Low signal-to-noise ratio speech classification with wavelet

Yeguang Huang¹, Yifen Peng¹, Wenjing Zhao² and Zhong Cao¹,³,⁴

¹School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China
²Lab Center, Network and Modern Educational Technology Center, Guangzhou University, Guangzhou 510006, China
³Institute of Computing Science and Technology, Guangzhou University, Guangzhou 510006, China
⁴E-mail: zhongc@gzhu.edu.cn

Abstract. This paper evaluates the potential of wavelet in low signal-to-noise ratio (SNR) speech classification. A wavelet feature extraction method is proposed, which includes wavelet analysis, wavelet denoising and sliding variance. The potential of wavelet feature extraction method is evaluated on two different datasets with Convolutional Neural Network. The results of our model and comparison show that the performance is significantly improving comparing with other state-of-the-art approaches.

1. Introduction

Speech recognition dates back to the 1920s, yet people make a major breakthrough in speech recognition in the 1960s. Since then, people use Hidden Markov Model (HMM) [1] and Gaussian Mixture Model (GMM) [2] to recognize speech. With the introduction of Deep Neural Network (DNN), people started to recognize speech by constructing Recurrent Neural Network (RNN) [3]. Due to the weakness of RNN [4], gradient descent is difficult, people proposed Long short-term memory (LSTM) [5, 6] to recognize speech and achieved better results.

In recent years, people started to construct Convolutional Neural Network (CNN) to recognize speech [7] because CNN succeeded in image classification. It is generally relying on feature extraction methods such as Mel Frequency Cepstral Coefficient (MFCC) [8, 9] and Perceptual Linear Prediction (PLP) [10, 11]. These methods can’t achieve good results in a low SNR environment, because MFCC and PLP can’t extract the features of speech well in the case of low SNR.

The feature extraction methods such as MFCC and PLP based on Fourier transform and Fourier transform can’t remove noise effectively in low SNR environments. The original signal will be masked by noise in the spectrum in low SNR environments, so the Fourier transform can’t remove the noise. But the wavelet transform is different from the Fourier transform [12], the wavelet transform can remove the noise better [13]. If wavelet transform is used for feature extraction, it can remove most noise in the original signal and then it can achieve better recognition accuracy. Whether wavelet transform can be used for speech classification with a low SNR environment well? This article is intended to provide an answer to this question.
2. Wavelet feature extraction

In the field of speech signal processing, wavelet analysis has many advantages compared with Fourier analysis. In essence, wavelet feature extraction bases on wavelet comparable to MFCC and PLP based on Fourier analysis.

2.1. Wavelet analysis

MFCC and PLP process speech signals with Fourier analysis. However, wavelet feature extraction processes speech signals with wavelet analysis. Wavelet analysis converts speech signals in time domain into wavelet features in wavelet domain. This process is wavelet analysis.

Wavelet analysis is a one-dimensional multi-scale wavelet decomposition of one-dimensional speech signals. Speech signal is decomposed and converted into wavelet decomposition vector. And then extract the wavelet coefficients of the wavelet decomposition vector. Wavelet coefficients consist of wavelet high frequency coefficient and wavelet low frequency coefficient. One-dimensional multi-scale wavelet decomposition:

\[ f(t) = \sum_k c_j^0(k) \varphi_{j_0,k}(t) + \sum_j \sum_{k=0}^{\infty} d_j^k(t) \psi_{j,k}(t), \]

and

\[ c_j^0(k) = \langle f(t), \varphi_{j_0,k}(t) \rangle = \int f(t) \varphi_{j_0,k}(t) dt, \]

\[ d_j^k(k) = \langle f(t), \psi_{j,k}(t) \rangle = \int f(t) \psi_{j,k}(t) dt, \]

\[ c_j^0(k) \] is wavelet low frequency coefficient of original speech signal, \( d_j^k(k) \) is wavelet high frequency coefficient of original speech signal. \( \varphi_{j_0,k}(t) \) is the scale function in the one-dimensional wavelet decomposition, and \( \psi_{j,k}(t) \) is the wavelet function in the one-dimensional wavelet decomposition.

