MULTICHANNEL AUDIO SIGNAL COMPRESSION BASED ON TENSOR DECOMPOSITION

Jing Wang, Chundong Xu, Xiang Xie, Jingming Kuang

School of Information Science and Technology, Beijing Institute of Technology, Beijing, China

ABSTRACT

This paper proposes a novel multichannel audio signal compression method based on tensor decomposition. The multichannel audio tensor space is established with three factors (channel, time, and frequency) and is decomposed into the core tensor and three factor matrices based on tucker model. Only the truncated core tensor is transmitted to the decoder which is multiplied by the factor matrices trained before processing. The performance of the proposed method is evaluated with approximation errors, compression degree and listening tests. When the core tensor is smaller, the compression degree will be higher. A very noticeable compression capability will be achieved with an acceptable retrieved quality. The novelty of the proposed method is that it enables both high compression capability and backward compatibility with little signal distortion to the hearing.

Index Terms— Multichannel, audio signal compression, tensor decomposition, tucker model, core tensor

1. INTRODUCTION

In the application field of digital audio, multichannel audio can provide more realistic surround experiences which stereo audio signals may fail to provide. As the need for enjoying high quality of digital audio signals, so does the need for more efficient audio compression technology. Many researches on digital audio coding are mono or stereo audio compression techniques. In order to generate surround effect with more channels, a number of multichannel audio storage techniques such as Dolby AC-3 [1], DTS (Digital Theatre System) [2], MPEG Surround (also known as SAC-Spatial Audio Coding) [3,4,5] have been proposed by using perceptual coding, transform or filter bank theory.

Researches towards developing lower rate coders for multichannel surround audio systems will be stronger in many multimedia interactive applications such as virtual reality, teleconference and 3D game playing. The data size of multichannel audio signals is much higher compared to that of the stereo audio signals mostly because of the large number of channels. The work in this paper focuses on incorporating multilinear analysis [6] and tensor decomposition [7] for representing and compressing multichannel audio signals, which has not been considered in the relative prior work.

Multilinear analysis has been proposed to manipulate the higher-order tensor structure of the observations by tensor algebra. The higher-order tensors are equivalents of multidimensional matrices, or multiway arrays, and have gained a lot of importance in the field of array data analysis. In our proposed method, the input multichannel audio signal is represented by 3-order tensor with three factors: channel, time and frequency. Then the tensor space is compressed into fewer channels and fewer frequency spectrum parameters by the way of low rank approximation based on tucker decomposition [7]. The multichannel audio can be recovered by the transmitted core tensor which is produced by tucker decomposition at the encoder and three factor matrices which are built from a set of training data. With decreasing the dimension of channel and frequency mode, high compression degree can be achieved with acceptable audio quality. This novel multichannel audio signal compression method is based on higher-order tensor algebra, satisfying both high compression capability and backward compatibility in itself without additional side information.

The remainder of this paper is organized as follows. Section 2 gives the preliminary tensor algebra that is used in this paper. Section 3 describes the tensor space construction of the multichannel audio signal. Section 4 presents the compression method based on tucker decomposition in detail. Section 5 shows the experiment designing and some test results; Section 6 concludes this paper.

2. PRELIMINARY TENSOR ALGEBRA

2.1. Basic tensor operations

A tensor is a multidimensional array which represents an element of *N*-order multifactor space. The order of tensor is the number of dimensions or factors, also known as mode or way. The two important and basic operations of tensor are mode-n unfolding and mode-n product.

The mode-n unfolding (mode-n flattening) is the matricization of the tensor $\mathrm{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$ in the subspace of mode n, which can be expressed as the flattened matrix $X_{(n)} \in \mathbb{R}^{I_n \times \bar{I}}$, where $\bar{I} = \prod_{m \neq n} I_m$.

