ELectronics EXpress

IEICE Electronics Express, Vol.6, No.12, 824–830

A new perceptually weighted distance measure for vector quantization of the STFT amplitudes in the speech application

Roghayeh Doost[1a)], **Abolghasem Sayadian**[1], **and Hossein Shamsi**[2]

[1] *Electrical Faculty, Amirkabir University of Technology, Tehran, Iran*

[2] *Electrical Faculty, K.N. Toosi University of Technology, Tehran, Iran*

a) *rdoost@aut.ac.ir*

Abstract: In this paper, a new perceptually weighted Euclidean distance function is proposed for the VQ of the speech signal. It employs a logarithmic preprocessing and utilizes the listening characteristics of the ear for the definition of the perceptually weighted Euclidean distance function. Simulation results show that the proposed VQ process has less spectral distortion (SD) than its conventional Euclidean counterpart does.

Keywords: speech processing, euclidean distance function, VQ, STFT, spectral distortion

Classification: Science and engineering for electronics

References

[1] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, 1980.

[2] Stephen So, "Efficient Block Quantization for Image and Speech Coding," Thesis for degree of Doctor of Philosophy, School of Microelectronic Engineering, Faculty of Engineering and Information Technology, Griffith University, Brisbane, Australia, BEng(Hons), 2005.

[3] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook Driven Short-Term Predictor Parameter Estimation for Speech Enhancement," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[4] Wai C. Chu, "Vector Quantization of Harmonic Magnitudes in Speech Coding Applications—A Survey and New Technique," *EURASIP J. Appl. Signal Process.*, pp. 2606–2613, 2004.

[5] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," *IEEE Trans. Speech Audio Process.*, vol. 1, pp. 3–14, Jan. 1993.

[6] Aki Harma, "Perceptual Aspects and Warped Techniques in Audio Coding," Thesis for degree of Master of Science, Helsinky University of Technology, Facaulty of Electrical Engineering, 1997.

[7] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 73–76, 2001.

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009

[8] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Sel. Areas Commun.*, vol. 10, pp. 819–829, June 1992.

1 Introduction

Vector quantization is a process whereby the elements of a vector are jointly quantized. The VQ derives a codebook of reference vectors from a training data. The code vectors are determined by minimizing a specific distance function, conventionally Euclidean distance function [1, 2].

In speech processing, the VQ is considered as a proper speech modeling that is applied on the parameters of the speech. In [3], making use of the VQ, the LPC parameters are quantized and the quantized parameters are used for the speech enhancement. In [4], the VQ is applied to the harmonic amplitude of the speech. Moreover, in [5], LSF parameters are quantized by a perceptually weighted distance function. It should be mentioned that the Euclidean distance function is commonly used in VQ processes [1].

In section 2 of this paper, a new perceptually weighted distance function is proposed for the VQ of the STFT parameters of the speech. Making use of a few MATLAB simulations, both the proposed approach and the conventional Euclidean method are evaluated in section 3 and finally the paper is concluded in section 4.

2 The proposed perceptually weighted distance measure

In the VQ process, the distance measure function has an important rule on the clustering of the training vectors. Conventionally the Euclidean distance function is utilized for this purpose [1]. Assume M vectors with the dimension of n that are defined as follows:

$$X^k = \left(x_1^k,\, x_2^k,\, \ldots,\, x_n^k\right) \quad k \in \{1, \ldots, M\} \tag{1}$$

The conventional Euclidean distance function between the vectors X^k and X^l are described as follows:

$$D\left(k,l\right) = \sum_{i=1}^{n} \left|x_i^k - x_i^l\right|^2 \quad k,l \in \{1, \ldots, M\} \tag{2}$$

The perceptual quality is an important criterion in the speech processing whereas the Euclidean distance function does not include this important principle in its definition. In this section, a new perceptually weighted distance function is proposed for the VQ of the STFT amplitudes. In other words, the STFT amplitudes are the feature vectors in the VQ process of this paper.

