AUDITORY CODING BASED SPEECH ENHANCEMENT

Yao Ren, Michael T. Johnson

Speech and Signal Processing Lab, Marquette University, Milwaukee, WI 53201 USA
{*yao.ren, mike.johnson*}*@marquette.edu*

ABSTRACT

This paper demonstrates a speech enhancement system based on an efficient auditory coding approach, coding of time-relative structure using spikes. The spike coding method can more compactly represent the non-stationary characteristics of speech signals than the Fourier transform or wavelet transform. Enhancement is accomplished through the use of MMSE thresholding on the spike code. Experimental results show that compared with the spectral domain logSTSA filter, both the subjective spectrogram evaluation and objective SSNR improvement for the proposed approach is better in suppressing noise in high noise situations, with fewer musical artifacts.

Index Terms— Speech enhancement, auditory coding, wavelet transforms, minimum mean square error methods

1. INTRODUCTION

Modern speech enhancement methods originated with the development of spectral subtraction [1] in the late 1970s. Rapid progress in the early 1980s saw the advent of two other enhancement methods: iterative Wiener filtering [2] and logSTSA filtering [3]. Most of the current speech enhancement methods are built and extended on these three baseline methods, which are based on the same mathematical tool, the short time Fourier transform (STFT), with the waveform divided into short frames during which the signal is assumed to be stationary. Non-stationary acoustic signals, however, are shift-sensitive to this block-based signal processing tool, due to their non-stationary transient structure [4].

As an alternative analysis tool of STFT, Wavelet Transform (WT) has the advantage of using an implicitly variable window size for different frequency components. This often results in better handling of non-stationary data like speech. The application of wavelets for signal enhancement is attracting more attention [5-7]. Like the Fourier transform, WT has both continuous WT (CWT) and discrete WT (DWT) implementations. The DWT method is based on decomposition by a quadrature mirror filter, and is sensitive to the selection and design of this filter. Additionally, the DWT is dyadic by nature and so its frequency scaling does not line up well with the perceptual frequency scaling desired for human speech. The CWT does not have this limitation, and can accurately represent speech structure through a good choice of Mother wavelet. However, implementation of the CWT is quite inefficient, requiring numerical integration techniques, and is often a highly redundant signal representation when done with fine frequency scaling.

A non-block based, time-relative representation method for auditory coding has been proposed in [8]. In this method, the speech signal is decomposed into sparse, shiftable acoustic spikes, represented by the kernel functions with a corresponding amplitude and temporal position, under the assumption that acoustic signal is encoded by spikes at the auditory nerve in the inner ear. This method has been shown to better characterize non-stationary structure in speech signal than Fourier transform. Motivated by this auditory coding system, the work presented here uses this coding technique, instead of the traditional STFT and WT. An MMSE thresholding technique is used to reduce noise and enhance speech, with the idea that the better representational capability of this coding method could lead to better enhancement results.

In section 2, we give a review of the spike auditory coding method. Section 3 presents the enhancement method based on this auditory coding system. Results are present and discussed in section 4, with a conclusion in section 5.

2. SPIKE AUDITORY CODING

2.1. Mathematical Model

Different from the block-based representation (Fourier Transform) and convolutional representation (CWT), spike coding is a sparse shiftable kernel representation, which is motivated by the assumption that speech signal is coded into spikes in the inner auditory nerves. In this model, a set of arbitrarily and independently positioned kernel functions $\phi_1,...,\phi_M$ are applied to code the signal $x(t)$, which is represented by the following mathematical form

$$x(t)=\sum_{m=1}^{M}\sum_{i=1}^{n_m} s_i^m \phi_m(t-\tau_i^m)+\varepsilon(t) \qquad (1)$$

where τ_i^m is the temporal position of the i^{th} instance of kernel function ϕ_m, s_i^m is its corresponding coefficient, n_m is the total number of kernel functions and $\varepsilon(t)$ is the coding error.

