SYNCHROINFO JOURNAL

www.media-publisher.eu/synchroinfo-journal

NOISY SPEECH COMMANDS RECOGNITION ALGORITHM BASED ON TEST SPECTRAL TRANSFORMATIONS OF INPUT SIGNAL

Chi Thien Nguyen [1]

[1] Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam

ABSTRACT

In practice, the result of signal recognition is degraded by noise. Training speech signals are usually noise-free, while testing speech signals are noisy. The presence of noise leads to a strong deviation of the spectra of the testing speech signals from the spectra of their standards in the training sample. Therefore, the quality of the recognition result against a background of noise drops sharply. The article proposes a trial amplification of the speech signal spectrum in the recognition process. A multiple algorithm for recognizing commands against a background of noise is compared with a single algorithm for recognizing speech commands. The problem of recognition of speech commands on the background noise is reviewed. The developed numerical algorithm of recognition is studied. The results of the experiments are reported on 11 speech commands from the TIDigits dataset.

KEYWORDS: *recognition of speech commands, noise, multiple algorithm*

DOI: 10.36724/2664-066X-2024-10-5-16-21

Received: 25.07.2024
Accepted: 14.09.2024

Citation: Chi Thien Nguyen, "Noisy speech commands recognition algorithm based on test spectral transformations of input signal" *Synchroinfo Journal* **2024**, vol. 10, no. 5, pp. 16-21

Licensee IRIS, Vienna, Austria.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Copyright: © 2024 by the authors.

Introduction

In practice, the speech signals recognition [1-6] result is degraded by noise. Training speech signals are usually noise-free, while testing speech signals are noisy. The presence of noise leads to a strong deviation of the spectra of the testing speech signals from the spectra of their standards in the training sample. Therefore, the quality of the recognition result against a noise background drops sharply [7]. In [7], to solve the problem of recognizing speech commands against a noise background, it is proposed to amplify the signal spectrum by a constant. This means that the values of the signal amplitude spectrum samples are increased by a constant. In [8], algorithms were proposed for determining optimal gain constants for each application condition and a single gain constant for different application conditions. The positive results of these studies confirm the improvement in recognition quality by amplifying the signal spectrum. This raises an important question about what to do if we do not have any a priori information about the noise (noise type, noise level).

Command recognition against a noise background.

A trial amplification of the speech signal spectrum during the recognition process is proposed. When executing the algorithm for recognizing speech commands against a noise background in [7], the speech signal is transformed with a fixed value of the gain constant c. We will call such an algorithm a single-shot recognition algorithm. If a trial transformation of the speech signal during the recognition process is taken into account, i.e. the constant c can change, then the algorithm for recognizing speech commands against a noise background becomes multiple. The steps of the multiple algorithm for recognizing speech commands against a noise background (MARSC) are given below:

1. Construct a sequence $A = (a_1, a_2, a_3,...)$ of short-term spectra [9] $\mathbf{a}_i$ ($a_i^k, 1 \le k \le N/2$) from a speech signal $Y = (y_1,..., y^T)$.

2. Take one value of the gain constant c from a predetermined range [0, 0.1,..., 1.9]. Increase the values of the amplitude spectra by the next value of the constant c. After "amplifying" the short-term amplitude spectrum by $c \ge 0$, a new sequence of amplitude spectra is obtained $\tilde{A} = \{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \tilde{\mathbf{a}}_3, ...\}$, where $\tilde{\mathbf{a}}_i = \{\tilde{\mathbf{a}}_i^k, 1 \le k \le N/2\}$, $\tilde{a}_i^k = \tilde{a}_i^k + c$.

3. Obtain a sequence $X = (x_1, x_2, x_3,...)$ of vectors of small-frequency cepstral coefficients [9] $\mathbf{x}_t$ ($x_t^m, 1 \le m \le M$) from the sequence $\tilde{A} = \{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \tilde{\mathbf{a}}_3, ...\}$, of short-term amplitude spectra.

