Voice Command Recognition Using Biologically Inspired Time-Frequency Representation and Convolutional Neural Networks

Roneel V. Sharan, *Senior Member, IEEE*, Shlomo Berkovsky, and Sidong Liu

***Abstract*—Voice command is an important interface between human and technology in healthcare, such as for hands-free control of surgical robots and in patient care technology. Voice command recognition can be cast as a speech classification task, where convolutional neural networks (CNNs) have demonstrated strong performance. CNN is originally an image classification technique and time-frequency representation of speech signals is the most commonly used image-like representation for CNNs. Various types of time-frequency representations are commonly used for this purpose. This work investigates the use of cochleagram, utilizing a gammatone filter which models the frequency selectivity of the human cochlea, as the time-frequency representation of voice commands and input for the CNN classifier. We also explore multi-view CNN as a technique for combining learning from different time-frequency representations. The proposed method is evaluated on a large dataset and shown to achieve high classification accuracy.**

I. Introduction

Natural user interfaces and voice interaction have been deployed in recent years in many applications, including healthcare technology. Voice interaction is of a particular importance in technologies designed for the elderly or people with disabilities, as it simplifies the interaction and flattens the learning curve of the technology.

Examples of voice interaction incorporated healthcare technologies include social assistive robots for health and psychological well-being of the elderly [1], robots and home automation technology for safety and wellbeing [2, 3], and electric wheelchairs [4] and myoelectric prostheses [5]. Another practical use case of voice command technology is situations in which users cannot control the technology using their hands or gestures, like voice controlled robotics during surgery that has been shown to save surgeon's time [6].

One of the pivotal components of such voice interfaces is speech recognition, as this facilitates the user-to-technology interaction channel. Typically, speech recognition is implemented by training cloud-based machine learning technologies on large datasets of spoken language. However, simple voice command technologies operating in a closed domain, e.g., wheelchair control or robotic surgery assistance, often need to recognize only a limited vocabulary. Furthermore, for practical reasons associated with streaming audio signal to the cloud, intermittent connection, high latency, and increased security risks, it would be advantageous to perform the speech recognition locally.

R. V. Sharan, S. Berkovsky, and S. Liu are with the Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia (e-mail: roneel.sharan@mq.edu.au, shlomo.berkovsky@mq.edu.au, and sidong.liu@mq.edu.au).

In a closed domain, voice command recognition can be cast as a speech classification task, where a range of feature extraction, machine learning, and deep learning methods can be deployed. In particular, convolutional neural network (CNN), an established image classification technique, has yielded promising results in applications such as speech [7] and acoustic event recognition [8].

The use of CNN for voice command recognition, however, requires the audio to be converted into an image-like representation, which is typically accomplished in audio signal processing through a time-frequency representation. Various time-frequency representations have been studied. This includes the conventional spectrogram representation and two other common approaches using frequency domain filterbanks. These are the moving average filters and mel-filters with the corresponding time-frequency representations referred to as smoothed-spectrogram and mel-spectrogram, respectively.

In this work, we propose to use a filterbank inspired by the human auditory model. Specifically, we utilize the gammatone filter, which models the frequency selectivity of the human cochlea [9]. The resulting time-frequency representation is called a *cochleagram*. Since the cochleagram representation differs from the smoothed- and mel-spectrograms in the spectral information at different frequencies and bandwidth, it remains unclear what representation leads to the best classification of a CNN for a closed domain vocabulary.

Furthermore, feature data combination has long been a technique for improving classification performance in various classification tasks. With CNN, this has been achieved by representing multiple images as channels in a single CNN. However, the use of a multi-view CNN [10] has shown to improve the robustness of 3-D shape recognition. This refers to a technique where learning from multiple CNNs trained on different 2-D images, which give multi-view representation of the object, is combined to recognize a 3-D shape.