Based on this model, the speech signal is decomposed with respect to these kernel functions and coded as discrete acoustic events, which is called a *spike code*, each of which has an amplitude and temporal position.

2.2. Encoding Algorithm

Three encoding algorithms have been introduced in [8] to compute the optimal values of τ_i^m and s_i^m for a given signal, to minimize the error $\varepsilon(t)$ and maximize coding efficiency. Here we use Matching Pursuit method for spike coding strategy. The idea of Matching Pursuit-based algorithm is to iteratively decompose the signal in terms of the kernel functions so as to best capture the signal structure, by projecting the coding residual signal of each iteration onto the kernel functions. The projection with the largest inner product is subtracted out and its coefficient and time instant are recorded. The signal is decomposed into kernel functions by

$$x(t) = \langle x(t) \cdot \phi_m \rangle \phi_m + R_x(t) \tag{2}$$

where $\langle \cdot \rangle$ indicates inner product and $R_x(t)$ is the residual signal after projecting $x(t)$ in the direction of ϕ_m. Iteratively projecting the signal in the direction to maximize the inner product $\langle x(t) \cdot \phi_m \rangle$ minimizes the power of $R_x(t)$, which can be generally expressed as

$$R_x^n(t) = \langle R_x^n(t) \cdot \phi_m \rangle \phi_m + R_x^{n+1}(t) \tag{3}$$

with the initialization of $R_x^0(t) = x(t)$. The best fitting projection is subtracted out, and its coefficient and time are recorded. Kernel function ϕ_m is selected by

$$\phi_m = \arg\max_m \langle R_x^n(t) \cdot \phi_m \rangle \tag{4}$$

The spike amplitude corresponding to the selected kernel function is calculated by

$$s_m = \langle x(t) \cdot \phi_m \rangle . \tag{5}$$

3. ENHANCEMENT TASK

The enhancement method presented here is based on applying an MMSE thresholding technique to spike coefficients. A block diagram of the overall approach is shown in Fig. 1. Spike coefficients Y_i^m are computed by (4) and (5). The kernel function calculation is discussed in detail in section 3.1 and thresholding method is discussed in section 3.2.

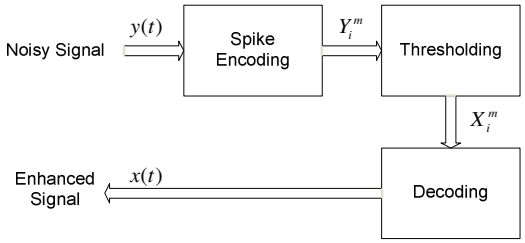

Fig. 1. Block diagram of the enhancement system

3.1. Morlet Wavelet Kernel Functions

Kernel function selection is a key part of spike coding, since the signal is encoded into spikes, each of which is represented by the corresponding kernel function located at a precise temporal position. Instead of using 64 Gammatone functions as in [8], we propose to use a set of scaled Morlet wavelet functions as the kernel functions.

The Morlet wavelet has the advantage of easy selection of its center frequency and quality factor. It was firstly used for speech coding tasks [9] , and has been successfully used for cochlear implants [10]. It has been argued that Morlet wavelet is an optimal speech representation solution [9], and that it is more suited for modeling phonemes.

The real Morlet wavelet is defined as

$$\varphi(t) = \exp\left(-\frac{t^2}{T_0}\right)\cos(\omega_0 t) \tag{6}$$

where we take $F_0 = 15{,}165.4Hz$ [10]. We keep this base frequency F_0, but recalculate time support T_0 to match the net time-frequency product of $T_0\omega_0$ of the standard Morlet wavelet. This time support would be $T_0 = 0.00007421$ [6].

