Spoken Digit Recognition Using Wavelet Scalogram and Convolutional Neural Networks

Roneel V. Sharan
Australian Institute of Health Innovation
Macquarie University
Sydney, Australia
roneel.sharan@mq.edu.au

***Abstract*—Spoken digit recognition finds numerous applications in digital technologies. Various feature engineering and classification strategies have been proposed for this purpose. This work explores the use of convolutional neural network (CNN) for spoken digit recognition. CNN is originally an image classifier and time-frequency representation of the spoken digit is used in this work to get an image-like representation. In particular, wavelet transform is used in forming the time-frequency representation as it provides better frequency localization for low frequency signals such as speech. The time-frequency representation is resized to a common dimension using bicubic interpolation and the resulting image-like representation, referred as scalogram, is used for recognizing spoken digits using CNN. In addition, late fusion is employed to combine the learning from scalogram representation and conventional time-frequency representations. The proposed approach is evaluated on a dataset containing 56,290 segments belonging to ten spoken digits, non-digits comprised of various other spoken words, and background noise. An overall validation and test error of 2.85% and 2.84% is achieved using the proposed method, outperforming various conventional methods.**

Keywords—bicubic interpolation, convolutional neural networks, late fusion, scalogram, spoken digit recognition, wavelet transform

I. INTRODUCTION

Handwritten digit recognition has received significant attention in the past decade. The availability of the MNIST dataset, a large public dataset of handwritten digits with a training set of 60,000 samples and a test set of 10,000 samples, provided an avenue for advancements in deep learning, a machine learning technique shown to perform particularly well on large datasets.

Similar to handwritten digit recognition, spoken digit recognition has various applications such as audio content analysis and retrieval, credit card number entry, voice dialing, and data entry [1, 2]. However, spoken digit recognition has received relatively less attention. Some related works in spoken digit recognition can be found in [3-5]. The TIDIGITS dataset is utilized in [3] which contains 2,412 training utterances and 1,144 test utterances. The OGI Multilanguage Corpus is utilized in [4] with 826 samples for training and 454 samples for testing. Their method utilizes mel-frequency cepstral coefficients (MFCC) as features, feature dimension reduction using principal component analysis (PCA), and classification using support vector machine (SVM).

In summary, a number of related works on spoken digit recognition utilize small datasets with conventional feature extraction and classification methods. Models trained on small datasets can be prone to overfitting and have poor generalization. In addition, most of these feature extraction, feature selection, and classification techniques have been superseded by deep learning techniques in recent times. One of the most comprehensive and recent related work that could be found is on the AudioMNIST dataset where deep learning techniques are utilized on a dataset of 30,000 audio digit samples [5]. This dataset, however, does not contain non-digit audio samples which is important in a realistic setting.

Convolutional neural network (CNN) is a deep learning technique for image classification. It has produced encouraging results in image classification tasks [6]. The robustness of CNN has seen its extension to non-image classification tasks, including audio signal classification where CNN outperformed conventional feature extraction and classification techniques [7, 8].

This work studies the use of CNN in spoken digit recognition. Being an image classifier, a key challenge is to find an appropriate image-like representation of the spoken digit signals. Various time-frequency representations of audio signals have been studied for use with CNN. The conventional time-frequency representation, the spectrogram, is probably the most common and has been used in spoken digit recognition [5]. Another common approach is the use of frequency domain filterbanks. Two commonly used filters are moving average filters and mel-filters, as used in the computation of MFCCs. The resulting time-frequency representations are referred as smoothed-spectrogram and mel-spectrogram, respectively. These representations have shown to be useful in speech and acoustic event classification [7-9].

In this work, the spectrogram, smoothed-spectrogram and mel-spectrogram form the baseline methods. In addition, this work investigates the formation of time-frequency representation using wavelet transform. Wavelet transform offers better frequency localization in the lower frequency range making it more suitable for speech classification tasks compared to conventional techniques. Furthermore, different time-frequency representations reveal spectral information at different frequencies. Combining the learning from these representations may help improve the classification performance. In this work, late fusion is utilized as a way to make a more informed prediction.

