A Novel 2kb/s Waveform Interpolation Speech Coder Based on Non-negative Matrix Factorization

Peng Zhang, Chang-chun Bao

Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100022, China

pengzhang@emails.bjut.edu.cn, baochch@bjut.edu.cn

10.21437/Interspeech.2007-468

Abstract

In this paper, a 2kb/s Waveform Interpolation speech coder is proposed based on non-negative matrix factorization (NMF). In characteristic waveforms (CWs) decomposition, band-partitioning initialization constraints were set to basis vectors before NMF was carried out. This decomposition method only requires speech signal from the current frame, and can yield high decomposition quality with low computational complexity. Besides, the high dimensional CWs matrix can be expressed by the low dimensional coding matrix, and this has facilitated the CWs quantization. The listening test shows that the proposed 2kb/s NMF-WI coder can give smooth speech with quality close to 2.4kb/s SVD-based WI coder.

Index Terms: Speech coding， waveform interpolation, characteristic waveforms decomposition, non-negative matrix factorization，vector quantization

1. Introduction

In Waveform Interpolation (WI) coder [1], how to decompose and quantize the characteristic waveforms with low delay, low complexity and high precision has always been a hot research topic. There have been three successful CWs decomposition methods proposed so far, i.e. simple FIR filtering decomposition [2], wavelet transform decomposition (WTD) [3] and singular value decomposition (SVD) [4,5]. The original CWs decomposition method of WI involves simple filtering of the characteristic waveforms surface in the evolution domain. The decomposition result is a slowly evolving waveform (SEW) representing the periodic component of speech, and a rapidly evolving waveform (REW) containing the random noise-like component [2]. To further decompose the signal, a multi-level wavelet decomposition mechanism is proposed [3]. In contrast with simple FIR filtering decomposition, advantages of this decomposition approach include scalability in quantization, multi-scale signal evolution analysis. Subsequently, another method exploits the evolution of CW surfaces but uses the decomposition characteristics of Singular Value Decomposition [4, 5] in place of linear filtering. This scheme requires only a single frame of speech and produces a scalable decomposition, and therefore, allows reconstruction accuracy to be varied according to the bit rate available. Although having many advantages in decomposing the CW surfaces of speech signal, the method of SVD has a very high level of computational complexity, which constrains its application. All these methods aim to decompose the CWs into periodic and noise-like components so that they all yield holistic representations of CWs. However this paper will show that NMF [6] can learn the parts of CWs magnitude spectra and give CWs parts-based representations in frequency domain. Besides, these earlier decomposition techniques respectively have their own drawbacks: simple FIR filtering and WTD all bring additional delay; and the computational complexity of SVD is very high. The proposed NMF-based decomposition only requires speech signal from the current frame. And by choosing proper factorization rank, it can achieve almost the same reconstruction accuracy as SVD-based method under keeping much lower computational complexity. The coding matrix of CWs magnitude spectrum is quantized by splitting matrix quantization method and the CWs phase spectrum is restored by fixed phase spectrum. Subjective A/B listening test indicates that the speech quality of our 2kb/s NMF-WI is close to that of 2.4kb/s SVD-WI [7], whose speech quality is better than 2.4kb/s MELP.

The rest of this paper is organized as follows: In section 2, some basic theory of non-negative matrix factorization is recalled. In section 3, the concept of the parts of speech spectra is explained. In section 4, the proposed NMF-based CWs decomposition method is introduced. The quantization scheme and the bit allocation are given in section 5. The subjective A/B listening test is finished in section 6, and finally, the conclusion and future work is summarized in section 7.