Inspired by the multi-view CNN technique, this work proposes combining the CNN learning from different time-frequency representations of an audio signal in voice command recognition. The smoothed-spectrogram, mel-spectrogram, and cochleagram are considered for this purpose. Each of these time-frequency representations captures spectral information at different center frequencies and bandwidth. We hypothesize that combining the learning from these three representations would to lead to improved classification accuracy.

We report of evaluation using a dataset of more than 56,000 audio segments, including ten voice commands, a combined class, and background noise [11].

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

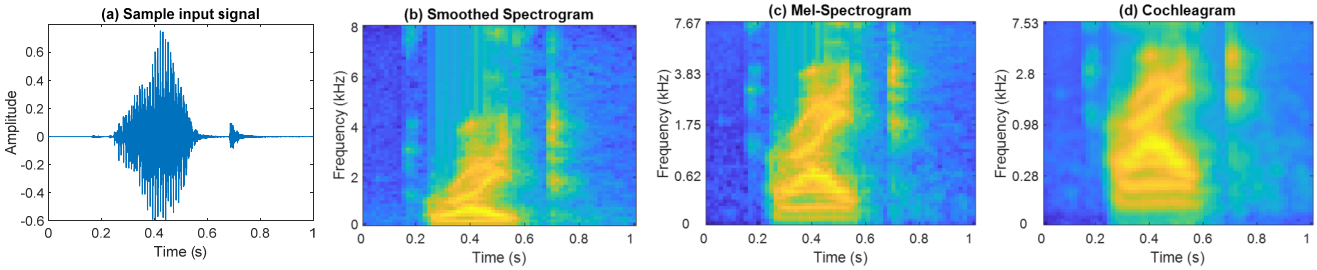

Figure 1. (a) Time plot of the voice command right, and three time-frequency representations: (b) smoothed-spectrogram, (c) mel-spectrogram, and (d) cochleagram.

The obtained results show that the cochleagram outperforms the baseline time-frequency representations in terms of the classification accuracy, achieving error rates of about 3.5%. Finally, we evaluate the performance of a multi-view CNN, which further improves the classification accuracy. In summary, our work demonstrates combination of audio signal processing and deep learning methods – namely, of the cochleagram representation and CNNs – that yields highly accurate command recognition with potential application in voice interfaces.

II. Methods

A. Dataset

This work utilizes the Speech Commands dataset [11] which has a total of 105,829 utterances across 35 classes, pronounced by 2,618 speakers. In this work, we used 38,546 voice commands across ten classes (*yes*, *no*, *up*, *down*, *left*, *right*, *on*, *off*, *stop*, *go*) and additional *other* class of words sampled from 25 classes (backward, bed, bird, cat, dog, eight, five, follow, forward, four, happy, house, learn, marvin, nine, one, seven, sheila, six, three, tree, two, visual, wow, zero). In addition, the dataset contains long background noise segments, which were used to generate 4,000 samples for the *noise* class. The audio segments used were one second long.

The final dataset contains a total of 56,003 audio segments. For the ten command and *other* classes, we used the validation and test data as per the dataset guidelines, while the noise class was randomly split into the training, validation, and test datasets using the 80%-10%-10% ratio. As such, the final dataset included 44,736 training segments, 5,401 validation segments, and 5,866 test segments. A sample time-domain command *right* is shown in Fig. 1(a).

B. Image-Like Representations

Analysis of audio signals is primarily carried out in time and/or frequency domains. While time-frequency analysis of audio signals is more popular, we considered an image-like representation of the raw time-domain signal using framing and interpolation techniques. We considered the conventional spectrogram image, as well as the commonly used smoothed- and mel-spectrogram representations, and the proposed cochleagram representation. The target image representation size of 64×64 was used in this work.

Time-Domain: To produce an image-like representation, the time-domain signal is divided into 64 equal parts and bicubic interpolation [12] is applied to resize it to a 64×64 representation.

