Judgement of unvoiced and voiced pronunciation based on sparse feature with DCT dictionary

Lianzi Wang¹, Nikos Mastorakis², and Xiaodong Zhuang^{1,*}

¹College of Electronics and Information, Qingdao University, 266071 Qingdao, China

²Technical University of Sofia, Bulgaria

Abstract. The judgment of unvoiced and voiced sound based on sparse representation in DCT dictionary is implemented. Human pronunciation can be mainly divided into unvoiced and voiced sound. Sparse representation can represent the signal with as few coefficients as possible on a set of over-complete vectors, which can reveal the most representative features of signals. In this paper, the difference between the sparse representation of unvoiced and voiced sound is studied, based on which a method is proposed to distinguish unvoiced and voiced sound in words. The experimental results prove that the proposed method is effective.

1 Introduction

Speech is produced when air flows stimulate the vocal tract and eventually radiate from the lips or nostrils or both. Depending on whether the vocal cords vibrate or not, the speech can be divided into voiced sound and unvoiced sound. The judgment of unvoiced and voiced sound is important in speech recognition and coding. The voiced and unvoiced sounds have different characteristics. The voiced sound has the characteristic of periodic signal, while the unvoiced has the characteristic of random noise. Voice has resonance peak structure in frequency domain and its energy is mainly concentrated in low frequency band. The amplitude value of unvoiced is relatively small and there is no obvious regularity in time domain and frequency domain.

The traditional methods of distinguishing voiced and unvoiced sound are including zero crossing rate, short-time energy, short time auto correlation function and short time average amplitude difference function (AMDF). The traditional ways of distinguishing voiced from unvoiced are still to be improved [1].

In this paper, a method is proposed that use sparse representation on dictionary to judge voiced and unvoiced sound. Sparse representation can reduce redundant and useless information and represent the most representative features of the signal, greatly improve the transmission efficiency and storage efficiency of the signal. It can be also used to speech enhancement and reconstruction.

2 Sparse Representation and DCT dictionary

2.1 Sparse representation of signals

In the processing of signal, the signal is usually represented in the form of a set of specific vectors. Sparse representation requires that the set of coefficient vectors be as sparse as possible. This is expressed as follows [2-3]:

$$y = \sum_{i=1}^n d_i c_i = DC \quad (1)$$

In Equation (1), the signal is represented by matrix y where each column as a speech frame ($y \in \mathbf{R}^{M \times N}$), dictionary represented as a matrix ($D \in \mathbf{R}^{M \times N}$) with each column as an atom. Sparse representation ($C \in \mathbf{R}^{N \times N}$) is represented as a matrix with each column as a sparse coefficient vector corresponding to each speech frame. The solution model of sparse representation solution model has two forms as follows [4-5]:

$$\hat{c} = \arg \min \|y - DC\|_2 \quad \text{s.t. } g(c) \leq K \quad (2)$$

$$\hat{c} = \arg \min g(c) \quad \text{s.t. } \|y - DC\|_2 \leq \varepsilon \quad (3)$$

where K is sparsity constraint, $g(c)$ is sparsity metric and $g(c)$ has two forms. When $g(c)$ is l_1 norm, the solution is a kind of convex optimization algorithm. It is shown in l_0 norm, the solution of sparse representation is a greedy optimization algorithm.

The second model is selected where $g(c)$ is l_1 norm and Basis Pursuit (BP) is chosen to solve it [6]. BP has high accuracy and requires more time to computation [7]. Because this study mainly deals with short speech that has low computational complexity and it require precise to distinguish voiced and unvoiced sound.

* Corresponding author: xdzhuang@qdu.edu.cn

2.2 The DCT dictionary

Dictionaries are divided into fixed dictionaries and learning dictionaries. Fixed dictionaries are constructed by mathematical transformations, while learning dictionaries are trained that based on the characteristics of a kind of signals. Signals can be represented sparsely in all dictionaries. Because DCT is a common fixed dictionary with simple structure, it is used in the proposed method for sparse signal representation. DCT is represented as a matrix with each column as an atom. It is defined as follows [8-9]:

$$y(k) = a_k \sum_{n=0}^{N-1} x(n) \cos[\pi(2n+1)k / 2N] \quad 0 < k < N-1 \quad (4)$$

$$\begin{cases} a_k = \sqrt{1/N}, k=0 \\ a_k = \sqrt{2/N}, 1 \leq k \leq N-1 \end{cases}$$

In Equation (4), $\mathbf{x}(\mathbf{n})$ is the signal vector. The DCT transform is essentially the inner product of the signal with each column of the DCT matrix. The DCT dictionary is the extension of DCT matrix, which has a smaller frequency discretization interval:

$$y(k) = a_k \sum_{n=0}^{N-1} x(n) \cos[\pi(2n+1)k / 2N] \quad 0 < k < L-1 \quad (5)$$

where L is larger than N .