Wavelet low frequency coefficient and wavelet high frequency coefficient represent the feature of the original speech signal in wavelet domain. Figure 1 shows the entire process of wavelet analysis.

Figure 1. Wavelet analysis.

2.2. Wavelet denoising

There is a lot of noise in the original speech signal usually. They will have a bad influence on the classification of speech by CNN. In order to reduce the influence, we need to try to reduce the noise in the speech signal. The wavelet denoising could reduce the noise of the wavelet coefficients. The selection method of wavelet coefficient threshold is Sqtwolog as follows:

\[ \lambda = \sigma \sqrt{2 \log(N)}, \]

where \( \lambda \) is threshold, \( N \) is the length of the wavelet coefficient, \( \sigma \) is the noise coefficient as follows:
\[ \sigma = \frac{MAD}{0.6745} = \frac{\text{median}(|\omega|)}{0.6745}, \]

where \( \omega \) is wavelet coefficient, and \( MAD \) is the median of the absolute value of the wavelet coefficients. We process the wavelet coefficients with the calculated thresholds. The threshold processing method is Soft Thresholding as follows:

\[
Y = \begin{cases} 
  x - \lambda, & x > \lambda \\
  0, & \text{otherwise} \\
  x + \lambda, & x < -\lambda 
\end{cases}
\]

where \( Y \) is processed wavelet coefficients, \( x \) is the original wavelet coefficient. Wavelet denoising can reduce the noise of the speech signal and improve the performance of CNN on speech classification.

2.3. Sliding variance

The high frequency coefficients have high resolutions in the time-domain, so the high frequency coefficients are more than the low frequency coefficients. There is a large number of coefficients after wavelet analysis. So we need to construct a big neural network for speech classification. Due to the high time cost for training a big model, constructing a big neural network is infeasible. Sliding variance is calculating the variance of wavelet coefficients. The variance represents degree of dispersion and it also represents the characteristics of the original signal. Sliding variance can compress the wavelet coefficients and keep the characteristics of the original signal.

Sliding variance is processed by dividing the one-dimensional wavelet coefficients into short segments. A short segment is a frame and make sure each frame overlap with the other when framing. And then calculate the variance for each frame. The number of the calculated variances is the same by adjusting the sliding window and stride. The calculated variances form a matrix vector. Matrix vector is input into CNN for speech classification. Figure 2 shows the entire process of sliding variance.

3. Dataset

To evaluate the potential of the wavelet feature extraction for low SNR speech classification in CNN, we recorded a dataset for evaluation.

We made an app and we recorded the voices of five people in a quiet environment. The voices of five people make up of a dataset. We call it homemade dataset.

The voice’s frequency is 16000Hz. Each voice is 2 seconds. The dataset consists of commands from five different people. Everyone has five kinds of commands, which are “GO”, “BACK”, “LEFT”, “RIGHT” and “STOP”. Each command is a category for each person. There are 25 categories and divide voice into 25 folders.
We also use a public dataset. It is Free-Spoken-Digit-Dataset [14] (FSDD). This dataset consists of 2000 voices from four different people. Everyone has ten kinds of commands, which are 0 to 9. Each command is a category for each person. There are 40 categories and divide voice into 40 folders.

We think the original dataset is the clean dataset. Gaussian white noise was added into the dataset for simulating a low SNR environment. The SNR of dataset added Gaussian white noise is 0db, -5db, -10db, -15db and -20db. Gaussian white noise is from the NoiseX-92 noise library [15].

We have two datasets in five environments. In each environment, we compare the neural network accuracy of the three methods on the two datasets.

4. Speech classification
We usually use MFCC and PLP to extract the features of speech and construct a CNN to classify the extracted feature vectors. This can achieve good results. If the SNR of speech is low, this method can’t achieve good results. However, the wavelet feature extraction proposed in this paper can achieve good results in the speech classification of CNN in the environment of low SNR.