Spectrogram: To form the spectrogram, the audio signal is divided into 64 frames with a 50% overlap. Discrete Fourier Transform (DFT) is performed using 128 points resulting in the 64×64 spectrogram image.

Smoothed-spectrogram: Similar to spectrogram, but DFT is performed using 1024 points resulting in a 512×64 spectrogram image. Then 64 non-overlapping moving average filters are applied along the frequency axis resulting in the 64×64 smoothed-spectrogram.

Mel-spectrogram: Utilizes the mel-filter used for computing the mel-frequency cepstral coefficients [13], while the outputs of mel-filterbank form the mel-spectrogram. Procedure similar to the smoothed-spectrogram is used and 64 mel-filters are applied to a 512×64 spectrogram to obtain the 64×64 mel-spectrogram.

Cochleagram: In contrast, the frequency components of the cochleagram time-frequency representation are modeled by a gammatone filter [9]. The bandwidth of the filter is determined using equivalent rectangular bandwidth (ERB), a psychoacoustic measure of the auditory filter width at each point along the cochlea [9]. The ERB filter model of [14] is used and the filter is implemented using Matlab's Auditory toolbox [15]. To form the cochleagram, the number of gammatone filters is set to 64 and the filtered signal is divided into 64 frames with a 50% overlap. The cochleagram is obtained by adding the energy in each frame.

An illustration of smoothed-spectrogram, mel-spectrogram, and cochleagram image representations for a sample voice command *right* is shown in Fig. 1(b)-(d). Each image is of size 64×64, with a frequency range of 0 – 8000 Hz, but with different center frequencies and bandwidth.

C. Convolutional Neural Network

The CNN is trained using adaptive moment estimation (Adam) [16]. The network contains five convolution layers, each of which includes a batch normalization layer [17] and a rectified linear unit (ReLU) [18]. The filter size in each convolution layer is 3×3. The number of filters in the first layer is 48, in the second layer – 96, and in the remaining three layers – 192. All the ReLU layers, except for the fourth, are followed by a max pooling layer [19]. The size of a max pooling layer is 3×3 and of the stride – 2×2. These are followed by a fully connected layer, a softmax layer [20], and an output layer.

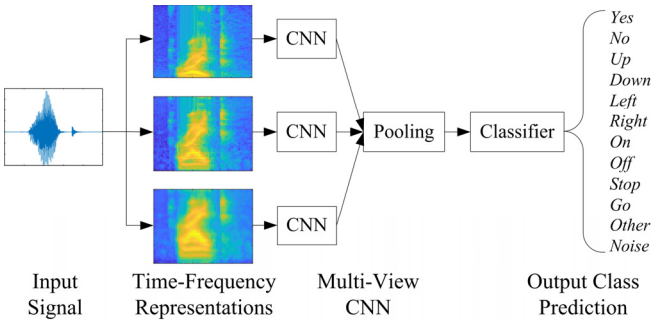

Figure 2. Multi-view CNN architecture.

The other parameters of the CNN are set as follows: initial learn rate = 0.0003, learn rate schedule = piece-wise, learn rate drop factor = 0.1, learn rate drop period = 20, L2 regularization = 0.05, mini batch size = 128, data shuffle = every-epoch, and max epochs = 25. These parameter values were optimized based on performance on the validation dataset. The training stops after the maximum number of epochs is reached.

Fig. 2 overviews the proposed multi-view CNN architecture. Each of the individual time-frequency representations (smoothed- and mel-spectrogram, and cochleagram) is trained using a single CNN. The three CNN outputs are pooled together to train a secondary classifier predicting the output class label.

