Recovery of Lost Speech Segments Using Incremental Subspace Learning

Jianjun Huang, Xiongwei Zhang, and Yafei Zhang

An incremental subspace learning scheme to recover lost speech segments online is presented. Our contributions in this work are twofold. First, the recovery problem is transformed into an interpolation problem of the time-varying gains via nonnegative matrix factorization. Second, incremental nonnegative matrix factorization is employed to allow online processing and track the evolution of speech statistics. The effectiveness of the proposed scheme is confirmed by the experiment results.

Keywords: Packet loss concealment (PLC), nonnegative matrix factorization (NMF), incremental subspace learning.

I. Introduction

The basic idea behind recovery of missing speech segments is to exploit redundant information embedded in neighboring segments and estimate the speech content of lost segments from their neighbors. The simplest predictive method is to replicate the previous speech segment instead of the missing segments. This method has low computational complexity and performs better than muting the signal, but it is not sufficient for high quality applications. An even better method is waveform substitution. This method selects a portion of the previous speech that can best approximate the lost samples and uses these samples to fill the lost segments. The International Telecommunication Union (ITU) has standardized a waveform substitution method for reconstructing the lost speech segments (G.711 Appendix I) [1]. In [2], Liang and others extend the work to estimate the lost segments through waveform similarity overlap-add (WSOLA). Techniques based on waveform substitution have proven to be very popular due to its simplicity. However, continual loss of segments leads to metallic-sounding artifacts [3].

Although differing in terms of speech model type and implementation details, conventional recovery algorithms are largely based on variations of signal repetition or parameter interpolation and extrapolation. Consequently, by focusing on very local signal statistics, conventional recovery algorithms always miss the larger context of speech statistics trends. This results in unwanted auditory artifacts in the recovered speech. In [4], Rødbro and others attempt to address the limitations of traditional recovery algorithms by using a hidden Markov model (HMM) to track the evolution of speech signal parameters. However, as described in [4], an HMM is an imperfect model of speech signals. Although differing from conventional algorithms, an HMM-based algorithm is still a "local" algorithm that is determined by the number of model states.

In this letter, we present an incremental nonnegative subspace learning scheme to recover missing speech segments using incremental subspace learning (ISL) [5]. To the best of our knowledge, ISL has not been previously used in the recovery of missing speech segments. In the proposed scheme, the recovery problem is transformed into a linear interpolation of projective parameters. Since the nonnegative subspace is updated when a new speech segment is available, the proposed algorithm is capable of tracking the evolution of speech statistics. In fact, we show that the recovery of the missing segments in the nonnegative subspace provides superior quality of speech to that provided by conventional algorithms. The proposed approach has a number of applications, such as restoration of archived speech recordings, estimation of lost speech packets due to dropouts over IP networks, and improvement of recognition accuracy in distributed speech

Manuscript received Sept. 22, 2011; revised Mar. 15, 2012, accepted Mar. 23, 2012.

Jianjun Huang (phone: +86 135 1250 1916, hjj954@gmail.com), Xiongwei Zhang (xwzhang@163.com), and Yafei Zhang (yfz55@gmail.com) are with the Institute of Command Automation, PLA University of Science and Technology, Nanjing, China.

http://dx.doi.org/10.4218/etrij.12.0211.0408

recognition systems over a communication channel.

II. Four Steps of Proposed Scheme

1. Magnitude Spectra Representation of Speech

The input speech signal $y(n)$ is transformed into the frequency domain by applying a window $h(n)$ to a frame of L samples of $y(n)$ and by computing the short-time Fourier transform (STFT) of size L on the windowed data. The window is shifted by R samples before the next STFT computation. The STFT analysis results in a set of frequency domain signals that can be written as

$$Y(k,t)=\sum_{n=0}^{L-1} y(tR+n)h(n)e^{-j2\pi kn/L}, \quad 0\le k\le L-1, \quad (1)$$

where k=0,…, L–1 is the discrete frequency index and t=0,…, T–1 is the frame index. Here, K is the number of frequency bins and T is the number of frames. In our implementation, we use a sampling rate of f_s=8,000 Hz and L=2R=256. Speech segment length is equal to the frame size. Phases are discarded by taking the absolute values of the STFT spectra, resulting in the magnitude spectrogram $M(k, t)=|Y(k, t)|$. To facilitate our notation, we discard the frequency index k from $M(k, t)$ and use $\mathbf{m}_t$ to denote the magnitude spectra vector of the current speech segment. Thus, the magnitude spectrogram of speech can be reformulated as

$$\mathbf{M}=[\mathbf{m}_0,\mathbf{m}_1,...,\mathbf{m}_{T-1}]. \quad (2)$$

2. Incremental Nonnegative Subspace Learning

By applying r rank nonnegative matrix factorization (NMF) to $\mathbf{M}$, the magnitude spectrogram matrix can be decomposed into the product of nonnegative matrices $\mathbf{A}$ and $\mathbf{X}$, which is given as

$$\mathbf{M}=\mathbf{AX}. \quad (3)$$

In (3), the rows of $\mathbf{X}$ denote the time-varying gains, while the column vectors of $\mathbf{A}$ span the nonnegative subspace. Because of its batch and static processing nature, conventional NMF is incapable of online processing.