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009

2.1 Logarithmic preprocessing

As a first step, the frequency domain vectors of X^k are preprocessed as follows where α is a parameter that adjusts the dynamic range of y_i^k.

$$y_i^k = \log_{10}\left(1 + \alpha \frac{x_i^k}{\max\limits_i \left(x_i^k\right)}\right) \qquad i \in \{1, \ldots, n\},\ k \in \{1, \ldots, M\} \tag{3}$$

This preprocessing adjusts the dynamic range of y_i^k to $[0 \quad \log_{10}(\alpha + 1)]$. As an example, for $\alpha = 10$ and $\alpha = 100$, the dynamic range of y_i^k will be $[0 \quad 1]$ and $[0 \quad 2]$, respectively.

Therefore, the distance function of the preprocessed vectors is defined as follows:

$$D_P(l, k) = \sum_{i=1}^{n} \left|y_i^k - y_i^l\right|^2 \quad k, l \in \{1, \ldots, M\} \tag{4}$$

2.2 Frequency sensitivity of the human ear

In this part, a few listening characteristics of the ear are described. They will be used for the definition of the perceptually weighted distance function. The ear of the human is more sensitive to the low frequency components rather than their high frequency counterparts. In other words, the low frequency components of the speech have more listening data than the high frequency components [6]. Moreover, in the telephony-band, 4 kHz, the frequency components smaller than 100 Hz and greater than 3800 Hz do not have any valuable listening data.

Based on the above facts, in this paper, the telephony-band is partitioned to low, middle, and high sub-bands. For this purpose, we divide the band of interest into three identical sub-bands in the Mel scale. The speech processing in the Mel scale takes into account the frequency sensitivity of the ear [7]. The equation of the Mel scale is shown as follows where f denotes the frequency in hertz [7].

$$mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{5}$$

Hence, the band of interest, 100 Hz-to-3800 Hz, is translated in the Mel scale as follows:

$$f_l^{mel} = mel(f_l = 100\,Hz) = 150 \tag{6}$$

$$f_h^{mel} = mel(f_h = 3800\,Hz) = 2100 \tag{7}$$

Therefore, the bandwidth of each sub-band in the Mel scale, BW^{mel}, is obtained as follows:

$$BW^{mel} = \frac{f_h^{mel} - f_l^{mel}}{3} = 650 \tag{8}$$

Consequently, the boundary frequencies of the sub-bands, shown in Fig. 1, are obtained as follows:

$$f_1^{mel} = f_l^{mel} + BW^{mel} = 150 + 650 = 800 \tag{9}$$

$$f_2^{mel} = f_1^{mel} + BW^{mel} = 800 + 650 = 1450 \tag{10}$$

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009

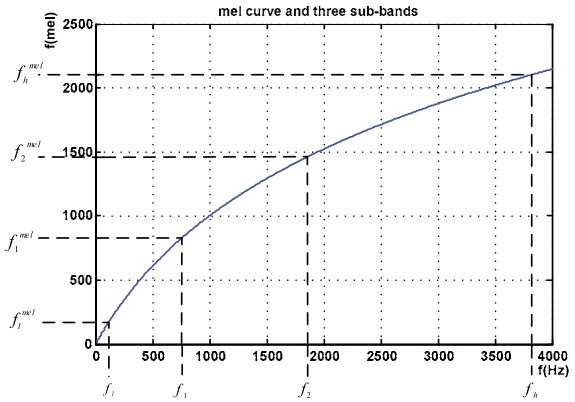

Fig. 1. sub-bands in hertz and Mel scales

Finally making use of relation (5), f_1 and f_2 are determined in hertz as follows:

$$f_1 = 723\,Hz \cong 750\,Hz,\ f_2 = 1835\,Hz \cong 1850\,Hz \tag{11}$$

Moreover, the bandwidth of each sub-band is obtained in hertz:

$$BW_l = 650\,Hz,\ BW_m = 1100\,Hz,\ BW_h = 1950\,Hz \tag{12}$$

In Fig. 1, the sub-bands are shown in both hertz and Mel scales. It is considered that the sub-bands have identical bandwidth in the Mel scale but they have different bandwidth in the hertz scale. It shows that the low frequency components of the speech are more important than the high frequency components. So, in order to define a perceptually weighted distance function, the low, middle, and high sub-bands are weighted with α_l, α_m, and α_h, respectively where the amount of these parameters is chosen according to the importance of the related sub-bands [5]. For this purpose, the weight of each sub-band is selected proportional to the reverse of its bandwidth in hertz. Therefore, we assign the highest value for $\alpha_l(\alpha_l = 1)$ and determine the normalized values of α_m and α_h as follows:

$$\begin{cases} \alpha_l = \dfrac{1}{BW_l} \times BW_l = 1 \\ \alpha_m = \dfrac{1}{BW_m} \times BW_l = 0.59 \\ \alpha_h = \dfrac{1}{BW_h} \times BW_l = 0.33 \end{cases} \tag{13}$$