III. Results

A. Individual Time-Frequency Representations

In this experiment, individual CNNs are trained on each type of image representation from all the 12 classes (ten voice command, *other*, and *noise*). The overall classification error obtained for the validation and test sets, and the evaluated image representations is shown in Table I. The highest classification error rates of 16.65% and 18.09% are produced by the time-domain image representation of the transformed audio signal. The accuracy is improved using the time-frequency image representations. Specifically, the error rates of spectrogram hover around the 6.7%-6.9% mark and smoothed- and mel-spectrogram representations achieve even lower error rates in the 3.75%-4.35% range. These results are further improved with the cochleagram representation, where the validation and test error rates were, respectively 3.65% and 3.39%. Hence, cochleagram fed into a CNN produces highly accurate voice command recognition.

Next, we turn to the classifications of individual voice commands and analyze the confusion matrix produced by the CNN using the cochleagram representation (see Table II). The rows represent the actual commands and the columns – the output of the classifier. With such a low error rate most commands are classified correctly, so that the diagonal values of the confusion matrix are 92.79% or greater. In particular, *yes*, *left*, *stop* and *other* achieve classification accuracy greater than 97% and *noise* is always classified correctly.

We observe a majority of misclassifications stemming from commands being misclassified as *other*. This can be explained by the composition of the *other* class, which contains words from 25 classes. With such a diverse data, some command pronunciations may be similar to training instances of *other*, which causes the misclassification. We also analyze the more pronounced combinations with misclassification rates greater than 1%. Here we observe two distinct groups. The first contains phonetically similar commands (e.g., *down-no*) that the classifier struggles to differentiate. The second contains short commands (e.g., *off-up* and *go-no*) that presumably yield similar cochleagram representations and also cause misclassifications.

TABLE I. Overall classification error across the 12 classes using time-domain and time-frequency representations and CNN

Image Representation	Validation Error (%)	Test Error (%)
Time-domain	16.65	18.09
Spectrogram	6.74	6.89
Smoothed-Spectrogram	4.35	3.94
Mel-Spectrogram	3.87	3.75
Cochleagram	3.65	3.39

TABLE II. Confusion matrix (values in %) for the cochleagram representation and test data set

	Yes	No	Up	Down	Left	Right	On	Off	Stop	Go	Other	Noise
Yes	99.28										0.72	
No	0.25	96.54		0.49	0.25					0.25	2.22	
Up			95.53				0.24	0.71	0.47	0.24	2.82	
Down		1.48		93.35						1.48	3.69	
Left	0.24		0.24		97.33			0.24			1.94	
Right			0.25		0.25	95.71					3.79	
On			0.76	0.51			94.44	0.51			3.54	0.25
Off			1.99				0.50	95.52			1.99	
Stop				0.24					98.78		0.97	
Go		2.74	0.50	0.25	0.25					92.79	3.48	
Other		0.36	0.22	0.29	0.29	0.22	0.36	0.14	0.07	0.50	97.56	
Noise												100

B. Multi-View CNN

Finally, we present the results obtained with a multi-view CNN, with various secondary classifiers (as per Fig. 2). We also present results using the conventional approach where the three time-frequency representations are treated as different channels on which a single multi-channel CNN is trained. For the latter, the classification results improve marginally, with an overall validation and test errors of 3.57% and 3.20%, respectively (see Table III).

For the multi-view CNN, the CNN outputs from the individual time-frequency representations are concatenated and a secondary classifier is trained to classify the audio segments. Four methods are exploited by the secondary classifier predicting the output: K-nearest neighbors (KNN), logistic regression (LR), softmax layer (SMX), and support vector machine (SVM). Classification error values using the multi-view CNN are given in Table II. As can be seen, the use of the smoothed- and mel-spectrogram representations improved the accuracy of the best-performing cochleagram. The best results were achieved by a multi-view CNN using a secondary SVM classifier, where the error for the validation and test sets was 2.93% and 2.90%, respectively.