4. Calculate the probability $p(X(c) \mid \lambda^v(c))$ of the sequence vectors X mel-frequency cepstral coefficients with respect to each signal class $v = 1,2,...,V$, where the parameter $\lambda^v(c)$ describes the v-th signal class after amplifying their spectrum by a constant c. In the database, each signal class $v = 1,...,V$ describes a set of standards $\lambda^v(c)$ with different levels of amplification of signal spectra $c \in [0, 0.1,..., 1.9]$..

5. Repeat teps 2-4 for all values of the constant c.

6. Among all set (c, v), find the set (c^*, v^*) that provides the maximum probability $p(X(c) \mid \lambda^v(c))$.

Thus, the corresponding number v* of the signal class is:

$$v^* = \arg \max_{c} \max_{\lambda^v} p(X(c) \mid \lambda^v(c)), \; v = 1,2,...,V.$$

Thus, when executing the MARKS algorithm, some value of the gain constant c is selected from the range [0, 0.1,..., 1.9]. In the general case, the use of some value of the gain constant c from this range does not mean at all that the mel-frequency cepstral representation of the input signal becomes closer to the mel-frequency cepstral representation of the reference signals. But, ultimately, such an optimal value of the gain constant c will be selected that will still improve the quality of recognition of the input signal, which means approaching the mel-frequency cepstral representation of the reference signals.

Algorithm study for recognizing commands against a noise background

A comparison of the multiple-shot command recognition algorithm against a background of noise MARKS and the single-shot speech command recognition algorithm (CRA) is performed. The CRA algorithm is a variant of the MARKS algorithm with the gain constant $c = 0$. Experiments were conducted on 11 speech commands from the TIDigits dataset [10-13]. The set of 440 speech signals from 40 speakers is randomly divided into two samples (each sample contains signals from 20 speakers who pronounced each command once). One sample plays the role of a training sample, the other is used as a test sample. The training sample is used to train the MARKS and CRA algorithms. Noise with a signal-to-noise ratio of R_{sn} (dB) was artificially added to the test speech signals.

For a given speech signal $\Psi = \{\psi_1,...,\psi_T\}$ and noise $\Xi = \{\xi_1,...,\xi_T\}$ with value R_{sn}, the noisy speech signal $Y = \{y_1,..., y_T\}$ is formed by the formula [14].

$$y_t = \psi_t + 10^{-\frac{R_{sn}}{20}} \xi_t \sqrt{\sum_{i=1}^{T} \psi_i^2 / \sum_{i=1}^{T} \varsigma_i^2}, \quad t = 1,...,T.$$

The recognition of speech signals contaminated with additive white Gaussian noise is considered. Figure 1 shows additive white Gaussian noise and its amplitude spectrum. For example, for additive white Gaussian noise with a noise level of R_{sn} = 6.9,12,15 dB for the model of signal classes as two-component random processes, recognition is performed by the MARKS and CRA algorithms with the number of recognition errors counted. Figure 2 shows the recognition result.

It turned out that for additive white Gaussian noise with a noise level of R_{sn} = 6.9,12,15 dB, on average, the use of the MARKS algorithm leads to a decrease in the number of recognition errors compared to the use of the CRA algorithm by 51.59%.

The recognition of speech signals contaminated with real environmental noise from the exhibition hall is considered [15].

Figure 3 shows the ambient noise from the exhibition hall and its amplitude spectrum. For example, for ambient noise from the exhibition hall with a noise level of R_{sn} = 6.9,12.15 dB, for the model of signal classes as two-component random processes, recognition is performed by the MARKS and CRA algorithms with the number of recognition errors counted. Figure 4 shows the recognition result. It turned out that for ambient noise from the exhibition hall with a noise level of R_{sn} = 6.9,12.15 dB, on average, the use of the MARKS algorithm leads to a decrease in the number of recognition errors compared to the use of the CRA algorithm by 31.25%.