The proposed approach is evaluated on a comprehensive dataset. While it is important for an automated spoken digit recognition system to be able to accurately detect spoken digits, it is also important to reject non-digits and other background noise. The audio dataset used in this work utilizes non-digit speech files and

978-1-7281-9052-5/20/$31.00 ©2020 IEEE

background noise, including complete silence, in addition to the spoken digit files. This makes it a very realistic and challenging dataset.

The rest of the paper is structured as follows. An overview of the dataset used in this work is given in Section II along with the method in time-frequency image formation and classification. The experimental results are presented in Section III and discussion and conclusion in Section IV.

II. Method

The dataset used in this work is described first followed by the technique used in the formation of wavelet scalogram, the CNN architecture, and late fusion.

A. Dataset

The Speech Commands dataset [10] is utilized in this work. The overall dataset has 38,908 spoken digit segments (0-9) and 66,921 non-digits segments (belonging to 25 classes), all of 1 second duration. In total, there are 105,829 utterances of 35 words from 2,618 speakers. In addition, there are 6 background sound files which vary in duration from 60s to 95s.

In this work, all the spoken digit segments available in the dataset are used together with about 20% of the non-digit segments (13,382), referred here as the *unknown* class. The validation and test data from digit and non-digit segments is determined as per the information provided with the dataset. Also, 4,000 segments of duration 1 second are taken from the background files. The volume of the background segments is rescaled from silent to loud and 80% are used for training, 10% for validation, and 10% for test.

The final dataset, therefore, has 45,013 training samples, 5,363 validation samples, and 5,914 test samples, a total of 56,290 samples. The distribution of the training, validation, and test data is provided in Fig. 1.

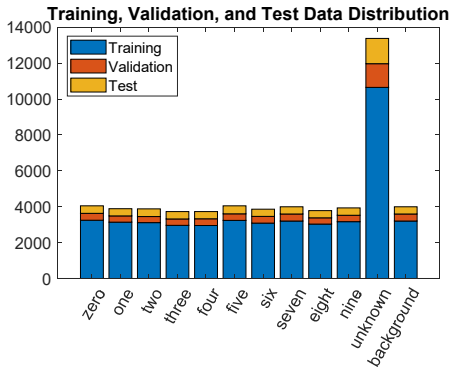

Fig. 1. Distribution of training, validation, and test data.

B. Wavelet Scalogram

The continuous wavelet transform for the time-domain audio signal $r(t)$ at scale s and position u is computed as

$$W_{\psi}(u,s)=\int_{-\infty}^{\infty} r(t)\frac{1}{\sqrt{s}}\psi*\left(\frac{t-u}{s}\right)dt \qquad (1)$$

where ψ is the mother wavelet [11]. The analytic wavelets [12] investigated in this work are Morlet wavelet, Morse wavelet, and bump wavelet.

The absolute value of the complex wavelet transform values is computed and the time-frequency representation is resized to a dimension of 64×64 which forms the input to the CNN. Resizing is carried out using interpolation, a commonly used technique in digital image processing. While various interpolation kernels are available for this purpose, bicubic interpolation has shown to be quite robust in time-frequency image resizing [13]. The interpolated surface using bicubic interpolation is given as

$$R(x,y)=\sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^{i}y^{j} \qquad (2)$$

which requires the computation of 16 coefficients a_{ij}. The interpolation can be computed by applying a convolution using the following kernel in both dimensions [14]

$$k(x)=\begin{cases}\frac{3}{2}|x|^3-\frac{5}{2}|x|^2+1, & |x|\le 1\\ -\frac{1}{2}|x|^3+\frac{5}{2}|x|^2-4|x|+2, & 1<|x|\le 2\\ 0, & \textit{otherwise.}\end{cases} \qquad (3)$$

A plot of time-domain signal for the spoken digit *zero* and its spectrogram and scalogram representations are given in Fig. 3.