4.1. Feature extraction
The two datasets with Gaussian white noise were extracted by three feature extraction methods, wavelet feature extraction, MFCC and PLP.

Figure 3 shows the entire process of wavelet feature extraction:

- Wavelet analysis is performed on speech firstly in wavelet feature extraction. The original speech signal is converted into wavelet coefficients in the wavelet domain. Wavelet coefficients represent the characteristics of speech signals in the wavelet domain.
- In order to reduce the influence of noise in original speech, wavelet denoising is performed on wavelet coefficients. Wavelet denoising can remove most noise in the original speech signal.
- Then the wavelet coefficients are arranged in an inverted triangle. And the vertical axis represents a change in frequency and the horizontal axis represents a change in time. It can better represent the wavelet features of original speech signal.
- Calculate the sliding variance of the coefficients.
- Normalize the variance and the normalization method is Min-Max normalization [16].
- To better classify in the CNN, the variance is converted to a square matrix vector.

Figure 3. Wavelet feature extraction.

The square matrix vector is the feature of the original speech after wavelet feature extraction. MFCC and PLP use rastanmat [17] implement to transform original speech into vectors. Figure 4 shows the entire process of MFCC and PLP.
Figure 4. MFCC, PLP and three methods’ feature vectors input to CNN for speech classification.

4.2. Convolutional neural network

Training a CNN involving many factors, two of which are important. One is the architecture of the CNN. The other is learning hyper-parameter.

Detailed evaluation of all combinations is not possible because training a complete model takes long time. So, most promising models are selected based on effective validation for the most important factors (layers and the number of kernels, shape of kernels, learning rate, dropout). Figure 5 shows the final CNN system used for evaluation:

- A first convolution ReLU layer consisted of 16 filters (3×3 size, 1×1 stride) and max-pooling (2×2 pool size, 2×2 pool stride).
- A second convolution ReLU layer consisted of 32 filters (3×3 size, 1×1 stride) and max-pooling (2×2 pool size, 2×2 pool stride).
- A third convolution ReLU layer consisted of 64 filters (3×3 size, 1×1 stride).
- A flattening layer.
- A first fully connected layer.
- A second fully connected layer.
- Output layer.

In order to prevent CNN from overfitting, L2 regularization [18] is introduced in the calculation of the CNN loss function and dropout learning is introduced in the convolutional layer.

The features of the two datasets in different environments were extracted by three extraction methods: wavelet feature extraction, MFCC and PLP. And training in CNN to get the results.
4.3. Results
The CNN model mentioned above was evaluated by speech features extracted in three methods. Table 1 is the classification accuracy of the homemade dataset and Table 2 is the classification accuracy of the public dataset FSDD. Figure 6 is a histogram of the results of the three methods in five environments of different SNR.

| Feature     | SNR(-20) | SNR(-15) | SNR(-10) | SNR(-5) | SNR(0)  |
|-------------|----------|----------|----------|---------|---------|
| Wavelet     | 24.33%   | 60.47%   | 82.67%   | 88.40%  | 92.07%  |
| MFCC        | 16.80%   | 40.33%   | 72.53%   | 85.53%  | 92.80%  |
| PLP         | 19.73%   | 54.80%   | 76.00%   | 90.06%  | 92.93%  |

| Feature     | SNR(-20) | SNR(-15) | SNR(-10) | SNR(-5) | SNR(0)  |
|-------------|----------|----------|----------|---------|---------|
| Wavelet     | 10.83%   | 32.67%   | 53.00%   | 66.50%  | 69.67%  |
| MFCC        | 8.83%    | 29.33%   | 42.83%   | 62.50%  | 69.83%  |
| PLP         | 13.33%   | 29.50%   | 44.33%   | 66.67%  | 74.00%  |

Figure 6. The accuracy of two datasets (left: homemade dataset, right: FSDD).