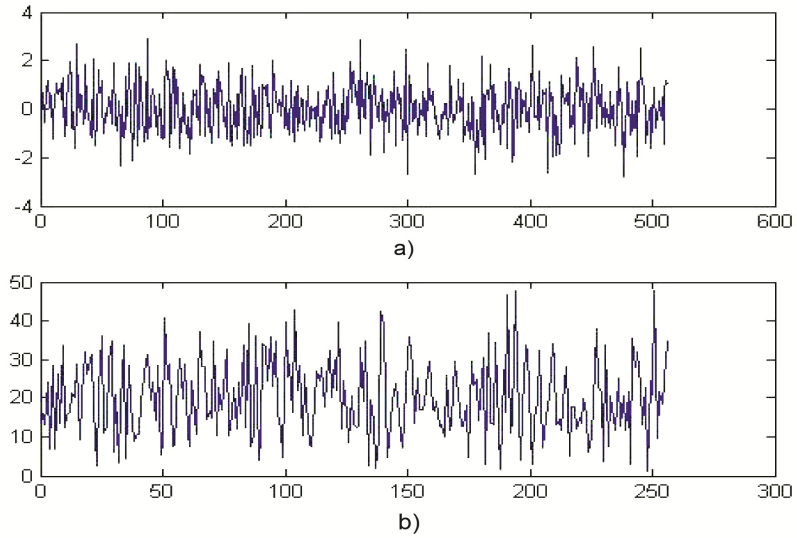

Figure 1. Additive white Gaussian noise a) and its amplitude spectrum b)

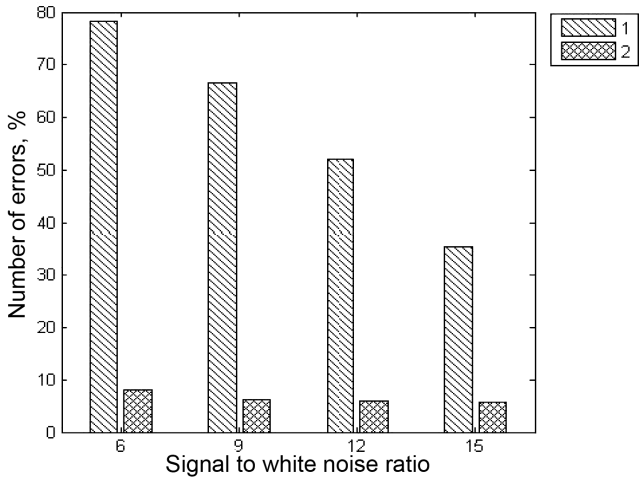

Figure 2. Number of recognition errors by algorithms: 1 – CRA; 2 – MARKS

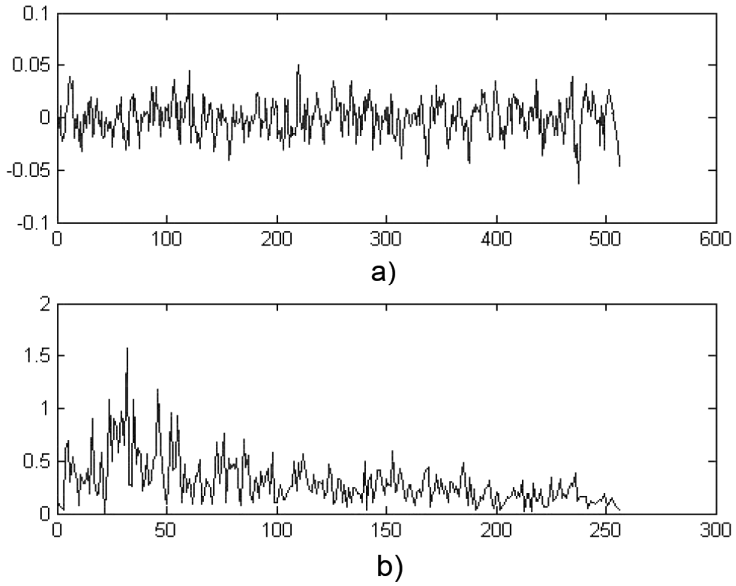

Figure 3. Ambient noise from the exhibition hall a), and its amplitude spectrum b)