To allow online implementation and track the evolution of speech statistics, the nonnegative subspace is learned and updated using the new proposed ISL technique [5]. The ISL is performed via incremental NMF (INMF). The aim of INMF is to update $\mathbf{A}$ and $\mathbf{X}$ by adding effects of the new arrival magnitude spectra $\mathbf{m}_t$. This process can be written as

$$\mathbf{m}_t=\mathbf{Ax}_t. \quad (4)$$

The time-varying gain $\mathbf{x}_t$ in time index t is updated by projecting magnitude spectra vector $\mathbf{m}_t$ onto nonnegative subspace $\mathbf{A}$ using (4). Equation (4) is also used to update the nonnegative subspace every time a new magnitude spectra vector $\mathbf{m}_t$ is available. Due to its dynamic updating nature, the nonnegative subspace is capable of tracking the trend of speech statistics.

The multiplicative update rules [5] for $\mathbf{x}_t$ and $\mathbf{A}$ are given as

$$\mathbf{x}_t \leftarrow \mathbf{x}_t \otimes (\mathbf{A}^T\mathbf{m}_t \odot \mathbf{A}^T\mathbf{Ax}_t), \quad (5)$$

$$\mathbf{A} \leftarrow \mathbf{A} \otimes (\beta\mathbf{MX}^T+\alpha\mathbf{m}_t\mathbf{x}_t^T) \odot (\mathbf{A}(\beta\mathbf{XX}^T+\alpha\mathbf{x}_t\mathbf{x}_t^T)), \quad (6)$$

where $\otimes$ denotes element-wise multiplication, $\odot$ represents element-wise division, and α and β are the weighting coefficients, which are chosen as 0.2 and 0.8, respectively, as described in [5].

3. Linear Interpolations of Time-Varying Gains

Transformation of the lost speech segments recovery problem into linear interpolation of time-varying gains is based on the observation that the time-varying gains always change slowly over time. Moreover, the smoothness of time-varying gains can be enhanced by incorporating a temporal smoothness constraint [6] into the conventional NMF framework.

Assuming only one frame of speech is available before and after the lost frames, the lost time-varying gains can be estimated by using linear interpolation. We define the general linear interpolating function as

$$\hat{\mathbf{x}}_t=\frac{\mathbf{x}_b(T_{\text{lost}}-t+1)+\mathbf{x}_a t}{T_{\text{lost}}+1}, \quad 1\le t\le T_{\text{lost}}, \quad (7)$$

where $\mathbf{x}_b$ is the first time-varying gain vector before the lost segment, $\mathbf{x}_a$ is the final time-varying gain vector after the lost segment, $\hat{\mathbf{x}}_t$ is the estimated time-varying gain vector of the lost segment, and T_{lost} is the consecutive lost frames number. This is illustrated in Fig. 1. Note that the subscript t, as seen in (7), represents the frame index. In this work, we choose linear

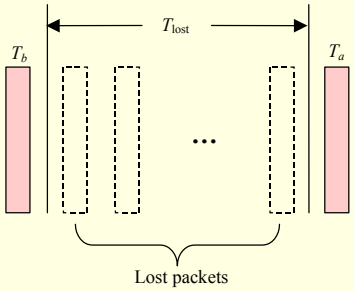

Fig. 1. Illustration of T_{lost} missing packets of speech signal.

interpolation just because it is computationally inexpensive and provides acceptable speech quality, in most cases.

4. Waveform Reconstructions via Spectrogram Inversion

After the time-varying gain vector of the lost segment is estimated, it is mapped back into the magnitude spectra representation, $\hat{\mathbf{m}}_t = \mathbf{A}\hat{\mathbf{x}}_t$. The time-domain speech signal of the lost segment is then reconstructed using the real-time spectrogram inversion algorithm proposed in [7]. In many signal processing applications, the STFT phase is lost or not available. Thus, it is desirable to reconstruct the time-domain signal for an STFT magnitude. The real-time spectrogram inversion algorithm presented in [7] is a method for estimating the time-domain signal from the STFT magnitude spectra **M** without the phase information. The proposed spectrogram inversion algorithm shows its superiority over current methods, and it is used to restore the waveform of the lost frames.

The overall procedure of the proposed reconstruction scheme is shown in Fig. 2.