As can be seen from the result, the wavelet feature extraction is generally able to achieve better classification accuracy in all environments. In the homemade dataset, when the SNR is -10db, -15db and -20db, it is obvious that the wavelet feature extraction method performs better than the other two methods. In the public dataset, the wavelet feature extraction method performs better than the other two methods when the SNR is -10db and -15db. The wavelet feature extraction method performs better than MFCC in all cases except 0db in both datasets. In the case of -5db and 0db, the wavelet feature extraction method is not better than the other two methods in both datasets, and even worse than the other two methods.

In summary, the wavelet feature extraction method has better performance than the other two feature extraction methods in low SNR environment (especially in the case of -10db and -15db). But in high SNR environment (such as in the case of 0db and -5db), the wavelet feature extraction method is not better than the other two methods.

5. Summary
The objective of this paper is to evaluate the performance of the wavelet feature extraction method in the speech classification with a low SNR environment.

From the results, it is indeed a feasible solution. The experiment shows that wavelet feature extraction is better than the other two methods in low SNR environment. However, in other cases,
such as when the SNR is not low, the wavelet feature extraction does not perform much better than the other two methods. What's more, this method provides a processing scheme for low SNR and increased choice for people in this extreme environment.

The proposed method in this paper increases the possibility of speech classification in a low SNR environment and seems to focus on speech processing in a low SNR environment.

Acknowledgements
This work is supported by the Humanities and Social Sciences Research Project of the Ministry of Education (18YJC880141) and Innovation training program for college students of Guangzhou University (S201911078048).

References
[1] Rabiner L R 1989 A tutorial on hidden Markov models and selected applications in speech recognition Proc of the IEEE 77(2) 257-286
[2] Reynolds D A 1995 Speaker identification and verification using Gaussian mixture speaker models Speech Communication 17(1-2) 91-108
[3] Graves A, Mohamed A and Hinton G 2013 Speech recognition with deep recurrent neural networks 2013 IEEE Int Conf on Acoustics, Speech and Signal Proc IEEE 6645-6649
[4] Bengio Y, Simard P and Frasconi P 1994 Learning long-term dependencies with gradient descent is difficult IEEE Transactions on Neural Networks 5(2) 157-166
[5] Hochreiter S and Schmidhuber J 1997 Long short-term memory Neural Computation 9(8) 1735-1780
[6] Sak H, Senior A and Beaufays F 2014 Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition preprint arXiv: 2014, 1402.1128
[7] Piczak K J 2015 Environmental sound classification with convolutional neural networks 2015 IEEE 25th Int Workshop on Machine Learning for Signal Proc (MLSP) IEEE 1-6
[8] Muda L, Begam M and Elamvazuthi I 2010 Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques preprint arXiv: 2010, 1003.4083
[9] Tiwari V 2010 MFCC and its applications in speaker recognition Int Journal on Emerging Technologies 1(1) 19-22
[10] Hermansky H 1990 Perceptual linear predictive (PLP) analysis of speech the Journal of the Acoustical Society of America 87(4) 1738-1752
[11] Dave N 2013 Feature extraction methods LPC, PLP and MFCC in speech recognition Int Journal for Advance Research in Engineering and Technology 1(6) 1-4
[12] Daubechies I 1990 The wavelet transform, time-frequency localization and signal analysis IEEE Transactions on Information Theory 36(5) 961-1005
[13] Soon Y, Koh S N and Yeo C K 1997 Wavelet for speech denoising Proc of IEEE TENCON'97. IEEE Region 10 Annual Conf. Speech and Image Technologies for Computing and Telecommunications (Cat. No. 97CH36162) Brisbane, Australia 2 479- 482
[14] Jackobovski 2017 Free-spoken-digit-dataset https://github.com/Jakobovski/free-spoken-digit-dataset
[15] Don J 2017 Noisex92. http://spib.linse.ufsc.br/noise.html
[16] Patro S and Sahu K K 2015 Normalization: A preprocessing stage preprint arXiv: 2015, 1503.06462
[17] Dan Ellis 2005 PLP and RASTA (and MFCC, and inversion) in Matlab http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/
[18] Ng A Y 2004 Feature selection, L1 vs. L2 regularization, and rotational invariance Proc of the 21st Int Conf on Machine Learning ACM 78