THE SUCCESSIVE MEAN QUANTIZATION TRANSFORM

Mikael Nilsson, Mattias Dahl, and Ingvar Claesson

Blekinge Institute of Technology
School of Engineering
Box 520, SE-372 25 Ronneby, Sweden
E-mail: mkn@bth.se, mdh@bth.se, icl@bth.se

ABSTRACT

This paper presents the Successive Mean Quantization Transform (SMQT). The transform reveals the organization or structure of the data and removes properties such as gain and bias. The transform is described and applied in speech processing and image processing. The SMQT is considered as an extra processing step for the mel frequency cepstral coefficients commonly used in speech recognition. In image processing the transform is applied in automatic image enhancement and dynamic range compression.

1. INTRODUCTION

A reliable and robust feature extraction is important for pattern recognition tasks. Problems of using different sensors and Analog-to-Digital (A/D) converters will have an impact on the performance of the system. These discrepancies may occur due to difference in the gain and bias in the signals. Another possibility is that the structure or the shape of the signal changes. The aim with the SMQT is to remove the disparity between sensors due to gain and bias. Additionally, the extraction of the structure of the data should be done in an efficient manner. This structure extraction problem can be seen as the problem of dynamic range compression [1]. The finding of structures in data has been proposed before, for example the Census Transform which extracts a binary structure from an image [2]. More recently the Modified Census Transform [3] emerged; this transform has similarities to a first level SMQT. However, these techniques reveal only one bit structures or structure kernels. The SMQT can be used to extend the structure representation to an arbitrary predefined number of bits on arbitrary dimensional data. This will be shown by applying the SMQT in both speech and image processing.

In speech recognition the Mel Frequency Cepstral Coefficients (MFCC) are commonly used as front end and the Hidden Markov Model (HMM) for pattern recognition [4, 5]. The mismatch between training and testing is a common problem. Techniques have been proposed for adjusting the parameters in Hidden Markov Models (HMMs) by Parallel Model Combination (PMC) to overcome this problem. For example, the Jacobian adaptation, fast PMC, PCA-PMC, log-add approximation, log-normal approximation, numerical integration and weighted PMC [6, 7]. These operations are necessary even if the signal changes its bias or gain since a gain or bias change in the signal will propagate to the MFCC. Hence, a SMQT as an extra step in the MFCC calculation will yield a separation between the structure and the level in the signal. These level-free coefficients, denoted SMQT-MFCC, will be compared with standard MFCC.

Producing digital images that render contrast and detail well is a strong requirement in several areas, such as remote sensing, biomedical image analysis and fault detection [8]. Performing these tasks automatically without human intervention is a particularly hard task in image processing. Different approaches and techniques have been suggested for this problem [8, 9, 10, 11]. The SMQT uses an approach that performs an automatic structural breakdown of information. This operation can be seen as a progressive focus on the details in an image.

2. DESCRIPTION OF THE SMQT

Let x be a data point and $\mathcal{D}(x)$ be a set of $|\mathcal{D}(x)| = D$ data points. The value of a data point will be denoted $\mathbf{V}(x)$. The form of the data points can be arbitrary, that is $\mathcal{D}(x)$ could be a vector, a matrix or some arbitrary form. The SMQT has only one parameter input, the level L (indirectly it will also have the number of data points D as an important input). The output set from the transform is denoted $\mathcal{M}(x)$ which has the same form as the input, i.e. if $\mathcal{D}(x)$ is a matrix then $\mathcal{M}(x)$ is also a matrix of same size. The transform of level L from $\mathcal{D}(x)$ to $\mathcal{M}(x)$ will be denoted

$$\mathrm{SMQT}_L : \mathcal{D}(x) \rightarrow \mathcal{M}(x) \tag{1}$$

The SMQT_L function can be described by a binary tree where the vertices are Mean Quantization Units (MQUs).

A MQU consists of three steps, a mean calculation, a quantization and a split of the input set.