TABLE III. OVERALL CLASSIFICATION ERROR USING MULTI-CHANNEL AND MULTI-VIEW CNN

Classification Method	Validation Error (%)	Test Error (%)
Multi-channel CNN	3.57	3.20
Multi-view CNN (KNN)	3.35	3.09
Multi-view CNN (LR)	3.11	2.98
Multi-view CNN (SMX)	3.31	3.20
Multi-view CNN (SVM)	2.93	2.90

IV. DISCUSSION AND CONCLUSION

This work focused on the recognition and classification of voice commands, which is pivotal for natural interaction and voice interfaces in a range of healthcare technologies.

In order to deploy individual and multi-view CNNs for recognition of voice command, we experimented with several time-frequency representations that converted the command audio signal into images. The cochleagram, a unique representation modeling the frequency selectivity of the human cochlea, was found to achieve the best classification accuracy. Cochleagram utilizes a gammatone filter that offers more frequency components in the lower frequency range than in the upper range. Since the considered audio segments are human voice commands with more spectral information in the lower frequency range (see Fig. 1(d)), the cochleagram representation reveals more information, which could explain its strong performance.

We analyzed the recognition errors of the cochleagram representations and discovered misclassifications primarily in commands that are either phonetically similar or too short and yield similar cochleagrams. To improve the recognition, we deployed a multi-view CNN that combined the cochleagram with two other time-frequency representations. Collectively, they covered a broader range of frequency components to convey more information. As a result, the classification accuracy improved and the lowest observed error rate was below 3%, indicating high-quality closed domain voice command recognition. Hence, our work provides an important contribution: it combines deep learning methods typically using images with signal processing methods converting voice commands into time-frequency image representations.

However, it should be mentioned that our evaluation involved a small vocabulary of commands. Thus, it is yet to be ascertained how the proposed cochleagram representation performs in less constrained and ecologically valid applications. We intend to evaluate its performance in simulations of a closed domain, e.g., robotic surgery assistant.

We are also interested to study personalized voice recognition applications. As different users have different pronunciations and accents, it is increasingly important to be able to adjust the voice recognition engines to the specific speaker. In this work, we trained accurate command recognition CNNs with relatively small data sets. This brings to the fore the opportunity to personalize voice recognition or alternatively tailor it to specific accents. We also plan to study the feasibility of this in the future.

REFERENCES

[1] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: A review," *Gerontechnology,* vol. 8, no. 2, pp. 94-103, 2009.

[2] T. Mukai, S. Hirano, H. Nakashima, Y. Kato, Y. Sakaida, S. Guo*, et al.*, "Development of a nursing-care assistant robot RIBA that can lift a human in its arms," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, 2010, pp. 5996-6001.

[3] M. Vacher, P. Chahuara, B. Lecouteux, D. Istrate, F. Portet, T. Joubert*, et al.*, "The sweet-home project: Audio processing and decision making in smart home to improve well-being and reliance," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, 2013, pp. 7298-7301.

[4] M. Rojas, P. Ponce, and A. Molina, "Skills based evaluation of alternative input methods to command a semi-autonomous electric wheelchair," in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, 2016, pp. 4593-4596.

[5] P. Fang, Z. Wei, Y. Geng, F. Yao, and G. Li, "Using speech for mode selection in control of multifunctional myoelectric prostheses," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, 2013, pp. 3602-3605.

[6] P. Kania and M. Salagierski, "Preliminary results of using a voice-controlled robotic camera driver during 3D laparoscopic radical prostatectomy," *Central European Journal of Urology,* vol. 71, no. 4, pp. 394-398, 2018.

[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. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, K. Horner, and L. Demany, Eds. Pergamon, 1992, pp. 429-446.

[10] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 945-953.

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

[12] 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.

[13] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 28, no. 4, pp. 357-366, 1980.

[14] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *Journal of the Acoustical Society of America* vol. 87, no. 6, pp. 2592-2605, 1990.

[15] M. Slaney, "Auditory Toolbox for Matlab," Interval Research Corporation, Technical Report 1998-010, 1998.

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

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

[18] 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.

[19] 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.

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