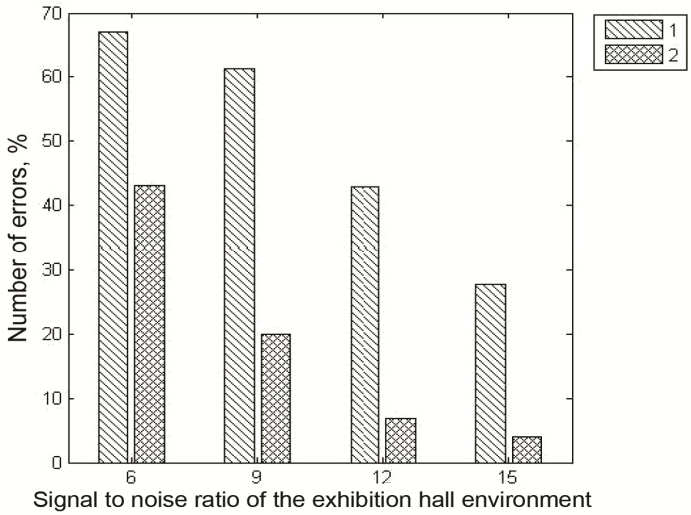

Figure 4. Number of recognition errors by algorithms: 1 – CRA; 2 – MARKS

Recognition of speech signals contaminated with real noise inside a moving subway train is also considered [15].

Figure 5 shows the noise inside a moving subway train and its amplitude spectrum.

As a conclusion after analyzing the literature, it can be said that multiplexers and demultiplexers are the basis for creating switches that control the flow of transmitted information in optical systems, while the synchronicity of their operation will be ensured by PLL devices implemented as part of specialized microcircuits manufactured using CMOS technology with submicron design standards.

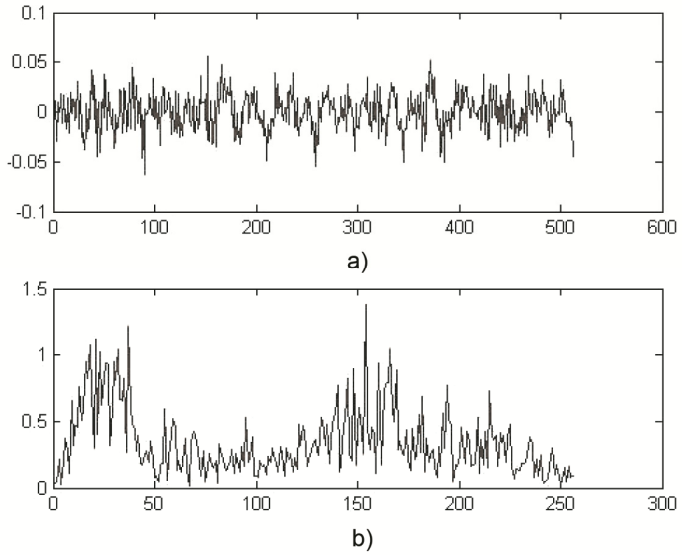

Figure 5. Noise inside a moving subway train a), and its amplitude spectrum b)

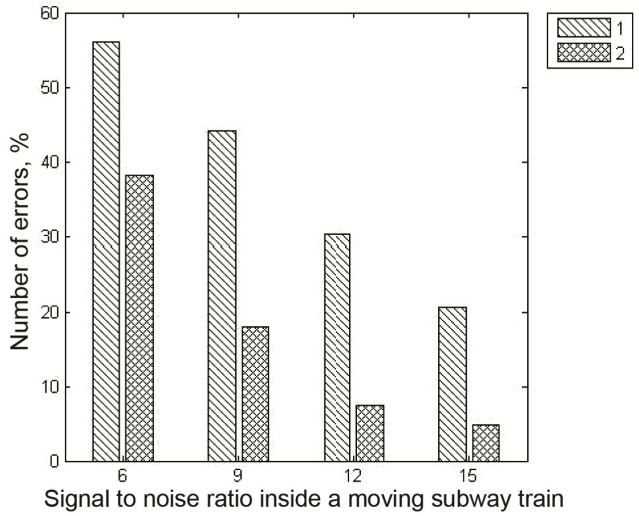

Figure 6. Number of recognition errors by algorithms: 1 – CRA; 2 – MARKS

For example, for the noise inside a moving subway train with a noise level of R^{sn} = 6.9,12.15 dB, for the model of signal classes as two-component random processes, recognition is performed by the MARKS and CRA algorithms with the number of recognition errors counted. Figure 6 shows the recognition result.