The mode-n product includes tensor timing vector, matrix and another tensor, which can be processed based on the tensor matricization. In this paper, we only make use of a tensor times a matrix. The mode-n product of an *N*-order tensor $\mathrm{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$ by a matrix $U \in \mathbb{R}^{J \times I_n}$ can be denoted by $\mathrm{Y} = \mathrm{X} \times_n U$, which is still an *N*-order tensor of size $I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots I_N$. And this kind of tensor product can be expressed by matrix product using the mode-n unfolding as follows

$$\mathrm{Y} = \mathrm{X} \times_n U \Leftrightarrow Y_{(n)} = UX_{(n)} \tag{1}$$

2.2. Relative tensor decompositions

There are many choices for tensor decompositions which generally combine a choice of orthonormal bases in the domain of tensor with a suitable truncation of its expansion. Two main kinds of tensor decompositions are CP (CANDECOMP/PARAFAC) and Tucker decomposition [7]. The latter one can be regarded as a multilinear generalization of the traditional matrix SVD (Singular Value Decomposition) [8] and plays an important role in tensor-based signal processing.

For an *N*-order tensor, a low rank-approximation with the truncated tucker decomposition is represented as

$$\mathrm{X} \approx \mathrm{G} \times_1 U^{(1)} \times_2 U^{(2)} \cdots \times_N U^{(N)} \tag{2}$$

Where $\mathrm{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$, $U^{(n)} \in \mathbb{R}^{I_n \times R_n} \left(n = 1, 2, ..., N; R_n \le I_n\right)$ are the truncated components or factor matrices (usually orthogonal matrices) in mode-1, mode-2 and mode-n subspaces, respectively. $\mathrm{G} \in \mathbb{R}^{R_1 \times R_2 \times \cdots R_N}$ is the core tensor whose entries show the level of interaction between the different components, and is computed with

$$\mathrm{G} = \mathrm{X} \times_1 U^{(1)\mathrm{T}} \times_2 U^{(2)\mathrm{T}} \cdots \times_N U^{(N)\mathrm{T}} \tag{3}$$

One advantage of tucker decomposition is that it can transform the original tensor into the core tensor with factor matrices. It is very useful in low rank approximation [8] and dimensionality reduction.

3. REPRESENTATION OF MULTICHANNEL AUDIO SIGNALS WITH TENSOR SPACE

In the latest 10 years, tensor algebra has been successfully used in the field of signal processing. When the signal information depends on more than one factor, a kind of tensor space can be established with different factors that stand for different subspaces. For example, TensorFaces can be used for face recognition [9,10] , which are established based on four factors: subject , expression, viewpoint and illumination. Recently, multilinear algebra with tensor representation has been attempted more and more in speech and audio signal processing [11,12,13,14].

This paper represents the mulitichannel audio space as 3-order tensor $\mathrm{T} \in \mathbb{R}^{N_c \times N_t \times N_f}$ with three factors: channel (*c*), time (*t*) and frequency (*f*). Here, N_c, N_t and N_f are the dimension of channel, dimension of time and dimension of frequency, respectively. There are more than two channels for each audio file, e.g. the signal with 5.1 channels has 6 channels including FL (front left), FR (front right), FC (front center), LFE (low frequency effects), BL (back left) and BR (back right). Each channel has a sequence of signal frames along the time axis. Each frame can be transformed into spectrum along frequency axis. There are several commonly used transforms, including the Discrete Cosine Transform (DCT), the Fourier Transform (FT), and the temporal-adaptive transform. This paper uses DCT to obtain positive frequency spectral values at the encoder and uses overlap-and-add technique [15] to remove the waveform discontinuity after IDCT (Inverse DCT) at the decoder.

Thus multilinear analysis can be carried out based on the multichannel audio tensor space. Most of the multichannel signals have spectrum correlations between channels and the spectrum amplitude correlations between frequencies. Some special audio signals also have spectrum correlations between frames, but should be seriously stationary along time. The work presented in this paper will focus on removing the interchannel and the intrachannel redundancy for multichannel audio signals based on low rank approximation with tucker decomposition.

4. THE PROPOSED COMPRESSION METHOD

4.1. General description of the scheme

The procedure of multichannel audio signal compression includes encoding and decoding. At the encoder, the input multichannel audio signal will firstly be transformed into frequency domain by DCT for all the channels and then tensor audio space (T) will be constructed and decomposed with tucker model which generates core tensor to be quantized and transmitted. The core tensor (S) looks like a kind of downmixed signal with fewer channels except that it is not an audio signal but stands for the compressed space of DCT spectrum parameters in the proposed scheme. Also the three factor matrices (U_c, U_t, U_f) will be generated by tucker decomposition, which need not to be transmitted to the decoder and are modeled by the training procedure. At the decoder, the received core tensor will be used to reconstruct audio tensor space based on the tucker model with the pre-trained factor matrices. At the end, the multichannel audio signal will be retrieved from the reconstructed tensor space through IDCT and overlap-add procedure. Fig.1 shows the multichannel audio signal encoding and decoding procedure based on tensor decomposition.