3 Sparse representation of voiced /unvoiced speech

The voiced and unvoiced pronunciation are recorded in a quiet laboratory environment and sampled at a rate of 16Khz with a precision 16 bits per sample. The frame size is 192ms and the overlapping of the frame is 32ms. DCT dictionary is a matrix with 192 dimensions and 384 columns. Unvoiced and voiced sounds are processed based on the DCT dictionary. And BP is selected to acquire sparse coefficients.

The $l_{0.5}$ norm of the sparse coefficient vectors is chosen to compare the difference of voiced and unvoiced. The sparsity of the l_0 norm of the sparse coefficient vectors is better than the l_1 norm of the sparse coefficient vectors, but it has complex calculation. To ensure the accuracy of the study, this method selects the $l_{0.5}$ norm of the sparse coefficient vectors as the feature value.

Experiments are carried out for signals of English phoneme pronunciations. Some typical experimental results are shown in Fig. 1 to Fig. 5. Comparing sparse coefficients of the unvoiced and voice sounds, it is found that unvoiced sound is not particularly sparse in the expansion of a DCT dictionary. The pronunciation structure of unvoiced pronunciation is similar to random noise, and the $l_{0.5}$ norm of sparse coefficients vectors of unvoiced sounds is larger than that of voiced sounds. However, the voiced sound has obvious sparsity in DCT dictionary, and the corresponding $l_{0.5}$ norm of sparse coefficients vectors is relatively small.

Fig.1. Results of the $l_{0.5}$ norm of sparse coefficient vectors of voiced “[a]”

Fig.2. Results of the $l_{0.5}$ norm of sparse coefficient vectors of voiced "[o]"

Fig.3. Results of the $l_{0.5}$ norm of sparse coefficient vectors of voiced "[u]"

Fig.4. Results of the $l_{0.5}$ norm of sparse coefficient vectors of unvoiced "[sh]"

Fig.5. Results of the $l_{0.5}$ norm of sparse coefficient vectors of unvoiced "devoice [e]"

4 Voiced/unvoiced judgement based on sparse feature

According to the sparse representation of voiced and unvoiced sounds in DCT dictionary, the threshold value of the $l_{0.5}$ norm of the sparse coefficient vectors is set to 600. A way that based on the sparse property of voiced and unvoiced has proposed to complete the judgement of voiced and unvoiced sounds. If the $l_{0.5}$ norm of sparse coefficients vectors are larger than 600, the frames are judged to be unvoiced and marked by 1 in the experiment. And if the $l_{0.5}$ norm of sparse coefficients vectors is less than 600, the frames are judged to be voiced and marked by 0. The flowchart of the proposed method is shown in Fig. 6.

Fig.6. Diagram of the judgment of voiced and unvoiced pronunciation based on sparse feature

Fig.7. The judgement of the unvoiced and voiced sound of the word “shape”

Fig.8. The judgement of the unvoiced and voiced sound of the word “show”

Fig.9. The judgement of the unvoiced and voiced sound of the word “face”

The judgement of unvoiced and voiced for some English words is implemented. Some of the typical results are shown in Fig. 7 to Fig. 9. Those frames with the $l_{0.5}$ norm of the sparse coefficient vectors less than 600 are judged as voiced sound, and are marked with 0 in these figures. The frames with the $l_{0.5}$ norm of the sparse coefficients vectors more than 600 are judges as unvoiced, and are marked with 1. There is a significant difference between the sparse representation of voiced and unvoiced pronunciation. And the results of the judgement are consistent with the actual pronunciation of words.

5 Conclusion

Sparse representation can reveal the most representative features of signals with few coefficients. A new method is proposed to distinguish voiced and unvoiced through utilizing their sparse coefficients based DCT dictionary in this study. The results show that unvoiced is similar to random noise and the $l_{0.5}$ norm of sparse coefficients vectors are generally large. Differently, the sparse representation of voiced sound has obvious sparsity. The $l_{0.5}$ norm of its sparse coefficients vectors is s mall. The experiments prove the effectiveness of the proposed method.

References

1. K. Pattanaburi, J. Onshaunjit, J. Srinonchat. International Symposium on Computer. IEEE, 2012.
2. H. Firouzi, M. Babaie-Zadeh, A. G. Sahebi, et al. Computer Science, 2008:1921-1924.
3. J. F. Gemmeke, T. Virtanen, A. Hurmalainen. IEEE Transactions on Audio, Speech, and Language Processing, 2011, **19**(7):2067-2080.
4. D. S. Williamson, Y. Wang, D. L. Wang. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.
5. Y. Tang, Y. Chen, N. Xu, et al. International Conference on Wireless Communications & Signal Processing. IEEE, 2015.
6. D. Wipf, S. Nagarajan. IEEE Journal of Selected Topics in Signal Processing, 2010, **4** (2):317-329.
7. Y. Yao, X. Xin, G. Ping. Neural Information Processing. 2012.
8. Ding H., Soon I. Y., Yeo C. K. IEEE Transactions on Audio Speech and Language Processing, 2011, **19**(8):2614-2623.
9. N. Sripriya, T. Nagarajan. 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013). IEEE, 2014.