Improving Keywords Spotting Performance in Noise with Augmented Dataset from Vocoded Speech

Ruohao Li* and Kaibao Nie†
*University of Washington, Bothell, USA
E-mail: ruohaoli@uw.edu
†University of Washington, Bothell, USA
E-mail: niek@uw.edu

Abstract— **While more and more electronic devices have an on-device speech recognition system, producing and deploying trained models for keyword(s) detection is becoming more and more demanding. The dataset preparation is one of the most challenging and tedious tasks in Keywords Spotting (KWS) since it requires a significant amount of time obtaining raw or segmented audio speeches. In this paper, we proposed a data augmentation strategy using a speech vocoder to artificially generate vocoded speech at different numbers of channels. A trained KWS system was first tested with vocoded speech and it showed consistent performance with studies from human subjects listening to vocoded speeches. Furthermore, the KWS system trained with the augmented dataset showed promising improvement evaluated at +10 dB SNR.**

I. INTRODUCTION

With the rapid development of deep-learning-based speech recognition technologies, speech-interacted electronics are becoming increasingly popular, for example, Amazon Echo, Google Home, and Apple Home. The personal assistants behind those devices such as Amazon's Alexa, Google Now and Apple's Siri are all utilizing speech recognition to interact with users. Such a KWS system aims to detect a pre-defined keyword or a set of keywords in a live stream of audio to enable hands-free experience to users. Also, a KWS system can be potentially implemented in hearing devices such as hearing aids and cochlear implants to allow voice-activated adjustment of map parameters, e.g. volume and sensitivity settings. However, for practical problems, the creation of a dataset for the KWS system is long and tedious. In order to have high recognition performance with desired keyword(s), the use of large datasets is required since it helps in preventing overfitting in complex prediction methods. Moreover, in the case of recognizing keyword(s) in noisy environments, a fair amount of noisy keyword(s) should be included in the datasets to introduce different features for classification, which further increases the time when preparing datasets. The data augmentation technique has the potential to artificially increase the size of the dataset. The early work from Chang et al, [1] used speaker modification techniques [2] in a speaker-independent keyword spotting task by increasing the size of the training set. More recent works [3-6] also used data augmentation for Automatic Speech Recognition. These works mainly take advantage of special audio datasets transformations for speeches such as vocal tract length perturbation. Also, data augmentation has been used to increase the quantity of training data, avoid overfitting and increase the robustness of models [7-8], while Raju et. al, [9] demonstrated improved performance in the living room with ambient noise from household appliances when mixing training datasets with music and TV/movie audio.

In this paper, we created a speech vocoder to generate vocoded speech in dataset augmentation with adjustable spectral resolutions (different number of channels) that will give different levels of distortion. The smaller number of channels, the more distortion will be introduced in vocoded speeches. We built the KWS system by adapting a Convolutional Neural Network (CNN) from Cui et al. [6], and trained it with Google's Speech Commands dataset [16] with selected 12 classes (10 keywords + background noise + unknown), and verified that it has high accuracy when classifying those 12 classes (best performance is ~95%). Then, we used the trained CNN to classify vocoded speeches at different numbers of channels varying from 1 – 32. The goal was to investigate how a trained KWS system can recognize vocoded speech. Data from human subjects [11-14] have shown that and 4-8 channels are sufficient for human subjects to reach a high level of performance in recognizing sentences or phonemes with vocoded speech. Another experiment we performed is mixing vocoded speeches with original speeches from Google's Speech Commands dataset, creating an augmented dataset, and then trained it with the same CNN used in this paper. We hypothesize that training with the augmented dataset can potentially improve its performance in classifying noisy keywords (SNR = 10dB) since vocoded speeches introduce new acoustic features from distorted speeches.

II. MODEL IMPLEMENTATION

The block diagram of the KWS system used in this paper is shown in Figure 1. First, audio samples will be processed by a vocoder to generate vocoded speeches. We mixed audio samples and vocoded speeches with a ratio of 1:1 to form an augmented training set. Then, in the feature extraction module, the 50-dimensional Log-Mel filter bank energies (LFBE) features were computed every 30 ms with a