2. Non-negative matrix factorization

NMF was first proposed by Lee and Seung. Starting with a non-negative n by m matrix $\mathbf{V}$, the goal of NMF is to approximate it as a product of two non-negative matrices $\mathbf{W}\in\mathbf{R}^{n\times r}$ and $\mathbf{H}\in\mathbf{R}^{r\times m}$, where $\mathbf{W}$ and $\mathbf{H}$ are called basis matrix and coding matrix respectively. Here r is the so-called factorization rank and is chosen so that $(m+n)r<mn$, which means $r < \min(m, n)$. So the product $\mathbf{WH}$ can be regarded as a compressed form of the data in $\mathbf{V}$. In NMF, no negative values are allowed in matrix factors $\mathbf{W}$ and $\mathbf{H}$. The non-negativity constraint is imposed in factorizing the data matrix $\mathbf{V}$ by limiting data manipulation only to additions; no subtractions are allowed. The reconstruction of an object is performed only by adding its representative parts collectively. Since the columns of $\mathbf{H}$ are one-to-one correspondence with the columns of $\mathbf{V}$, the results can be interpreted as each column of $\mathbf{V}$ being described as weighted of a few basis vectors, the weights being the corresponding column of $\mathbf{H}$. In this paper, we use the square of the Euclidean distance between $\mathbf{V}$ and $\mathbf{WH}$ as the cost function, which can be written as:

$$\left\|\mathbf{V}-\mathbf{WH}\right\|^2=\sum_{ij}(\mathbf{V}_{ij}-(\mathbf{WH})_{ij})^2 \qquad (1)$$

Lee and Seung have proved that the square of Euclidean distance between $\mathbf{V}$ and $\mathbf{WH}$ is nonincreasing under the

This work was supported by the Science and Technology Project of Beijing Municipal Education Commission (KM200710005001), the National Natural Science Foundation of China, and the Natural Science Foundation of Beijing under Grant 4042009.

following update rules [8]:

$$\mathbf{H}_{a\mu} \leftarrow \mathbf{H}_{a\mu} \frac{(\mathbf{W}^T\mathbf{V})_{a\mu}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{a\mu}} \tag{2}$$

$$\mathbf{W}_{ia} \leftarrow \mathbf{W}_{ia} \frac{(\mathbf{V}\mathbf{H}^T)_{ia}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T)_{ia}} \tag{3}$$

3. Parts of speech spectra

It is obvious that the "parts" for facial images are several versions of mouths, noses and other facial parts, but what may the "parts" be when referring to the speech spectra? As we know, the human inner ear consists of the cochlea, which is a fluid-filled chamber partitioned by the basilar membrane, and cochlear nerve [9]. The mechanical vibrations impinging on the oval window at the entrance to the cochlea create standing waves (of the fluid inside the cochlea) that cause the basilar membrane to vibrate at frequencies commensurate with the input wave frequencies and at a place along the basilar membrane that associate these frequencies [9]. So the basilar membrane is characterized by a set of frequency responses at different points along the membrane. Hence, the cochlea can be modeled as a bank of filters (appropriately called cochlea filters) [9]. So it can be concluded that our auditory perception is parts-based in frequency domain. From this perspective, the "parts" of speech spectra can be regarded as some localized pulse-like harmonics along the frequency axis and the speech CWs can be represented by these various "parts" in frequency domain.

4. Finding speech spectra parts by NMF

In this paper, the characteristic waveform sampling rate was 400 Hz and 10 CWs were extracted per frame. After the CWs extraction, the CWs were represented by Fourier-series, and then converted to magnitude and phase spectra by polar coordinates transformation:

$$\begin{aligned} &| CW_k(n) | = \sqrt{DTFS_A_k^2(n) + DTFS_B_k^2(n)} \\ &k = 1, 2, \cdots, P(n)/2 \end{aligned} \tag{4}$$

Here, the $p(n)$ is the pitch of the current sub-frame of CW. The NMF training procedures were then carried out on the CWs magnitude spectra. In the earlier decomposition frameworks, i.e. simple FIR filtering, WTD and SVD, the CWs should be well aligned before decomposition in order to obtain smooth and highly correlated CWs. However, since NMF is parts-based and the goal is to find the localized "pulse-like" harmonics, the proposed NMF-based decomposition removed the CWs alignment that has great computational complexity.