The first step of the MQU finds the mean of the data, denoted $\overline{\mathbf{V}}(x)$, according to

$$\overline{\mathbf{V}}(x) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbf{V}(x) \tag{2}$$

The second step uses the mean to quantize the values of data points into $\{0, 1\}$. Let a comparison function be defined as

$$\xi\left(\mathbf{V}(y), \overline{\mathbf{V}}(x)\right) = \begin{cases} 1, & if\ \mathbf{V}(y) > \overline{\mathbf{V}}(x) \\ 0, & else \end{cases} \tag{3}$$

and let $\bigotimes$ denote concatenation, and then

$$\mathcal{U}(x) = \bigotimes_{y \in \mathcal{D}} \xi\left(\mathbf{V}(y), \overline{\mathbf{V}}(x)\right) \tag{4}$$

is the mean quantized set. The set $\mathcal{U}(x)$ is the main output from a MQU. The third step splits the input set into two subsets

$$\begin{aligned} \mathcal{D}_0(x) &= \{x \mid \mathbf{V}(x) \leq \overline{\mathbf{V}}(x), \forall x \in \mathcal{D}\} \\ \mathcal{D}_1(x) &= \{x \mid \mathbf{V}(x) > \overline{\mathbf{V}}(x), \forall x \in \mathcal{D}\} \end{aligned} \tag{5}$$

where $\mathcal{D}_0(x)$ propagates left and $\mathcal{D}_1(x)$ right in the binary tree, see Fig. 1.

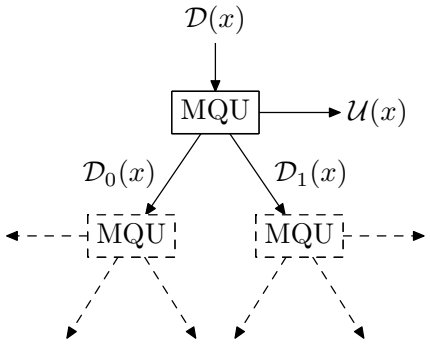

Fig. 1. The operation of one Mean Quantization Unit (MQU).

The output set $\mathcal{U}(x)$ from a MQU is not a value or similarity coefficient as in linear transforms. Instead, $\mathcal{U}(x)$ can be interpreted as the structure of $\mathcal{D}(x)$. Hence, the MQU is independent due to gain and bias adjustments of the input.

The MQU constitutes the main computing unit for the SMQT. The first level transform, SMQT_1, is based on the output from a single MQU, where $\mathcal{U}$ is the output set at the root node. The outputs in the binary tree need extended notation. Let the output set from one MQU in the tree be denoted $\mathcal{U}_{(l,n)}$ where $l = 1, 2, \ldots, L$ is the current level and $n = 1, 2, \ldots, 2^{(l-1)}$ is the output number for the MQU at level l, see Fig. 2.

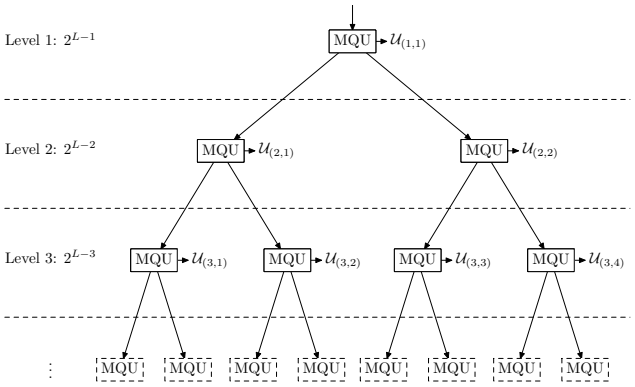

Fig. 2. The Successive Mean Quantization Transform (SMQT) as a binary tree of Mean Quantization Units (MQUs).

Weighting of the values of the data points in the $\mathcal{U}_{(l,n)}$ sets are performed and the final SMQT_L is found by adding the results. The weighting is performed by 2^{L-l} at each level l. Hence, the result for the SMQT_L can be found as

$$\begin{aligned} \mathcal{M}(x) &= \{x \mid \mathbf{V}(x) = \textstyle\sum_{l=1}^{L} \sum_{n=1}^{2^{l-1}} \mathbf{V}\left(u_{(l,n)}\right) \cdot 2^{L-l}, \\ &\quad \forall x \in \mathcal{M}, \forall u_{(n,l)} \in \mathcal{U}_{(l,n)}\} \end{aligned} \tag{6}$$

As a consequence of this weighing the number of quantization levels, denoted Q_L, for a structure of level L will be $Q_L = 2^L$.

The MQU is insensitive due to gain and bias. The MQU is the basic building block of the SMQT. Hence, inductively the SMQT is also insensitive to gain and bias.

3. SMQT IN SPEECH PROCESSING

Acoustic mismatch between the training and test data degrades the performance of Automatic Speech Recognition (ASR) systems [12]. The MFCC are frequently used as a speech parametrization in speech recognizers. The MFCC are calculated from speech frames, hence the mismatch is directly affected by the disparity in these speech frames. The difference between the frames can occur due to different gain and bias in the speech signal. Another possible difference is that the structure or the shape of the speech frame changes. The motivation for the Successive Mean Quantization Transform - Mel Frequency Cepstral Coefficients (SMQT-MFCC) is to remove the gain and bias disparity between training and testing. The basic steps for the calcula-

tion of the MFCC and the SMQT-MFCC can be found in Fig. 3.