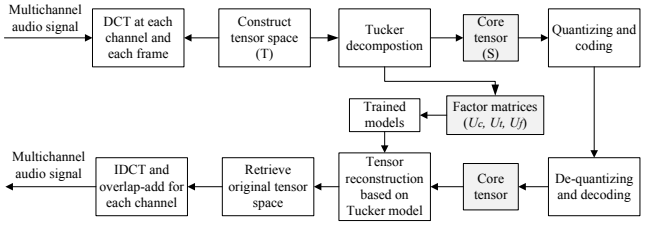

Fig.1. The proposed compression scheme.

4.2. Low rank approximation with tucker decomposition

In order to compress the large information in multichannel audio signal, we propose to construct the audio tensor space and approximate it with low-rank tucker model which decomposes a tensor into a core tensor multiplied by a matrix along each mode.

The 3-order audio tensor space is $\mathrm{T} \in \mathbb{R}^{N_c \times N_t \times N_f}$ and the low rank approximation with tucker decomposition is

$$\mathrm{T} \approx \mathrm{S} \times_1 U_c \times_2 U_t \times_3 U_f = [\![\mathrm{S}; U_c, U_t, U_f]\!] \qquad (4)$$

Here, $U_c \in \mathbb{R}^{N_c \times R_c} \left(R_c \le N_c \right)$, $U_t \in \mathbb{R}^{N_t \times R_t} \left(R_t \le N_t \right)$, and $U_f \in \mathbb{R}^{N_f \times R_f} \left(R_f \le N_f \right)$ are the factor matrices (columnwise orthogonal) and $\mathrm{S} \in \mathbb{R}^{R_c \times R_t \times R_f}$ are the core tensor.

And we will set up the above tucker model through the truncated HOSVD (Higher-Order SVD) [7] which computes the leading left singular vectors of the flattened matrix in each mode. When the factor matrices have been decided, the core tensor S can be obtained according to the equation (3).

4.3. Training of factor matrices in different modes

As is shown in Fig. 1, the three factor matrices (U_c, U_t, U_f) can be pre-trained before coding procedure. They stand for the principle components of the mode-n unfolding matrix and need not to be transmitted to the decoder. When training, the original dimensions (N_c, N_t and N_f) will be truncated to be lower dimensions (R_c, R_t and R_f). The different combination of dimensions (N_c, N_t, N_f, R_c, R_t and R_f) will lead to different trained models which should meet with the requirements when multiplying the core tensor S.

In this paper, we use some multichannel audio signals which are different from the test signals to train the factor matrices obtained from tucker decomposition. For each training sample, a series of factor matrices will be generated based on the tucker decomposition. In order to eliminate the influences of music types, the different series of factor matrices obtained from different training samples will be averaged to generate the final trained models used for tensor reconstruction at the decoder.

4.4. Transmission of core tensor information

As is shown in Fig. 1, only core tensor should be transmitted to the decoder and be multiplied by the factor matrices with the trained models. The core tensor is smaller than the original audio tensor, i.e. has lower dimensions (R_c, R_t and R_f) along different modes. This paper will mainly investigate the compression degree caused by the tensor decomposition and will use 16 bits uniform PCM (pulse code modulation) method for the parameters' quantizing and coding.

5. EXPERIMENT RESULTS

5.1. Experiment design

Multichannel audio sources can be roughly classified into three categories [16]. Audio of class III consists of material recorded in a real space with multiple microphones such as DVD-audio and has considerably larger redundancy inherent among channels than that of class I and class II. In order to investigate the removal of interchannel redundancy, we only use the materials that belong to class III. We randomly selected 16 multichannel audio files from HIFI music soundtracks that stored in 5.1 channels DVD. In which, 10 files are used for training and 6 files are used for testing. The experimental results will be averaged among the test files.