Because the dimension of the CWs varies along with the pitch period, which ranges from 20 to 120 in our WI coder, we first classified the CWs into 9 categories according to their maximum sub-frame pitch periods. For the first 8 categories, the spans of pitch period were all set to be 10 samples, i.e. pitch from 20 to 30 classified as the first category, 30 to 40 as the second and so on, and the maximum sub-frame pitch period in the 9th category ranges from 100 to 120. After CWs classification, the NMF training procedures were carried out on the spectra data base of each category, which means we will obtain totally 9 different basis matrices for all the categories. It is straightforward that CWs classification before training will make the basis matrices be more adaptive to the variation of harmonic numbers.

Before training, we should decide the factorization rank r, which is expected to be as small as possible. Lee and Seung in [6] have suggested that the rank r could be chosen so that $(n+m)r<nm$. To meet this, r should not be bigger than 8 in the CWs decomposition, however, experiments showed that even rank 8 was too small to reconstruct the original CWs spectra with tolerable errors. So in this paper, the rank was chosen to be 16 as a trade-off between computational complexity and decomposition quality.

Figure1: Basis vectors trained by NMF with no initialization constraints
X-axis: harmonic number Y-axis: amplitude

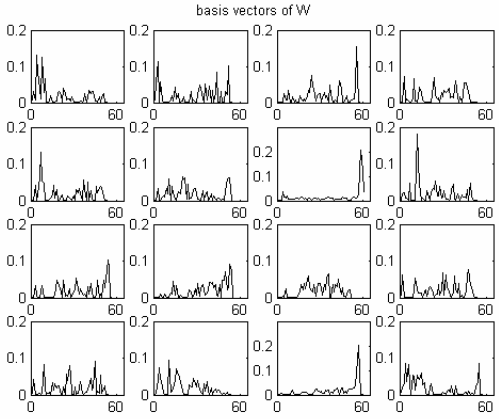
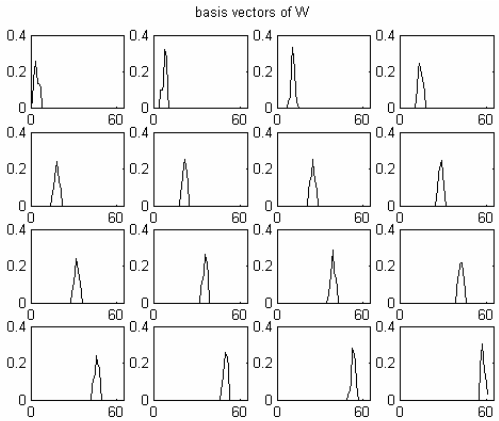


Figure2: Basis vectors trained by NMF with band-partitioning initialization constraints
X-axis: harmonic number Y-axis: amplitude



When training the matrix **W**, differing from the original NMF, in which all the elements of the basis matrix **W** were randomly initialized, our method gave band-partitioning constraints on the column vectors (also named as basis vectors) of **W**. Inspired by the cochlea filters, the basis vectors were initialized to be non-negative random values only within a certain frequency band (called useful band for convenience) and zeros in other bands. The widths of the useful bands were almost uniform, and between adjacent basis vectors useful bands were overlapping. Taking the 8th category of CWs for example, they were initialized as follows: for the 1st basis vector, the useful band covers between the 1st~6th samples; for the second one, the useful band covers between the 4th~9th samples, i.e. stepping forward 3 samples

and overlapping 3 samples with the former, and so do the rest vectors. After initialization, iterative algorithm for NMF, i.e. the multiplicative update rules (2) and (3) described in section 2, would be taken until the Euclidean distance between **V** and **WH** converged. Taking the 9th category of CWs for example, the basis vectors trained with no initialization constraints and with band-partitioning constraints are shown in figure 1 and figure 2, and each of them contains 16 sub-figures depicting 16 column vectors of **W** in turn.