Two sets of kernel functions are designed for this enhancement experiment: 1) 22 Morlet wavelet functions, with logarithmic spaced center frequencies to match cochlear frequency warping curve, following [10], and 2) 64 Morlet wavelet functions, with uniformly spaced frequencies across the frequency range. These pre-determined center frequency are accomplished by the discretization of the scale variable a in the Morlet wavelet function. The calculations of scale factor a_m , center frequency f_m and corresponding kernel function ϕ_m are addressed in (7) and (8). A similar calculation is used for uniformly spaced kernel functions.

$$a_m = (1.1623)^{m+6},\ f_m = F_0 / a_m,\ m = 1,\ldots,22 \tag{7}$$

$$\phi_m = \varphi\left(\frac{t}{a_m}\right) = e^{-\left(\frac{t}{a_m T_0}\right)^2}\cos\left(\frac{2\pi f_0 t}{a_m}\right) \tag{8}$$

Fig. 2 illustrates the comparison of spectrogram and spikegram of a word pronounced as /aa b aa/, using 22 Morlet kernel functions. In the spikegram, the size of each point indicates the amplitude of spike.

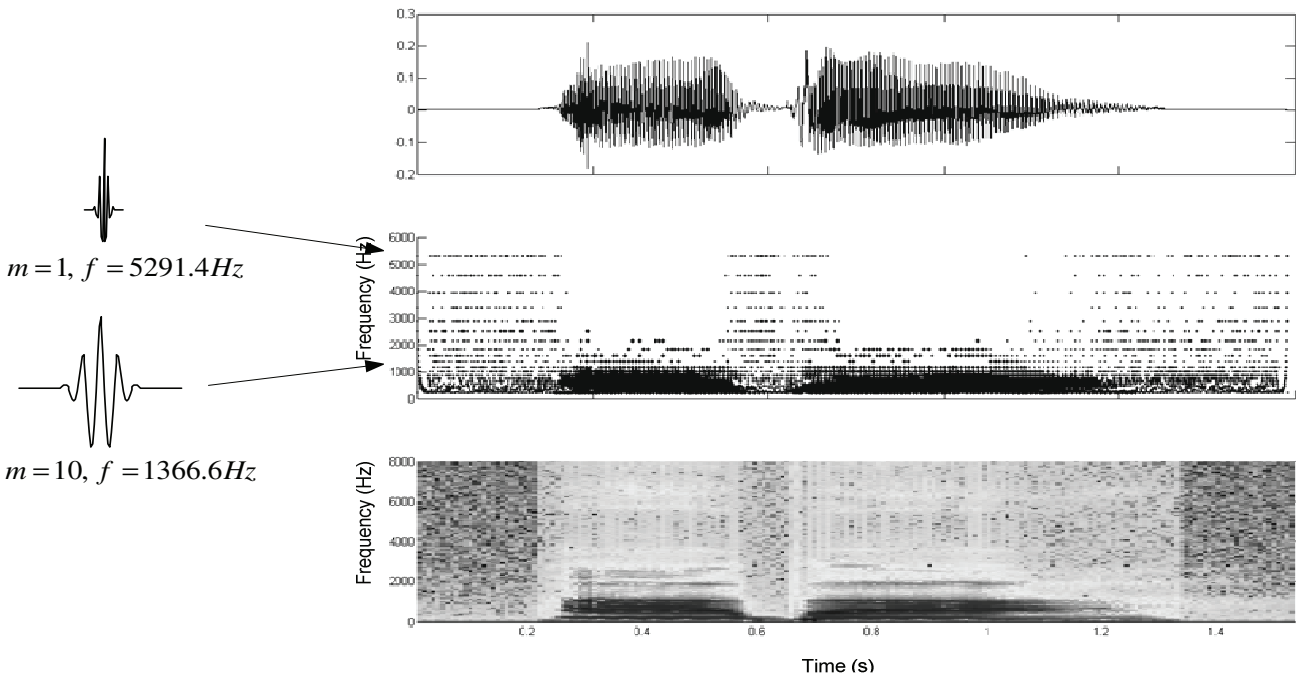

Fig. 2. Three representation of word /aa b aa/: upper, a time domain waveform; middle, spikegram; lower, spectrogram

3.2. Thresholding

Given the encoding structure, an MMSE estimator is applied to threshold the spike coefficients. This estimator is an optimally modified LSA estimator [11] which has been used for wavelet denoising.