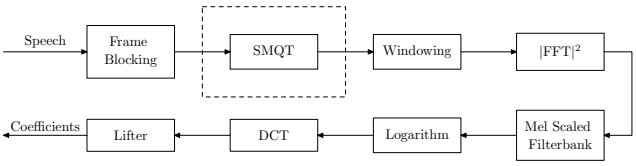

Fig. 3. The steps from speech to coefficients by MFCC and SMQT-MFCC.

A correctly band-limited speech signal sampled at 16 kHz is used as comparison between the MFCC and the SMQT-MFCC. The speech comes from a male speaker pronouncing the word "one". This signal undergoes modifications by gain, bias and additive white Gaussian noise at three Signal-to-Noise (SNR) levels. The MFCC and SMQT-MFCC are calculated for these cases, see Fig. 4. Speech frames of 20 ms are used and a SMQT_8 is applied to the frames.

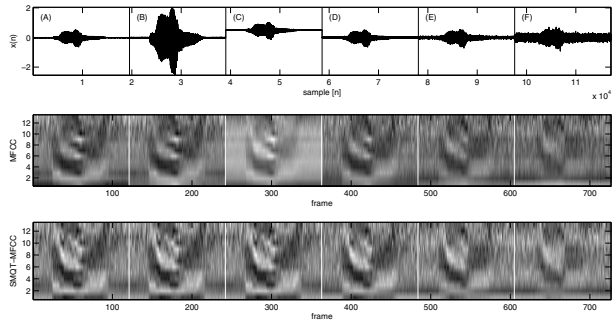

Fig. 4. Comparison between MFCC and SMQT-MFCC. (A) original speech signal $s(n)$, (B) $4s(n)$, (C) $s(n) + 0.5$, (D) $s(n) + w_{20}(n)$ where $w_{20}(n)$ is white Gaussian noise and subindex indicates the SNR level in dB, (E) $s(n) + w_{10}(n)$ and (F) $s(n) + w_0(n)$. Coefficients have been normalized for better visualization.

.

The SMQT-MFCC have an exact match between (A), (B) and (C) while the MFCC does not. This since the SMQT is independent of gain and bias. However, both the SMQT-MFCC and the MFCC are affected in different degrees by the white Gaussian noise.

It should be emphasized that it is the separation between the structure and the levels of a signal that can be useful in speech processing. Hence, an augmentation of the SMQT-MFCC with information about the level of the signal might be desired.

4. SMQT IN IMAGE PROCESSING

Transforming a whole image at different levels gives an illustrative description of the operation for the SMQT, see Fig.5.

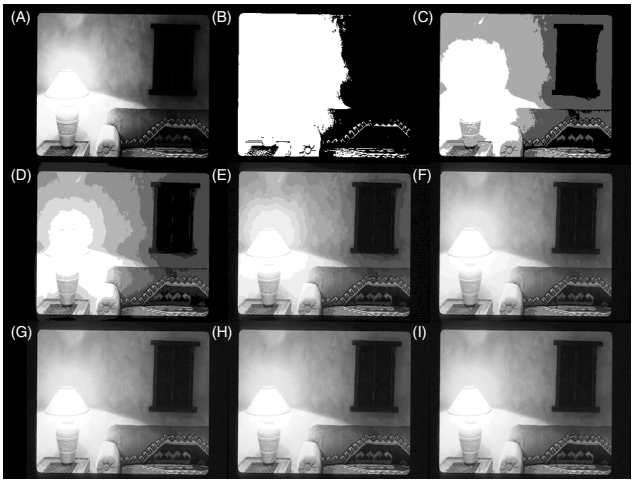

Fig. 5. Example of SMQT on whole image. (A) original image, dynamic range 0-255 (8 bit). (B)-(I) corresponds to $\mathrm{SMQT}_1 - \mathrm{SMQT}_8$ of image.

The level L in the SMQT_L denotes the number of bits used to describe the transformed image. Hence, a SMQT_1 of the image has a one bit representation $\{0, 1\}$ and a SMQT_2 of the image has two bits $\{0, 1, 2, 3\}$ see (B) and (C) respectively. Choosing a level of the transform lower than the number of bits in the original image yield a dynamic range compressed image. A SMQT_8 of an image, which has a dynamic range represented by 8 bits, will yield an uncompressed image with enhanced details. A comparison with a histogram equalization [9] is conducted, see Fig. 6.