The time complexity of the proposed ISL scheme is mainly introduced by the computation of the INMF algorithm. The time complexity of the INMF algorithm is approximately $O(L\times r)$ for each new frame, where L is the frequency bin number and r is the rank of the nonnegative subspace. Moreover, the time complexity of the NMF algorithm is $O(L\times r\times T)$ [8] at each iteration, where T is the overall frame number. Thus, the computational cost of the proposed scheme for large scale datasets is reduced compared to the cost of the NMF algorithm. In addition, as the incremental learning process performs in an online manner, the proposed scheme is suitable for online implementation.

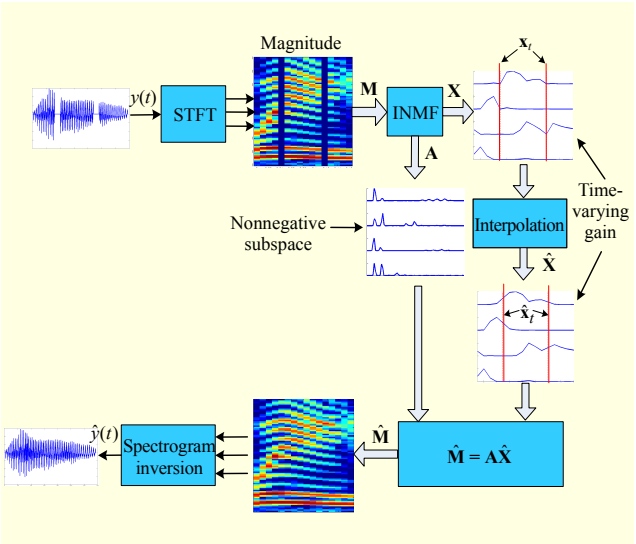

Fig. 2. Overall procedure of proposed reconstruction scheme.

III. Experiments and Results

The Gilbert model [3] is used to introduce frame loss in speech signals. After introducing the gaps in speech signals, each signal is reconstructed using the proposed algorithm. The performance of the proposed recovery algorithm is compared with the WSOLA method [2] and the standard recovery algorithm in ITU-T G.711 [1]. Silence substitution (SS), also known as zero stuffing, is also evaluated.

As performance measures, the perceptual evaluation of speech quality (PESQ) score [9] and log-spectral distance (LSD) are adopted. The results are calculated and averaged for a test set of approximately 100 sentences randomly selected from the TIMIT database. A random frame loss (generated by the Gilbert model) test is performed at loss rates ranging from 5% to 40%. According to the Gilbert model, the average consecutive frames lost numbers (T_{lost} equals lost number) corresponding to loss rates of 5%, 10%, 20%, 30%, and 40% are 1.05, 1.11, 1.25, 1.43, and 1.67, respectively.

Figure 3 shows the averaged PESQ and LSD evaluation results. It is observed that the proposed algorithm always outperforms other algorithms for all percentages of packet losses. As depicted in Fig. 3(a), the PESQ score of the proposed algorithm at a 20% frame loss rate is 0.15 higher than that of the WSOLA method and 0.39 higher than that of the

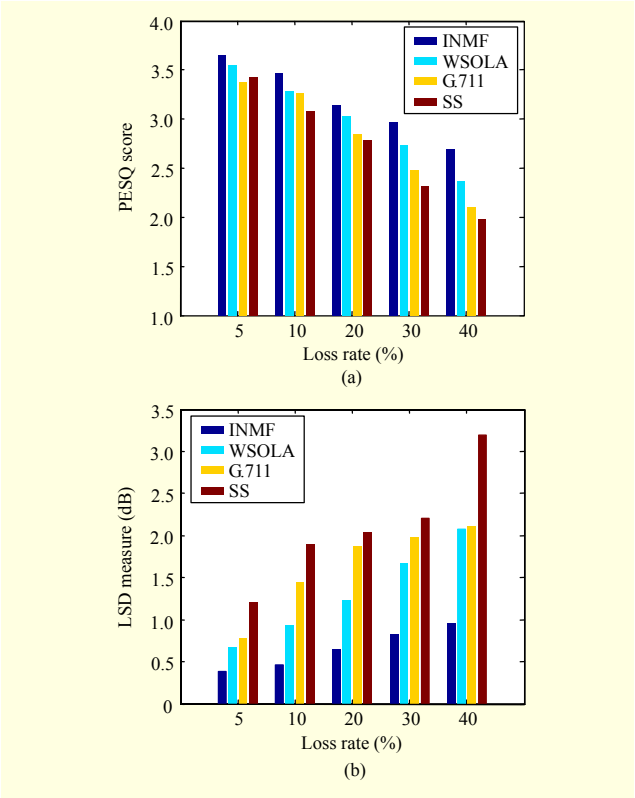

Fig. 3. Performance comparison of different recovery algorithms.