These audio files have 6 channels and 48 kHz sampling rate at a typical coding rate of 64 kbit/sec/ch. The window length is set to 960 samples (20ms) with 50% overlapping and the 960-points DCT will be used to get the spectral parameters. There are at most 899 overlapped frames in each channel and each audio file. When constructing the tensor space $\mathrm{T} \in \mathbb{R}^{N_c \times N_t \times N_f}$, the original dimensions are set to be N_c=6, N_t =899, and N_f=960.

We have found that the available truncating in the way of time is very different for different audio samples and will seriously affect the retrieved quality of the multichannel audio signal. In the experiments, we only investigate the performance of tensor truncating in the way of channel and frequency. Thus the truncated dimension R_t is set to be equal to the original dimension N_t along the time axis. The other two truncated dimensions R_c, R_f can be set to be different values to investigate the performances below.

5.2. Relative approximation error

In the applications of tensor decomposition, the relative approximation error between the low rank tensor $\mathrm{T_r}$ and the original one T can be expressed in Frobenius norm

$$e = \left\| \mathrm{T} - \mathrm{T_r} \right\| / \left\| \mathrm{T} \right\| \qquad (5)$$

Here, the Frobenius norm of tensor $\mathrm{T} \in \mathbb{R}^{N_c \times N_t \times N_f}$ is the square root of the sum of the squares of all its elements z, i.e.,

$$\|\mathrm{T}\| = \sqrt{\sum_{i_1=1}^{N_c}\sum_{i_2=1}^{N_t}\sum_{i_3=1}^{N_f} z_{i_1 i_2 i_3}^2} \tag{6}$$

Table 1 Approximation errors with different R_c and R_f

R_c \ R_f	800	400	200
6	0.002982	0.054510	0.151910
4	0.002989	0.054516	0.151916
2	0.003004	**0.054517**	0.151917
1	0.493783	0.599367	0.620283

From Table 1, we can see that the approximation errors are very small when only measuring the tensor reconstruction performance not the audio quality. And the errors are very close among 2 channels, 4 channels and 6 channels, which means that 5.1 channels of class III can be truncated to 2 channels with little distortion. Also the approximation errors can be acceptable (below 0.1) when there are at least 2 channels and 400 DCT points.

5.3 Compression capability

In order to further understand the storage space savings of the multichannel audio information， the compression degree is calculated according to the following formulas

$$\begin{aligned} r &= \frac{b1-b2}{b1}\times 100\% \\ &= \frac{N_c \times N_t \times N_f - R_c \times R_t \times R_f}{N_c \times N_t \times N_f}\times 100\% \\ &= \left(1-\frac{R_c \times R_f}{N_c \times N_f}\right)\times 100\% \,(\text{with } N_t = R_t) \end{aligned} \tag{7}$$

Where $b1$ and $b2$ are the bits for encoding parameters in the original and the improved scheme, respectively. In order to evaluate the compression performance caused by the tensor decomposition, the original audio tensor and the low rank tensor are both quantized with PCM method. Thus the compression degree can be expressed with the dimension of each mode as shown in equation (7). With R_c and R_f smaller, the compression degree is higher.

Table 2 Compression degree with different R_c and R_f

R_c \ R_f	800	400	200
6	16.7%	58.3%	79.2%
4	55.6%	72.2%	86.1%
2	72.2%	**86.1%**	93.1%
1	13.9%	93.1%	96.5%

From Table 2, we can see that higher compression degree is obtained when the truncated dimension is lower which means the core tensor is smaller. For example, the compression degree arrives at a very high degree of 86.1% at the condition of 2 channels and 400 DCT points.

5.4. Subjective quality test

We carried out MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) [17] subjective listening test to evaluate the retrieved audio quality at different conditions with 5.1 channels loudspeakers. Fig.2 shows the mean MUSHRA score and the 95% confidential interval with 6 test audio files and 10 listeners.

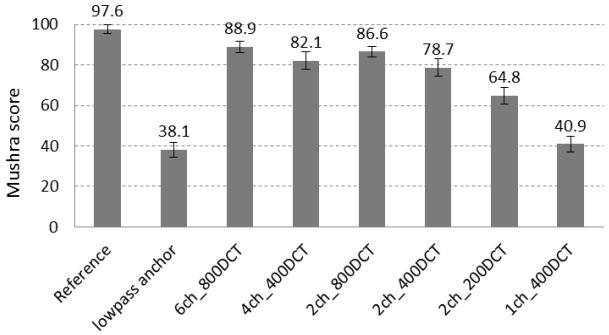

Fig.2. MUSHRA test results of different conditions.