Let Y_i^m be the spike coefficient corresponding to the kernel function ϕ_m after encoding processing. An estimate for the clean coefficient, which minimize the mean-square error, results in

$$X_i^m = \frac{\lambda_i^m p_i^m}{\lambda_i^m + (\sigma_i^m)^2} Y_i^m \tag{9}$$

where the signal variance is given by using the decision-directed method of logSTSA filter

$$\lambda_i^m = \alpha \left| X_{i-1}^m \right| + (1-\alpha) \max\left[\left| Y_i^m \right| - \sigma_i^m, 0 \right] \tag{10}$$

p_i^m is a parameter of signal presence uncertainty which is calculated through the equation

$$p_i^m = \left\{ 1 + \frac{1+\xi_i^m}{(q_i^m)^{-1}} \exp\left[-\frac{\upsilon_i^m}{2} \right] \right\}^{-1} \tag{11}$$

where ξ_i^m is the *a priori* SNR,

$$\upsilon_i^m = \frac{1}{1+\xi_i^m} \gamma_i^m, \ \gamma_i^m = \frac{\left(Y_i^m\right)^2}{\lambda_{di}^m} \tag{12}$$

and q_i^m is the a priori probability for signal absence, which is estimated by

$$\hat{q}_i^m = 1 - \begin{cases} \dfrac{\log(\xi_i^m / \xi_{\min})}{\log(\xi_{\max} / \xi_{\min})} & if\ \xi_{\min} \le \xi_i^m \le \xi_{\max} \\ 0 & if\ \xi_i^m \le \xi_{\min} \\ 1 & otherwise \end{cases} \tag{13}$$

4. EXPERIMENT RESULTS

To evaluate the performance of the this method, logSTSA enhancement and the proposed spike coding based enhancement are performed over 10 speech utterances taken from TIMIT database [12]. For logSTSA, a frame size of 32ms with 75% overlap is used. 10 iterations are used for Matching Pursuit method in spike encoding part. White noise is added to each utterance at an Segmental SNR(SSNR) level form -25dB to +10dB. The noise spectrum is estimated by averaging the first 3 frames of each noisy utterance.

Evaluation of the method was done by comparing the objectively measured quality of the enhanced signal through SSNR improvement. Objective evaluation results are shown in Fig. 3. The averaged SSNR improvement from 10 utterances show that the proposed spike coding based enhancement method has significant improvement over the logSTSA method in low SNR situation, but is not as effective in less noisy situations. Two interesting points in these results:

1) 64 kernels do not provide better results than just 22 kernels. For coding system, the more kernel functions, the better the speech quality (also the higher the bit rate), which

is not the case for enhancement task. Too many kernel functions may result in an insufficient number of coefficients for each individual kernel, preventing accurate statistical measures for MMSE thresholding.

2) In low noise conditions, the proposed method does not work as well. In a noisy environment, a better representation of the signal can facilitate the extraction of the signal information out of the noisy signal; however, when the signal is relatively clean, the spike coding together with the thresholding method may over-denoise the spike coefficients and cause some signal distortion.

An example spectrogram in -5dB white noise is shown in Fig. 4. It can be seen that the 22 kernel function thresholding suppresses a significant amount of background noise compared with the logSTSA method. Acoustically, there is also a reduced level of musical artifact. Although the 64 kernel function thresholding reduces more ambient noise, it also suppresses more vocalization information.