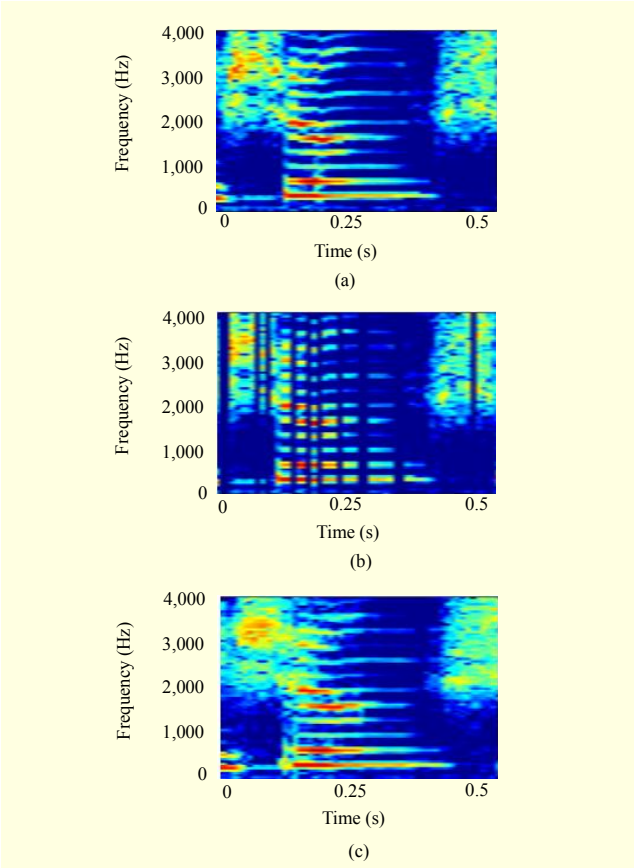

Fig. 4. Spectrogram of sample signal with introduction of 30% frame loss and the restored versions: (a) original speech, (b) distorted speech, and (c) reconstructed speech using INMF.

G.711 method. As the loss rate becomes higher, the PESQ score improvement by the proposed algorithm becomes greater. The advantages of the proposed algorithm are also indicated by the LSD measure in Fig. 3(b). The second most efficient recovery method is the WSOLA method, followed by G.711 method. The SS method shows the worst performance. The performance of both the WSOLA method and the G.711 method deteriorates very quickly as the loss rate increases. This is because these conventional methods are largely based on variations of signal repetition or parameter interpolation. Consequently, these methods miss the larger context of speech statistics trends and result in metallic-sounding speech, as is validated by subjective listening tests. Conversely, the nonnegative subspace is able to track the speech signal's statistics evolution. Thus, recovery of missing speech segments in the nonnegative subspace can utilize more redundant information to reconstruct the lost speech samples than can conventional recovery algorithms. This reason contributes to the robustness of the proposed algorithm.

To investigate in detail, we draw the spectrograms of a sample signal, its silence substitution, and its recovered version, as seen in Fig. 4. From Fig. 4, we can see that the proposed scheme successfully restores the lost segments.

IV. Conclusion

An effective and online recovery algorithm was proposed, and its power verified by experiments. Compared to conventional recovery algorithms, the proposed algorithm produces improved perceptual quality of speech. The experiment results encourage the use of the proposed algorithm in many practical applications. The intention for future work is to employ a more sophisticated interpolation method to estimate the time-varying gains.

References

[1] Appendix I: A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711, *ITU-T Recommend G.711*, Sept. 1999.

[2] Y.J. Liang, N. Färber, and B. Girod, "Adaptive Playout Scheduling and Loss Concealment for Voice Communication over IP Networks," *IEEE Trans. Multimedia*, vol. 5, no. 2, June 2003, pp. 532-543.

[3] E. Zavarehei and S. Vaseghi, "Interpolation of Lost Speech Segments Using LP-HNM Model with Codebook Post-Processing," *IEEE Trans. Multimedia*, vol. 10, no. 3, Apr. 2008, pp. 493-502.

[4] C.A. Rødbro et al., "Hidden Markov Model-Based Packet Loss Concealment for Voice over IP," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, Sept. 2006, pp. 1609-1623.

[5] S.S. Bucak and B. Gunsel, "Incremental Subspace Learning via Non-negative Matrix Factorization," *Pattern Recognition*, vol. 42, no. 5, May 2009, pp. 788-797.

[6] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, Mar. 2007, pp. 1066-1074.

[7] X. Zhu, G.T. Beauregard, and L.L. Wyse, "Real-Time Signal Estimation from Modified Short-Time Fourier Transform Magnitude Spectra," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, July 2007, pp. 1645-1653.

[8] G. Zhou et al., "Online Blind Source Separation Using Incremental Nonnegative Matrix Factorization with Volume Constraint," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, Apr. 2011, pp. 550-560.

[9] Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, *ITU-T Recommendation P.862*, 2001.