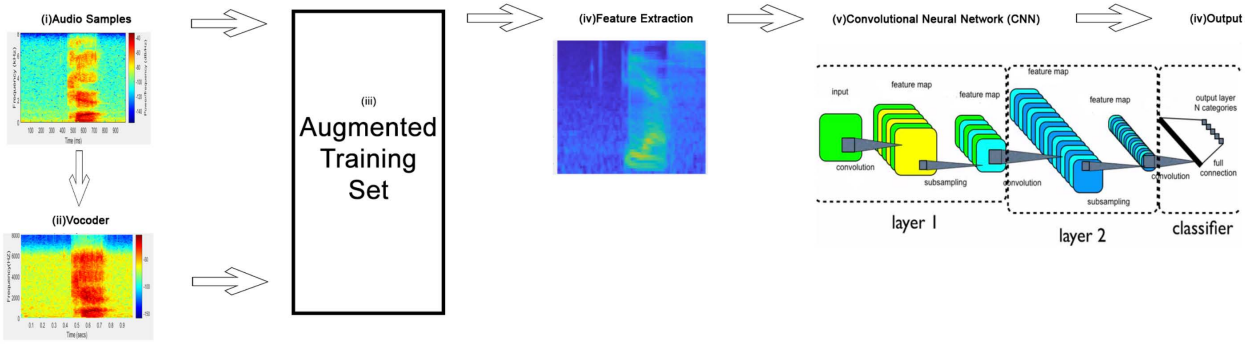

Figure 1: Block diagram of the KWS system. From left to right: (i) Input audio samples and the spectrogram of a keyword "cat". (ii) the spectrogram of a vocoded keyword "cat". (iii) Augmented training set contains both original audio samples and vocoded speeches with a ratio of 1:1, e.g. if the total number of input audio aamples is 100, then the augmented training set will have 200 samples in total, with 100 samples from the vocoded speeches. (iv) Feature extraction from a keyword "cat". (v) Convolutional Neural Network (CNN). (vi) Output probability for classfication.

10 ms frameshift at 512 FFT length. The extracted speech feature matrix was fed into a Convolutional Neural Network (CNN) which is adapted from [10]. The classifier can generate the probabilities for the output classes in training or validation.

A. Speech Vocoder

The Speech Vocoder we used in this paper is inspired by the acoustic simulation algorithms for cochlear implants as described in [15], which are commonly used to simulate the sound perceived by cochlear implant subjects. The vocoder has five major steps to process speech signals, as shown in Figure 2. The first step is to use a number of band-pass filters (BPF) (N >= 1) to filter the input signal with logarithmically equal frequency bands. Since the frequency of human voice is primarily under 5 kHz, a low-pass filter is commonly applied before BPF. In this paper, the frequency range for vocoding is from 80 Hz to 6 kHz. The second step is

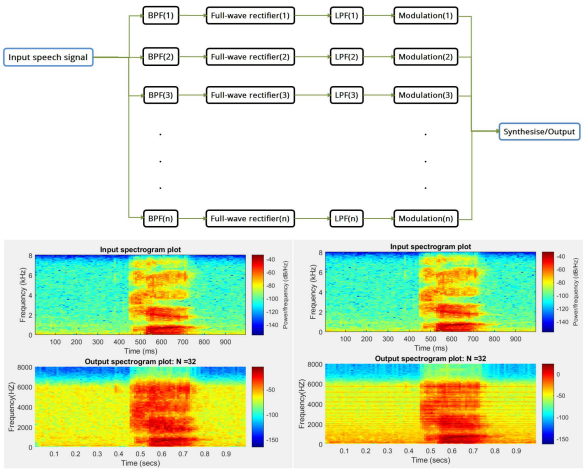

Figure 2 shows the flowchart of an n-channel Speech Vocoder. The spectrogram of "cat" with white noise modulation (left), the spectrogram of "cat" with sine modulation (right).The spectrogram of "cat" with white noise modulation has smoother transitions.

to apply a full-wave rectifier in each channel after the BPF to extract temporal envelopes from each channel. The third step is to apply a low-pass filter (LPF) in each channel after the full-wave rectifier to filter out temporal fine structures higher than the upper limit of envelope frequency. Two different modulations were used to synthesize vocoded sounds, a sine-wave modulation that produces robotic sound and a white noise modulation that produces more human-like sounds. The last step is to sum all signals from n channels after modulation. All vocoded sounds were further normalized by calculating their RMS (Root-Mean-Square) which is required for the generated vocoded speeches because the amplitude of the vocoded speeches typically dropped by around 60% compared to the input speech signal.

B. CNN Architecture

In Figure 3, we adapted the CNN architectures from [10] with increased convolutional layers to five to deal with larger feature sets. After feature extraction, the KWS system produces 4900 (50x98) feature matrixes for one-second of audio, while [10] only has 1960 (49×40) features. Each convolutional layer is followed by a batch normalization layer, a ReLU layer, and a max-pooling layer to reduce feature maps. By adding a final global max-pooling layer, the number of parameters required in the fully-connected layer will be significantly reduced. To prevent the network from overfitting, a 0.2 of dropout was added to the input to the last fully-connected layer.

III. EXPERIMENTS

A. Training on the Speech Commands Dataset

We trained and evaluated our model using Google's Speech Commands dataset [16], an established dataset for benchmarking KWS systems. The dataset consists of 65k one-second-long speech audio of 30 different keywords spoken by 1881 different speakers. The CNN proposed in this paper is trained to classify the following 12 keywords: "yes", "no", "up", "down", "left", "right", "on", "off", "one", "two", along with unknown and background noise. Background noise data is generated from the background noise catagory inside the