The above values of α_m and α_h are chosen as initial values in an optimization algorithm. The optimization algorithm varies α_m and α_h around the above initial values and satisfies the constraint of $0 < \alpha_h < \alpha_m < \alpha_l = 1$. In order to determine the optimum values of α_m and α_h, the optimization algorithm utilizes the spectral distortion (SD) criterion [4, 8]. In other words, the optimum values of α_m and α_h lead to the least spectral distortion in our simulations. The optimum values of α_l, α_m, and α_h are as follows:

$$\begin{cases} \alpha_l = 1 \\ \alpha_m = 0.7 \\ \alpha_h = 0.4 \end{cases} \tag{14}$$

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009

Finally the frequency components of each sub-band are weighted as follows where f_s denotes the sampling rate, 8 kHz, and the value of fw_i is determined as shown in relation (16). Because of the symmetry of the Fourier transform, for $\frac{n}{2} < i \leq n$, the amount of fw_i is found in this way: $fw_i = fw_{n-i+1}$.

$$FW = (fw_1,\ fw_2, \ldots, fw_n) \tag{15}$$

$$fw_i = \begin{cases} \alpha_l \ \ldots\ldots\ldots 100\,Hz < \dfrac{i}{n}.f_s < 750\,Hz \\ \alpha_m \ldots\ldots\ldots 750\,Hz < \dfrac{i}{n}.f_s < 1850\,Hz \\ \alpha_h \ \ldots\ldots\ldots 1850\,Hz < \dfrac{i}{n}.f_s < 3800\,Hz \\ 0 \quad \ldots\ldots\ldots otherwise \end{cases} \quad , \quad i \in \left\{1, 2, \ldots, \frac{n}{2}\right\} \tag{16}$$

Thus, the perceptually weighted distance function is proposed as shown bellow:

$$D_F(k,l) = \sum_{i=1}^{n} fw_i \left| y_i^k - y_i^l \right|^2 \ldots\ldots\ldots k, l \in \{1, \ldots, M\} \tag{17}$$

2.3 Amplitude sensitivity of the human ear

The frequency components with large amplitude have more valuable listening data than the frequency components with small amplitude. Based on this fact, another perceptually weighted distance function is also proposed in this paper. For this purpose, the average power of the frequency components is calculated in each sub-band as shown bellow:

$$p_1^{k0} = \frac{1}{b-a} \sum_{i=a}^{b-1} \left| x_i^k \right|^2 \ldots, \ldots p_1^k = \log_{10}(1 + p_1^{k0}) \tag{18}$$

$$p_2^{k0} = \frac{1}{c-b} \sum_{i=b}^{c-1} \left| x_i^k \right|^2 \ldots, \ldots p_2^k = \log_{10}(1 + p_2^{k0}) \tag{19}$$

$$p_3^{k0} = \frac{1}{d-c} \sum_{i=c}^{d-1} \left| x_i^k \right|^2 \ldots, \ldots p_3^k = \log_{10}(1 + p_3^{k0}) \tag{20}$$

where a, b, c, and d are:

$$a = \left[\frac{f_l}{f_s} n\right],\ b = \left[\frac{f_1}{f_s} n\right],\ c = \left[\frac{f_2}{f_s} n\right],\ d = \left[\frac{f_h}{f_s} n\right], \quad [x] = floor\,(x) \tag{21}$$

Making use of the average power of the frequency components, we define another weigh vector as follows:

$$PW^k = \left(pw_1^k, pw_2^k, \ldots\ldots, pw_n^k\right) \tag{22}$$

$$pw_i^k = \begin{cases} p_1^k \ldots\ldots a \leq i < b \\ p_2^k \ldots\ldots b \leq i < c \\ p_3^k \ldots\ldots c \leq i \leq d \\ 0 \ldots\ldots otherwise \end{cases} \quad , \quad i \in \left\{1, 2, \ldots, \frac{n}{2}\right\} \tag{23}$$

$$pw_i^k = pw_{n-i+1}^k \qquad\qquad , \quad i \in \left\{\frac{n}{2}+1, \frac{n}{2}+2, \ldots, n\right\}$$