By subjective listening test, we can see that the retrieved audio quality can be acceptable when the channels are truncated very largely. Referring to Table 2, when the compression degree is increased to 86.1%, the subjective audio quality can also be acceptable with an average score of 78.7 at the condition of 2ch_400DCT.

6. CONCLUSIONS

This paper proposed a novel method for multichannel audio signal compression by using tensor decomposition for deriving the truncated core tensor to be transmitted and the factor matrices to be pre-trained. The multichannel audio signal was decomposed and retrieved with tucker model. The experiment results showed that the proposed tensor-based compression method can achieve noticeable high compression capability with acceptable listening quality. For further improvements of the multichannel audio signal compression, the optimization of tensor decomposition and the optimized tensor rank are the key issues to be investigated. Also, the parameters' coding method and the factor matrices' training method are important to advance the compression capability.

7. ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their suggestions. The work in this paper is supported by National Natural Science Foundation of China (No.11161140319 and No.61001188) and Doctoral Fund of Ministry of Education (No. 20101101110020).

8. REFERENCES

[1] C. C. Todd, G. A. Davidson, M. F. Davis, L. D. Fielder, B. D. Link, and S. Vernon, "AC-3: Flexible Perceptual Coding for Audio Transmission and Storage," *96th AES Convention*, preprint 3796, 1994.

[2] DTS company, http://www.dts.com/, 2012.

[3] C. Faller, "Coding of Spatial Audio Compatible with Different Playback Formats," *117th AES Convention*, San Francisco, 2004.

[4] ISO/IEC JTC1/SC29/WG11, "Tutorial on MPEG Surround Audio Coding," *ISO/IEC*, 2005.

[5] ISO/IEC 23003-1:2007, "Information technology-MPEG Audio Technologies -- Part 1: MPEG Surround," *ISO/IEC*, 2007.

[6] L. De Lathauwer, B. De Moor, J. Vandewalle. "A Multilinear Singular Value Decomposition," *SIAM J. Matrix Anal. Appl.*, vol.21, pp. 1253–1278, 2000.

[7] T. G. Kolda, B. W. Bader. "Tensor Decomposition and Applications," *SIAM REVIEW*, vol. 51, no. 3. pp. 455–500, 2009.

[8] S. Weiland, F. van Belzen. "Singular Value Decompositions and Low Rank Approximations of Tensors," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp.1171-1182, MARCH 2010.

[9] M. Alex O. Vasilescu, D. Terzopoulos, "Multilinear Analysis of Image Ensembles: Tensorfaces" in *the Proc. of the European Conference on Computer Vision (ECCV'02)*, Copenhagen, Denmark, pp.447-460, 2002.

[10] M. Alex O. Vasilescu, D. Terzopoulos, "Multilinear Image Analysis for Facial Recognition," in *the Proc. of the International Conference on Pattern Recognition (ICPR'02)*, Quebec City, Canada, pp. 511-514, 2002.

[11] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of Speech from Nonspeech based on Multiscale Spectro-temporal Modulations," *IEEE Transations on Audio, Speech and Language Processing*, vol. 14, no. 3, pp.920-930, 2006.

[12] Q. Wu, L. Zhang and G. Shi, "Robust Multifactor Speech Feature Extraction Based on Gabor Analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol.19, no.4, pp. 927-936, 2011.

[13] Y. Jeong, "Speaker Adaptation based on the Multilinear Decomposition of Training Speaker Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, USA: IEEE, pp.4870-4873, 2010.

[14] D. Saito, K. Yamamoto, N. Minematsu and K. Hirose, "One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space," *Interspeech2011*, Florence, Italy, pp.653-656, 2011.

[15] B. Edler, "Coding of Audio Signals with Overlapping Block Transform and Adaptive Window Functions," *Frequenz*, vol. 43, pp. 252-256, 1989 (in German).

[16] Dai Tracy Yang, Chris Kyriakakis, and C.-C. Jay Kuo, *High-Fidelity Multichannel Audio Coding*, Hindawi Publishing Corporation, New York, USA, 2006.

[17] ITU-R Recommendation BS.1534, " Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems, " *International Telecommunication Union*, Geneva, Switzerland, June 2001.