The histogram equalization has some problems with oversaturation and artifacts in several areas area in the images. Notice how the histogram equalized images have a tendency to get washed out or unnatural. These effects do not occur, or are very limited, in the SMQT enhanced images. The SMQT also has less computational complexity and fewer adjustments compared to more advanced enhancement techniques such as [8, 10, 11].

5. CONCLUSIONS

This paper presents a new transform, denoted the successive mean quantization transform. The SMQT has properties that reveal the underlying organization or structure of data. The transform extracts the structure in a robust manner which makes it insensitive to changes in bias and gain in the signal.

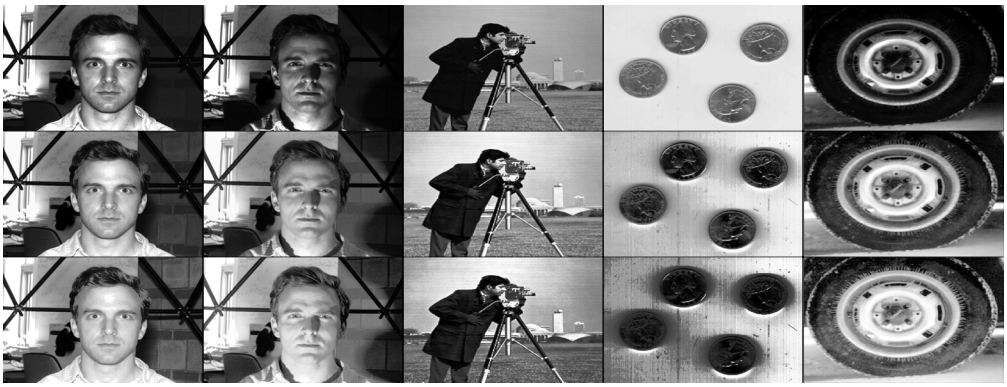

Fig. 6. **Top** row original image, **middle** SMQT_8 of image and **bottom** histogram equalization of image. Face images are from Yale Face Database B [13]. Zooming in the pdf images is recommended for detail studies. Note the differences of the forehead in the face images.

.

The transform has been applied as an extension of the MFCC and the SMQT-MFCC are introduced. A comparison between the MFCC and the SMQT-MFCC has been conducted. The benefit of the SMQT-MFCC is that it extracts only the structure and ignores the level in the signal. This implies that the SMQT-MFCC are robust to bias and gain dissimilarities in speech signals. The transform has also been applied for automatic enhancement of images. A comparison with a histogram equalization has been performed, which showed the advantage of the SMQT.

6. REFERENCES

[1] R.J. Cassidy, "Dynamic range compression of audio signals consistent with recent time-varying loudness models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004, vol. 4, pp. 213–216.

[2] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *ECCV (2)*, 1994, pp. 151–158.

[3] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, May 2004, pp. 91–96.

[4] Deller John R. Jr., Hansen John J.L., and Proakis John G., *Discrete-Time Processing of Speech Signals*, IEEE Press, 1993, ISBN 0-7803-5386-2.

[5] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993, ISBN 0-13-015157-2.

[6] H. K. Kim and R. C. Rose, "Cepstrum-domain model combination based on decomposition of speech and noise for noisy speech recognition," in *Proceedings of ICASSP*, May 2002, pp. 209–212.

[7] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for hmm recognition in noise. speech communication," *Speech Communication*, vol. 12, no. 3, pp. 231–239, July 1993.

[8] C. Munteanu and A. Rosa, "Towards automatic image enhancement using genetic algorithms," *Proceedings of the 2000 Congress on Evolutionary Computation*, vol. 2, pp. 1535–1542, July 2000.

[9] William K. Pratt, *Digital Image Processing*, John Wiley & Sons, 3rd edition, 2001.

[10] Z. Rahman, D.J. Jobson, and G.A. Woodell, "Multi-scale retinex for color image enhancement," *International Conference on Image Processing*, vol. 3, pp. 1003–1006, September 1996.

[11] D.J. Jobson, Z. Rahman, and G.A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Transactions on Image Processing*, vol. 6, pp. 451–462, March 1997.

[12] Y. Gong, "Speech recognition in noisy environments: A survey," in *Speech Communication*, 1995, vol. 16, pp. 261–291.

[13] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Generative models for recognition under variable pose and illumination," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 277–284.