C. Convolutional Neural Network

An overview of the CNN architecture used in this work is provided in Fig. 2 [9]. The network consists of five convolution layers. The filter size in each convolution layer is 3×3 with 12 filters in the first convolution layer, 24 filters in the second convolution layer, and 48 in the remaining three layers. Each convolutional layer is followed by a batch normalization layer [15] and rectified linear unit (ReLU) [16]. The inclusion of the batch normalization layer, in particular, was seen to improve the classification results. A max pooling layer [17] of size 3×3 and stride 2×2 is included after the ReLU layers, except the fourth. The final pooling layer is followed by a fully connected layer, softmax layer [18], and classification layer.

Adaptive moment estimation [19] is used for training the network with an initial learn rate of 0.0003, L2 regularization of 0.05, mini batch size of 128, and a maximum of 25 epochs. These parameters were optimized based on the training and validation performance.

D. Late Fusion

In addition to wavelet scalogram, spectrogram, smoothed-spectrogram, and mel-spectrogram time-frequency representations are also experimented with. These representations utilize different center frequencies and bandwidth and, therefore, may capture slightly different spectral information in the signal. Combining information from these representations may help improve the classification performance. In this work, late fusion is employed for this purpose whereby the output scores of the

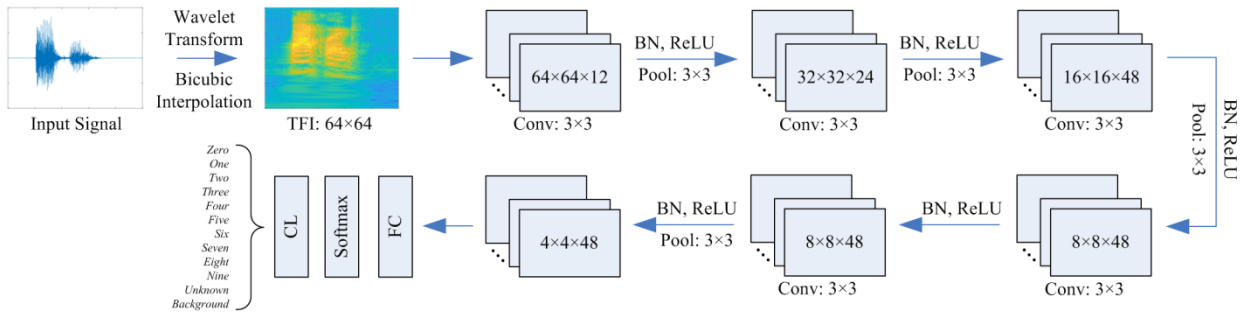

Fig. 2. An overview of the CNN architecture used in this work.

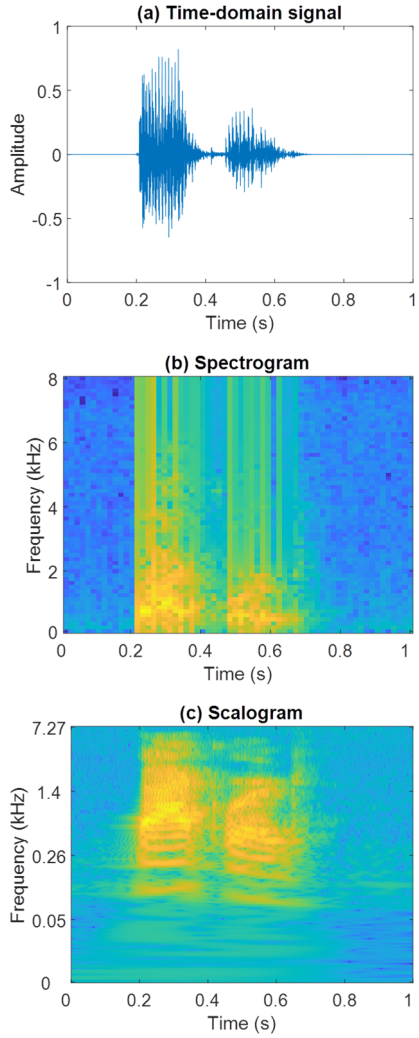

Fig. 3. (a) Time-domain signal for the spoken word "zero" and its time-frequency representations, (b) spectrogram representation using short-time Fourier transform and (c) scalogram representation using wavelet transform (bump wavelet).