It turned out that for the noise inside a moving subway train with a noise level of R_{sn} = 6.9,12.15 dB, on average, the use of the MARKS algorithm leads to a decrease in the number of recognition errors compared to the use of the CRA algorithm by 20.68%.

Thus, the experiments show that the MARKS algorithm effectively improves the quality of speech command recognition against a noise background.

REFERENCES

[1] G. K. Berdibayeva, A. N. Spirkin, O. N. Bodin and O. E. Bezborodova, "Features of Speech Commands Recognition Using an Artificial Neural Network," *2021 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, Yekaterinburg, Russia, 2021, pp. 0157-0160, doi: 10.1109/USBEREIT51232.2021.9455111.

[2] Daniel-S. Arias-Otalora, Andrés Florez, Gerson Mellizo, C. H. Rodríguez-Garavito, E. Romero, J. A. Tumialan, "A Machine Learning Based Command Voice Recognition Interface", *Applied Computer Sciences in Engineering*, vol.1685, pp.450, 2022.

[3] A. R B, V. R C, V. K, S. Chikamath, N. S R and S. Budihal, "Limited Vocabulary Speech Recognition," *2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, 2024, pp. 1-5, doi: 10.1109/INOCON60754.2024.10511500.

[4] A. Kuzdeuov, S. Nurgaliyev, D. Turmakhan, N. Laiyk and H. A. Varol, "Speech Command Recognition: Text-to-Speech and Speech Corpus Scraping Are All You Need," *2023 3rd International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, Singapore, Singapore, 2023, pp. 286-291, doi: 10.1109/RAAI59955.2023.10601292.

[5] Aditya Kulkarni, Vaishali Jabade, Aniket Patil, "Audio Recognition Using Deep Learning for Edge Devices", *Advances in Computing and Data Sciences*, vol.1614, pp.186, 2022.

[6] A. Yasmeen, F. I. Rahman, S. Ahmed and M. H. Kabir, "CSVC-Net: Code-Switched Voice Command Classification using Deep CNN-LSTM Network," *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Kitakyushu, Japan, 2021, pp. 1-8, doi: 10.1109/ICIEVicIVPR52578.2021.9564183.

[7] C. T. Nguyen, “Solution of the problem of speech command recognition against a noise background,” *Bulletin of Tula State University. Technical sciences*. Issue 11. Tula: Tula State University Publishing House, 2013, pp. 241-250.

[8] C. T. Nguyen, “Optimization of the parameters of the heuristic model of speech signals in order to improve the quality of their recognition,” *Bulletin of Tula State University. Technical sciences*. 2014. Issue 1, pp. 44-50.

[9] J. Benesty et al., “Handbook of speech processing.” Springer, 2008. 1159 p.

[10] G. Leonard, G. Doddington, TIDigits [Electronic resource]. Linguistic Data Consortium, Philadelphia, 1993. URL: https://catalog.ldc.upenn.edu/LDC93S10 (date of access: 23.03.2024).

[11] H. Aghakhani et al., "Venomave: Targeted Poisoning Against Speech Recognition," *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, Raleigh, NC, USA, 2023, pp. 404-417, doi: 10.1109/SaTML54575.2023.00035.

[12] L. Guo et al., "Transformer-Based Spiking Neural Networks for Multimodal Audiovisual Classification," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 3, pp. 1077-1086, June 2024, doi: 10.1109/TCDS.2023.3327081.

[13] S. Xiang et al., "Neuromorphic Speech Recognition with Photonic Convolutional Spiking Neural Networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 29, no. 6: Photonic Signal Processing, pp. 1-7, Nov.-Dec. 2023, Art no. 7600507, doi: 10.1109/JSTQE.2023.3240248.

[14] K. Wojcicki, “Add noise to a signal at a prescribed SNR level,” URL: http://www.mathworks.com/matlabcentral/ (date of access: 10.03.2024).

[15] http://labrosa.ee.columbia.edu/sounds/noise/ (date of access: 15.03.2024).