation," in *Proc. of ICSLP'02*, Denver, Colorado, September 2002, pp. 409–412.

[12] S. Molau, F. Hilger, D. Kersey, and H. Ney, "Enhaced histogram normalization in the acoustic feature space," in *Proc. of ICSLP'02*, Denver, Colorado, September 2002, pp. 1421–1424.

[13] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 202 050 Recommendation*, 2002.