best performing representations are averaged before predicting the class label.

III. Experimental Evaluation

The performance of the scalogram-CNN approach is evaluated against the spectrogram, smoothed-spectrogram, and mel-spectrogram representations, the computational details of which can be found in [8, 9]. The target time-frequency representation size in all cases is 64×64.

TABLE I. Overall classification error for spoken digit, non-digit, and background noise recognition using different time-frequency representations of size 64×64 and CNN.

Time-frequency representation	Validation Error (%)	Test Error (%)
Spectrogram	6.02	6.88
Smoothed-spectrogram	4.16	4.23
Mel-spectrogram	3.93	3.96
Scalogram (Morlet)	3.90	3.84
Scalogram (Morse)	3.86	3.60
Scalogram (Bump)	**3.52**	**3.52**
Late fusion of CNN output: smoothed-spectrogram + mel-spectrogram + scalogram (bump)	**2.85**	**2.84**

The overall validation and test results in spoken digit, non-digit, and background noise recognition are given in Table I. A validation and test error of 6.02% and 6.88% is achieved using the conventional spectrogram image. The results improve with the use of smoothed-spectrogram and mel-spectrogram representations. With the smoothed spectrogram, the validation and test error improve to 4.16% and 4.23%, respectively. Similarly, the validation and test error improve to 3.93% and 3.96% with the mel-spectrogram representation.

However, the best results using individual representations are achieved using the wavelet scalograms. While all three scalograms give reduction in the error rate compared to the baseline time-frequency representations, with a validation error of 3.52% and a test error of 3.52%, the best results are achieved using bump wavelet. The performance is further improved using late fusion. Here, the CNN output from the smoothed-spectrogram, mel-spectrogram, and wavelet scalogram (bump) representations are considered as they reveal spectral information using different techniques. The outputs of these three methods are averaged before making the final prediction. Late fusion achieves a validation error of 2.85% and a test error of 2.84%.

The confusion matrix for the test results using the late fusion approach are given in Table II. Background sounds, silence and noise, is the best performing class with 0% error and the least number of misclassifications are into this class, only 0.22% misclassification from class five. All the spoken digits, except one, four, and nine, have an error

TABLE II. CONFUSION MATRIX FOR TEST RESULTS USING LATE FUSION.

	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Unknown	Background	Error (%)
Zero	97.61			0.24			0.24	0.72			1.20		2.39
One		95.24			0.25					1.25	3.26		4.76
Two	0.71		97.41						0.24		1.65		2.59
Three			1.48	97.28					0.25		0.99		2.72
Four				0.25	95.00						4.75		5.00
Five						97.75			0.22		1.80	0.22	2.25
Six							98.22		0.51	0.25	1.02		1.78
Seven	0.25						0.99	98.03			0.74		1.97
Eight			0.25			0.25			96.08	0.25	3.19		3.92
Nine	0.25					0.98				95.59	3.19		4.41
Unknown	0.07	0.50	0.14	0.28	0.71	0.50		0.07	0.21	0.21	97.30		2.70
Background												100.00	0.00
Overall													2.84

TABLE III. COMPARISON OF OUR SPOKEN DIGIT RECOGNITION METHOD AND RESULTS AGAINST SOME EARLIER WORKS.