When decomposing new CWs, the CWs should be firstly classified into one of the 9 categories, and then take the iterative algorithm (2) to update its coding matrix **H** for 10 times, which is the maximum iteration number required to converge. The histogram in figure 3 shows that the maximum iteration number 10 is sufficient for convergence when decomposing new CWs.

Figure3: Maximum iteration number required for convergence

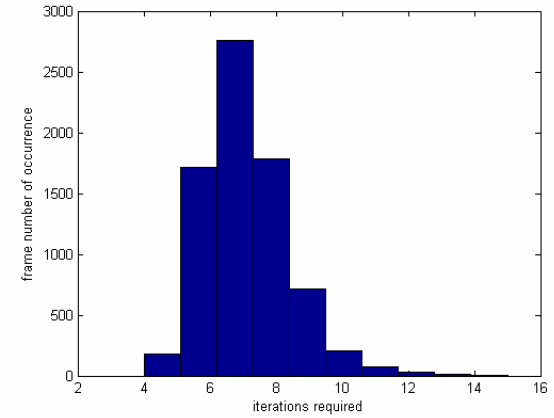
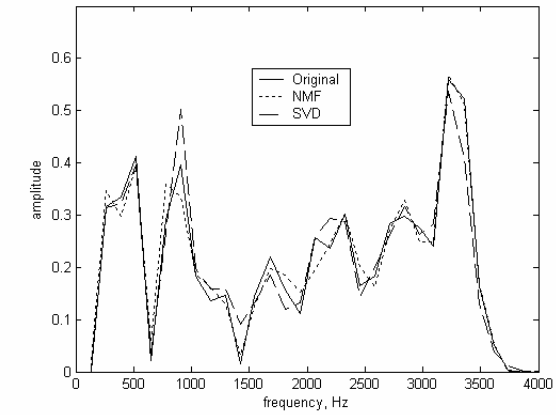


Figure4: One sub-frame CW reconstructions by NMF-based and SVD-based methods

———Original CW spectrum

……Reconstruction made by proposed NMF-based method

- - - - -Reconstruction made by two ranks of singular values



The test speech, including 100000 frames, 50000 of which were female speech and the rest were male speech, were decomposed by SVD and NMF (with two kinds of W initialization), and then the CWs were reconstructed. Taking the square Euclidean distance between the original CWs and the reconstructed CWs as error criterion, the mean errors were as follows: NMF with no initialization constraints was 2.1513; NMF with band-partitioning initialization constraints was 1.8596, which was better than the standard NMF; that was 1.8292 by 2 ranks of singular values' reconstruction, and 2.4853 by 1 rank. Figure 4 shows one sub-frame CW reconstructions by 2 ranks of singular values and by proposed initialization constrained NMF. They obtained almost the same reconstruction accuracy.

The computational complexity of proposed NMF-based decomposition method is $O(mn)$, while the SVD-based method is $O(mn^2)$, where m and n are row and column dimensions of the CWs. So the proposed method is of less complexity than the SVD-based by removing CWs alignment procedures.

5. CWs magnitude spectra quantization

Our NMF-WI coder diagram is shown in figure 5. Differing from earlier WI coders, our NMF-based WI coder has removed the CWs alignment procedures, which has greatly reduced the computational complexity of WI coder [10]. Besides, the high dimensional CWs amplitude spectra can be expressed by low dimensional coding matrix **H** after non-negative matrix factorization, and this will facilitate the CWs magnitude spectra quantization at low bit rate.