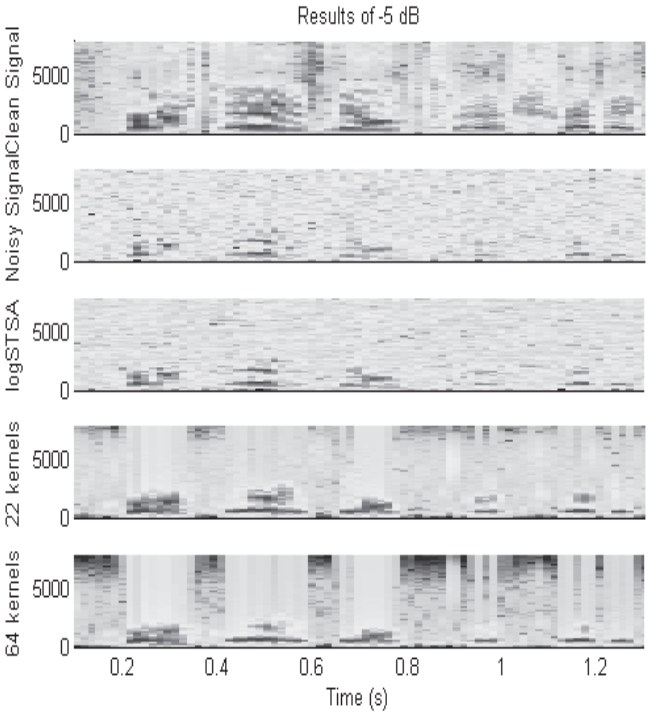

Fig. 4. Spectrogram of enhanced signal

5. CONCLUSION

This paper has introduced a novel spike coding based speech enhancement approach, distinctly different from traditional Fourier transform and wavelet transform based speech enhancement methods, in that the waveform is encoded as a discrete set of acoustic events rather than transformed in entirety. Results indicate that the new approach gives better results than standard logSTSA estimation in high noise.

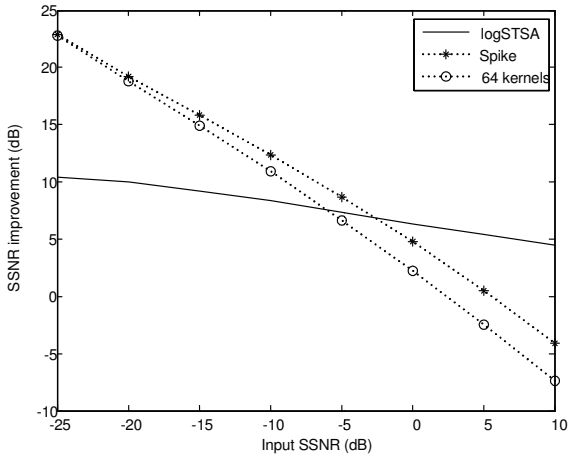

Fig. 3. SSNR evaluation of spike coding based enhancement

6. REFERENCE

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process*, vol. ASSP-27, pp. 113-120, 1979.

[2] J. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 26, pp. 197-210, 1978.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process*, vol. ASSP-32, pp. 1109-1121, 1984.

[4] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356-363, 2002.

[5] M. Jansen, *Noise Reduction by Wavelet Thresholding*: Spinger, New York, 2001.

[6] M. T. Johnson, X. Yuan, and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Commun.*, vol. 49, pp. 123-133, 2007.

[7] I. Cohen, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 80, pp. 2403-2418, 2001.

[8] E. C. Smith and M. S. Lewicki, "Efficient coding of time-relative structure using spikes," *Neural Computers*, vol. 17, pp. 19-45, 2005.

[9] W. Walker and S. Foo, "Optimal Wavelets for Speech Signal Representations," *Journal of Systemics, Cybernetics and Informatics*, vol. 1, No. 4, pp. 44-46, 2003.

[10] J. Yao and Y. T. Zhang, "Bionic wavelet transform: a nre time-frequency method based on an auditory model," *IEEE Trans. Biomed. Eng.*, vol. 48, pp. 856-863, 2001.

[11] I. Cohen and B. Berdugo, "Speech enhancement for non-statoinary noise enviroments," *Speech Processing*, vol. 81, pp. 2403-2418, 2001.

[12] J. Garofolo, L. Lamel, and W. Fisher, "TIMIT Acoustic-Phonetic Continuous Speech Corpus: Linguistic Data Consortium," 1993.