Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 87 (2007) 2431–2445
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

Enhancement of Chinese speech based on nonlinear dynamics

Junfeng Sun, Nengheng Zheng[1], Xinlong Wang*

Institute of Acoustics, Nanjing University, Nanjing 210093, China

Received 21 September 2005; received in revised form 25 March 2007; accepted 30 March 2007
Available online 14 April 2007

Abstract

Based on recently observed nonlinear dynamic features of human speech, the local projection (LP) method, originally developed for noisy chaotic time series, is generalized and adapted to the enhancement of Chinese speech. The analysis of minimum embedding dimensions estimated by the false nearest neighbor algorithm shows that all the basic phonemes and syllables in Chinese can be faithfully embedded in some low-dimensional phase space. Over-embedding is applied to reconstruct the dynamics of continuous speech in some extended phase space of higher dimension, thus solving the problem of nonstationarity in continuous speech. A generalization of the LP method, named the *local subspace method*, is presented for speech enhancement in the phase space. It is demonstrated that, the local subspace method is essentially an extension of the well-known linear subspace technique in the local phase space, and the LP method is the least square case of this generalization. Noise reduction is then carried out in the local phase space. Results show that the LP method, with 2 or 3 iterations, achieves better performances than the local subspace method. For both isolated and continuous speech with additive white noise, experiments show the superiority of the LP method over two other popular algorithms.

PACS: 43.72.Ew; 43.60.Ac; 43.25.Ts

Keywords: Speech enhancement; Local projection; Nonlinear time series analysis

1. Introduction

Speech enhancement, which aims at eliminating noise components from contaminated speech signal, is one of the most important technologies for speech processing. In the past several decades, a variety of approaches for speech enhancement have been proposed, mostly based on the classical linear acoustical model of speech [1]. Two popular directions of these studies are noise suppression in the frequency domain and noise elimination in the signal subspace. The frequency domain methods take advantage of the perceptual function of the short-time spectrum of speech signal. Typical algorithms include the spectral subtraction [1,2], Wiener filtering [1,3], and that proposed by Ephraim and Malah (E–M) [4], which has been commonly recognized to be the best one of this class. The subspace methods assume the separability of speech and noise in some properly reconstructed space, i.e., the existence of two orthogonal subspaces: (1) the noise subspace containing components from the noise process only, and (2) the signal

*Corresponding author. Tel.: +86 25 83593663.
E-mail address: xlwang@nju.edu.cn (X. Wang).
[1]Present address: Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

0165-1684/$ - see front matter
doi:10.1016/j.sigpro.2007.03.020

subspace containing the dominant speech signal, plus a certain amount of noise as well. Enhancement is performed by removing the components in the noise subspace and estimating the clean signal from the remaining components in the signal subspace. The estimation of signal subspace can be achieved using either the singular value decomposition (SVD) of the data matrix [5,6], or the eigenvalue decomposition (EVD) of the data covariance matrix [7–13]. The more popular framework based on EVD was first established by Ephraim and Van Trees (E–V) [7], and then extended to an adaptive one [8] and to colored noise [9–11]. The hearing property and perceptual mechanism of human beings were also introduced into this framework for further improved performance [12,13]. Although these algorithms perform comparatively well, they all, especially those approaches in the frequency domain, suffer from different levels of annoying residual noise, e.g., the "musical noise".

The conventional linear model of speech overlooks the underlying nonlinearity in various stages of speech production, e.g., the nonlinear vocal cords vibration, the nonlinear modulation and resonance when sound waves propagate through vocal tract, the strong turbulent motions in producing fricatives, etc. The nonlinear characteristics of the voiced sounds were recently revealed by investigating the speech production mechanism [14–16]. On the other hand, nonlinear analysis of speech signal in different languages [17–19], including Chinese [20], disclosed the chaos-like dynamic features in most phonemes, especially the voiced ones, albeit the continuous speech may be highly nondeterministic and nonstationary. Surrogate tests on the simple vowels of US English showed that the linear prediction analysis widely applied in current speech technologies does not fully exploit the dynamic characteristics of speech signal [21]. These results call for nonlinear or linear/nonlinear hybrid models to characterize the nonlinear phenomena in speech.

In recent years, various techniques based on nonlinear dynamics have been developed for speech analysis and processing, for example, the synthesis technique for voiced sounds [22], the short-term prediction in speech coding [23,24], the nonlinear analysis of speech [25], the classification of isolated phonemes [26], and the noise reduction for noisy speech [20,27,28]. Hegger et al. [27] first attempted to adapt the local projection (LP) method, originally developed for chaotic signal [29,30], to enhance the sustained vowel /a/ with additive white noise. Their result showed the superiority of the LP method over E–M's algorithm [4] in the sense of signal-to-noise ratio (SNR). Zheng et al. revealed the nonlinear dynamic characteristics of Chinese speech by estimating the dynamic features (e.g., correlation dimension, Lyapunov exponents) of some selected vowels and voiced consonants, and an initiative attempt was made to adapt the LP method to enhance some typical isolated phonemes of Chinese speech [20]. Motivated by the encouraging works as mentioned, we attempt to make a comprehensive survey on speech enhancement based on the nonlinear dynamics of speech in this paper. Specifically, we investigate on these issues: (1) the phase-space embedding of Chinese phonemes and syllables, through which the nonlinear dynamics can be clearly reconstructed; (2) the generalization of the LP method and its relationship with the linear subspace technique; and (3) the applicability of the LP method to the enhancement of Chinese speech, both for isolated and continuous speech. In this study, we assume that the speech is contaminated by additive white Gaussian noise, while leaving the general approach in real noise environment as our next-step work.

The rest of this paper is organized as follows. Section 2 briefly reviews the embedding theorem and demonstrates the embedding of Chinese phonemes and syllables in the reconstructed phase space. It is shown that all Chinese phonemes and syllables can be well embedded in some low-dimensional phase space, and the nonstationarity in continuous speech can be tackled by the so-called over-embedding technique. With these preparations, noise reduction can be performed in the reconstructed phase space, being an alternative to the linear subspace noise reduction. In Section 3, the LP method is introduced and generalized to the local subspace method. The relationship between the LP, the local subspace, and the general linear subspace methods is also discussed. Section 4 compares the speech enhancement results on isolated Chinese speech by several approaches, i.e., the LP, the local subspace, E–M's, and E–V's methods. It is shown that the LP method outperforms others in the sense of SNR. Section 5 demonstrates the superiority of the LP method over E–M's and E–V's algorithms in the continuous speech enhancement via both objective and subjective evaluation. Finally, we draw conclusions of this study in Section 6.