Figure 5: Block diagram of the NMF-WI Coder

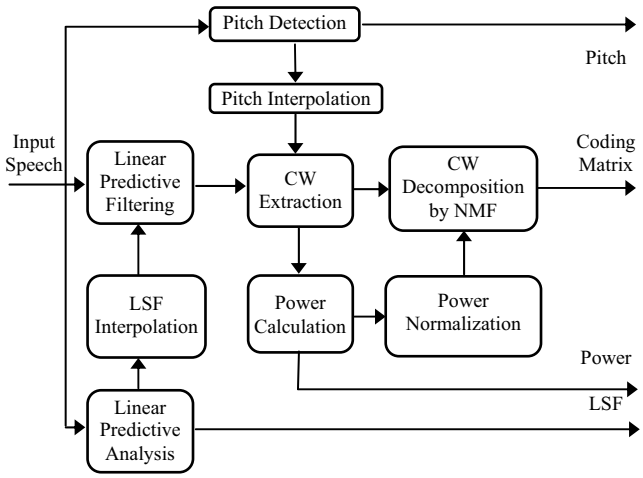
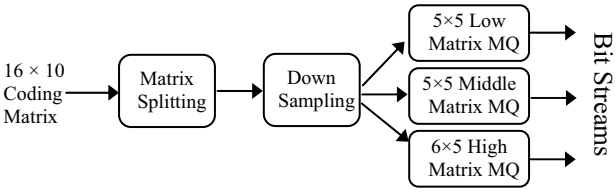


Figure 6: Block diagram of the coding matrix quantization



The basis vectors of **W** trained by NMF with proposed band-partitioning initialization constraints are pulse-like harmonics from low frequency to high frequency. Each column of **V** can be described as weighted of these basis vectors, and the weights are the corresponding column of **H.** So it is straightforward that the low dimensional elements in **H** are the weights for low frequency "pulse-like harmonics" (or called as spectra "parts"), and high dimensional elements in **H** are the weights for high frequency spectra "parts". According to this character of coding matrix, we can split the

H matrix into 3 sub-matrices and flexibly allocate the bits for each sub-matrix. In terms of 16×10 **H** matrix, the first 5×10 sub-matrix, which is used to weight for low frequency basis vectors, is called as low matrix, while the second 5×10 sub-matrix is called as middle matrix, and the last 6×10 sub-matrix is called high matrix. According to perceptual characteristic of human audio system, the perception of hearing is more sensitive to the low frequencies [11]. So when quantizing the coding matrix, we can allocate more bits to low matrix than middle matrix or high matrix.

Before quantizing these sub-matrices, we first down sample their columns into 5-dimension, and in decoder we can restore the columns discarded in coder by linear interpolation. After the down sampling, we use the matrix quantization (MQ) [12, 13] scheme to quantize these sub-matrices. The "coding matrix" quantization scheme is given in figure 5.

And other parameters in our NMF-WI coder, i.e. LSF coefficients, pitch period and CWs power, were all quantized in the same way as [7,14] does. Table 1 shows the bit allocation for our 2kb/s NMF-WI coder.

Table 1. Bit allocation for 2kb/s NMF-WI coder

Parameter	Update rate (Hz)	Bit allocation
LSF coefficients	40	20
Pitch period	40	7
CWs power	40	7
Low matrix	40	7
Middle matrix	40	5
High matrix	40	4

6. Performance evaluation

In order to evaluate the performance of the proposed NMF-based WI coding scheme, the subjective A/B listening test comparing the NMF-based 2kb/s WI coder with the SVD-based 2.4kb/s WI coder [7,14] was finished. The test data sampled by 8 kHz include 16 speech sentences in Chinese, spoken by female and male speakers. The listening test results listed in table 2, which indicate that the speech quality of 2kb/s NMF-WI is close to that of 2.4kb/s SVD-WI, whose speech quality is better than 2.4kb/s MELP [7,14]. And the main disadvantage of our 2kb/s NMF-WI coder is slightly strong correlation between speech frames, which is due to the CWs phase spectra being restored by fixed phase.