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009

Thus, another perceptually weighted distance function is proposed as follows:

$$D_A(k,l) = \sum_{i=1}^{n} pw_i^k \left| y_i^k - y_i^l \right|^2 \ldots\ldots\ldots k,l \in \{1,\ldots,M\} \tag{24}$$

2.4 Definition of the final distance function

In order to propose an appropriate perceptually weighted distance function, we should simultaneously consider both the frequency and amplitude sensitivity of the human ear. For this purpose, similar to [5], we multiply pw_i^k to fw_i, and define the overall weight as shown in following relations:

$$W^k = \left(w_1^k, w_2^k, \ldots\ldots, w_n^k\right) \tag{25}$$

$$\mathrm{w_i^k} = pw_i^k . fw_i \tag{26}$$

So, the final proposed perceptually weighted distance function is described as follows:

$$D_w(k,l) = \sum_{i=1}^{n} w_i^k \left| y_i^k - y_i^l \right|^2 \ldots\ldots\ldots k,l \in \{1,\ldots,M\} \tag{27}$$

where y^k is the training vector, and y^l is the approximated vector (codevector) in the VQ process.

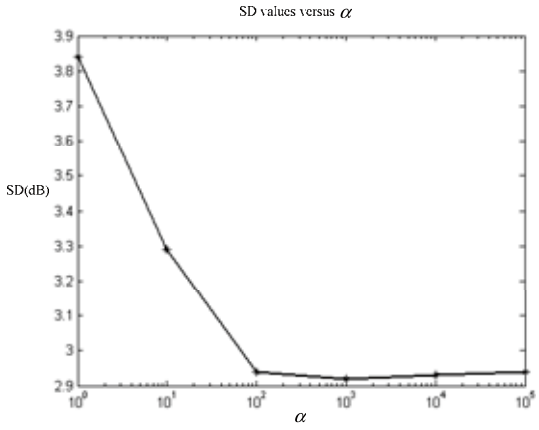

Fig. 2. SD values versus α (parameter of preprocessing)

3 Simulation results

In order to evaluate the proposed distance function, a few MATLAB simulations are performed in this section. In these simulations, the Farsi speech database, Farsdat, is employed. 60 (4×15) sentences from four speakers are utilized for training and consequently a codebook of 1024 vectors is designed. The sampling rate is reduced from 22.5 kHz to 8 kHz. The frame length and frame shift are identical to 20 ms and 10 ms, respectively. Besides; in order to determine the STFT amplitudes, the hamming window and 128-point FFT are utilized. After designing the codebook, the spectral distortion (SD) criterion [4, 8] is utilized for the evaluation of the VQ process. In the evaluation phase, 20 (4×5) different sentences of same speakers are used.

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009

In Fig. 2 the SD values are depicted versus α. It helps the designer perform the preprocessing of the feature properly. So, $\alpha = 1000$ is chosen for the preprocessing of the feature vectors.

In table. 1, the SD is extracted for several VQ processes where each VQ process employs a specific distance function. The simulation results prove that the proposed perceptually weighted distance functions reduce the SD and consequently improve the VQ process.

Table I. SD values for various distance function

Distance measure type	SD (dB)
Simple Euclidean distance measure(D)	3.98
Distance measure by preprocessing of features(D_p)	2.92
Perceptually weigthed distance measure(D_F)	2.60
Perceptually weigthed distance measure(D_A)	2.51
Final Perceptually weigthed distance measure(D_w)	2.19

4 Conclusion

In this paper, a new perceptually weighted distance function is proposed for the VQ of the STFT parameters of the speech signal.

This method performs a preprocessing. Then it partitions the telephony-band to three sub-bands and proposes a weight vector, FW. It also measures the average power of the frequency components of each sub-band and proposes another weight vector, PW^k. Multiplying FW to PW^k, it proposes the final weight vector, W^k.

Making use of the preprocessed features and employing the final weight vector, W^k, the conventional Euclidean distance function is modified and a new perceptually weighted distance function is proposed.

DOI: 10.1587/elex.6.824
Received March 17, 2009
Accepted June 01, 2009
Published June 25, 2009