Method	Dataset	Classification Performance
GTCC and HTM [3]	Dataset: TIDIGITS Corpus Training samples = 2,412 Test samples = 1,144 (Spoken digit data only)	Error = 8.57%
MFCC, PCA, and SVM [4]	OGI Multilanguage Corpus Training samples = 826 Test samples = 454 (Spoken digit data only)	Error = 5.10%
Spectrogram and CNN [5]	AudioMNIST Dataset Training/test samples = 30,000 (Spoken digit data only)	Cross-validation error = 4.18%
This work: late fusion of CNN outputs from smoothed-spectrogram, mel-spectrogram, and wavelet scalogram	Speech Commands Dataset Training samples = 45,013 Validation samples = 5,363 Test samples = 5,914 (Dataset has spoken digit data, non-digit data, and background silence and noise)	Validation error = 2.85% Test error = 2.84%

rate of less than 4%. Spoken digit four has the highest error rate of 5.00%. Most misclassifications, 4.75%, are into the class unknown. While the error rate of class unknown is also less than 4%, the highest number of digit misclassifications is into this class.

IV. DISCUSSION AND CONCLUSION

Of the different time-frequency representations used to recognize digit, non-digit, and background signals using CNN in this work, the wavelet scalogram representations produced the smallest error. Wavelet transform offers good frequency localization in the lower frequency range. Since the audio signals considered here have more spectral information in the lower frequency range than the upper frequency range, as shown in Fig. 3, the wavelet scalogram representations help reveal more spectral information which could explain its superior performance. Furthermore, the bump wavelet offers wider variance in time and narrower variance in frequency which performed slightly better than the other two analytic wavelets considered in this work. Further analysis of the results shows that most of the misclassification is into the class unknown. This class comprises of several non-digit classes which could explain the high number of misclassifications into this class.

Table III summarizes the performance of this work along with some related works. It can be deduced that the strength of this work are the comprehensiveness of the dataset and the robustness of the classification performance, in particular the scalogram-CNN approach and late fusion. Also, the proposed method doesn't require significant feature engineering, relying on CNNs ability to learn distinguishing characteristics directly from the time-frequency representations.

REFERENCES

[1] L. R. Rabiner, "Applications of speech recognition in the area of telecommunications," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, USA, 1997, pp. 501-510.

[2] P. Salmela, M. Lehtokangas, and J. Saarinen, "Neural network based digit recognition system for voice dialling in noisy environments," *Information Sciences,* vol. 121, no. 3, pp. 171-199, 1999.

[3] J. v. Doremalen and L. Boves, "Spoken digit recognition using a hierarchical temporal memory," in *9th Annual Conference of the International Speech Communication Association (INTERSPEECH-2008)*, Brisbane, Australia, 2008, pp. 2566-2569.

[4] I. Bazzi and D. Katabi, "Using support vector machines for spoken digit recognition," in *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, 2000, pp. 433-436.

[5] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *arXiv preprint arXiv:1807.03418,* 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097-1105.

[7] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 22, no. 10, pp. 1533-1545, 2014.

[8] R. V. Sharan and T. J. Moir, "Acoustic event recognition using cochleagram image and convolutional neural networks," *Applied Acoustics,* vol. 148, pp. 62-66, 2019.

[9] R. V. Sharan, S. Berkovsky, and S. Liu, "Voice command recognition using biologically inspired time-frequency representation and convolutional neural networks," in *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Montreal, QC, Canada, 2020, pp. 998-1001.

[10] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209,* 2018.

[11] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way.* United States: Academic Press, 2009.

[12] I. Daubechies, *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 1992.

[13] R. V. Sharan and T. J. Moir, "Time-frequency image resizing using interpolation for acoustic event recognition with convolutional neural networks," in *IEEE International Conference on Signals and Systems (ICSigSys)*, Bandung, Indonesia, 2019, pp. 8-11.

[14] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 29, no. 6, pp. 1153-1160, 1981.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167,* 2015.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 807-814.

[17] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *IEEE 12th International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 2146-2153.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer, 2006.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.