Table 2. Results of subjective A/B test

	2kb/s NMF-WI	2.4kb/s SVD-WI	No Preference
Female	18.75%	15.625%	65.625
Male	9.375%	28.125%	62.5%
Total	14.0625%	21.875%	64.0625%

7. Conclusion and future work

The proposed 2kb/s NMF-based WI coder used non-negative matrix factorization techniques to decompose CWs in WI coder, and this decomposition method has learnt the "parts-based" representations of the CWs amplitude spectra. The main advantages of our 2kb/s NMF-WI coder are as follows: (1) NMF-WI doesn't need additional delay; (2) NMF-WI doesn't need CWs alignment procedures, so about 25.968 MOPS was reduced in WI coder [10]; (3) NMF-WI has much lower computational complexity in CWs decomposition than the SVD-based decomposition scheme, but can obtain almost the same CWs reconstruction accuracy as the SVD-based method.

In our NMF-based CWs decomposition framework, the original high dimensional CWs can be expressed by low dimensional coding matrix **H**, and experiments show that this has indeed facilitated the CWs magnitude spectra quantization in low bit rate WI coder. Experiments, however, also show that if the CWs phase spectra were removed in our NMF-WI coder and restored by fixed phase and random phase spectra in decoder, the speech synthesized will have a little of buzz due to slightly strong correlation between speech frames. So how to effectively quantize the CWs phase spectra in our NMF-based WI coding scheme will be an interesting research topic in future work.

8. References

[1] W. B. Kleijn and J. Haagen, "Waveform Interpolation for Coding and Synthesis". Speech coding and Synthesis. Elsevier Science, 175-207,1995.

[2] W. B. Kleijn and J. Haagen, "Transformation and decomposition of the speech signal for coding", IEEE signal processing letters, 1994, 1(9), pp.136–139, 1994.

[3] N. R. Chong, I. S. Burnett and J. F. Chicharo, "A new waveform interpolation coding scheme based on pitch synchronous wavelet transform decomposition", IEEE Transactions on Speech and Audio Processing, Vol.8, Issue 3, pp. 345–348, May 2000.

[4] J. Lukasiak and I. S. Burnett, "Scalable decomposition of speech waveforms", 2002 IEEE Speech Coding Workshop Proceedings, Tsukuba City, Ibaraki, Japan, pp.135–137, 2002.

[5] Gui-ping Wang and Chang-Chun Bao, "High quality low complexity decomposition for the characteristic waveform of speech signal", ISCSLP, pp.145–148, 2004.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", Nature, Vol. 401, pp.788–791, 1999.

[7] Gui-ping Wang, Chang-chun Bao and Peng Zhang, "Low bit rates Waveform Interpolation speech coding based on singular value decomposition", ACTA ELECTRONICA SINICA, Vol.34, No.1, pp.135-140, 2006.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", In Advance in Neural Information Processing 13 (*Proc. NIPS 2000*), MIT Press, 2001.

[9] L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[10] Qi Feng-yan, Bao Chang-chun, "Fast algorithms on Characteristic Waveform representation and alignment in Waveform Interpolation speech coding", Journal of Beijing University of Technology, Vol.32, No.6, pp.514-519,2006.

[11] Chang-chun Bao, "Fundamentals of Digital Speech Coding at Low Bit Rates", Publishing House of Beijing University of Technology, Beijing, 2001.

[12] C. Tsao and R. M. Gray, "Matrix quantizer design for LPC speech using the generalized Lloyd algorithm", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-33, pp. 537–545, Jun. 1985.

[13] Jani Nurminen, Ari Heikkinen and Jukka Saarinen. "A novel quantization scheme for the noise-like component in waveform interpolation speech coding". IEEE ICASSP, Vol. 1, pp.649-652, Orlando, Florida, USA, 2002.

[14] Gui-ping Wang, "Research on low bit rates waveform interpolation speech coding based on singular value decomposition", Master thesis by Gui-ping Wang, Beijing University of technology, Beijing, 2005.