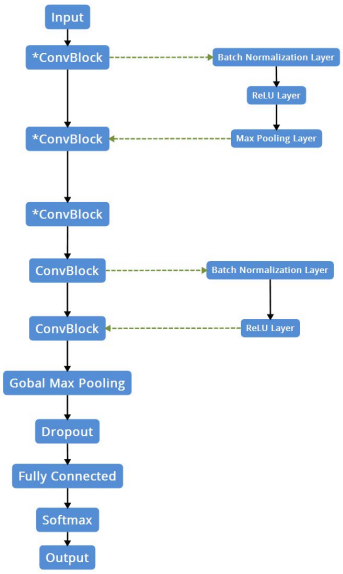

Figure 3. Overall CNN architecture used in this paper. Each *ConvBlock is followed by a batch normalization layer, a ReLU layer, and a max-pooling layer. Each ConvBlock is followed by a batch normalization layer and a ReLU layer. Global max pooling is followed by the second ConvBlock.

dataset and segmented into one-second-long audios, thus 4500 background noise data samples will be generated. The remaining 20 keywords from the dataset are labeled as "unknown". However, there are many more samples for the "unknown" that leads to an imbalance in the dataset. Therefore, we only take 15% of the "Unknown" from the remaining 20 keywords. The original dataset is split into training, validation, and test sets in the ratio of 8:1:1, while making sure that the audio clips from the same person stay in the same set.

The KWS model is trained for 40 epochs with the Adam optimizer [17] with a mini-batch size of 128 and an initial learning rate of 1e-4 to reduce the learning rate by a factor of 10 after 30 epochs; the model with the highest validation accuracy is saved to evaluate accuracy on the test set.

B. Experimental Setup for Vocoded Speech Dataset

The Vocoded Speech Dataset is generated from the same Google's Speech Commands dataset [12] by using a speech vocoder with the same dataset split configuration as introduced previously. We generated six vocoded speech datasets at different numbers of channels: n = 1, 2, 4, 8, 16, 32. Since the characteristic of 2-channel vocoded speech is similar to the 1-channel vocoded speech, we dropped the 2-channel vocoded speech when evaluating the performance of vocoded speech with numbers of channels: n = 1, 4, 8, 16, 32 by training with the original dataset, and validating with vocoded speech with the proposed KWS. On the other hand, 1 and 2-channel vocoded speech are highly distorted sounds, so we dropped them when evaluating the performance of vocoded speech in noise with conditions at the numbers of channels: n = 4, 8, 16, 32 by training with 50% original datasetand 50% vocoded speech with various channels, and validating with SNR = 10 dB noise corrupted original dataset in the proposed CNN.

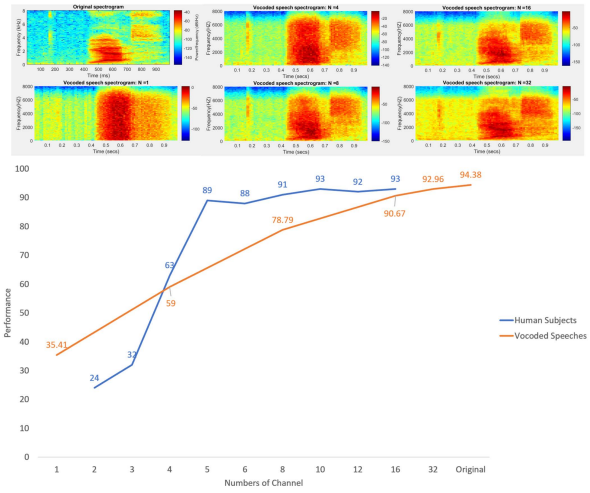

Figure 4. The top six spectrograms show the difference between original speech and various channels of vocoded speech from the keyword "yes". The orange curve indicates the recognition performance (correct percentage) of the trained CNN as the number of channels is increased from 1 to 32 tested with the vocoded speeches. The CNN was trained with the Original Dataset and validated with the vocoded speeches. The "Original" in the "Numbers of channel" means trained with the Original Dataset and validated with the Original Dataset. The blue curve indicates the the recognition performance of the human subjects with different channels from 2 to 16.

IV. RESULTS AND DISCUSSION

The base model of CNN using the original dataset in both training and validation achieved the desired accuracy of 94.38% on the Google's Speech Commands dataset. Figure 4 shows the performance of vocoded speech tested with the trained CNN (orange curve) and the sentence recognition performance from human subjects (blue curve) listening to vocoded sounds from a previous study [14], as the number of channels for vocoding is increased. It is worth to notice that the performance of KWS for vocoded speech is in a similar trend to sentence understanding performance from human subjects. Better recognition performance is achieved when the number of channels increases. The number of channels required to achieve reasonable performance is around 8 for the trained KWS system. However, the performance of sentence understanding from human subjects does not tend to change much from 5-channel to 16-channel. The performance at N = 32 shows that 32-channels vocoded speech is very similar to the original dataset. 8-32 channels of vocoded speech can be considered as augmented samples for training KWS or automatic speech recognition systems. The dataset augmentation method may have the potential to significantly expand the training dataset by multiple times.

Figure 5 summarizes the performance of vocoded speech in SNR = 10 dB noise corruption with numbers of

channels: n = 4, 8, 16, 32 as compared to how the original dataset performed in the same condition. The original dataset has lower performance compared to vocoded speech with the number of channels at n = 4, 8, 16, and 32. And also, the performance of KWS with vocoded speech improves when the number of channels in the vocoder increases.

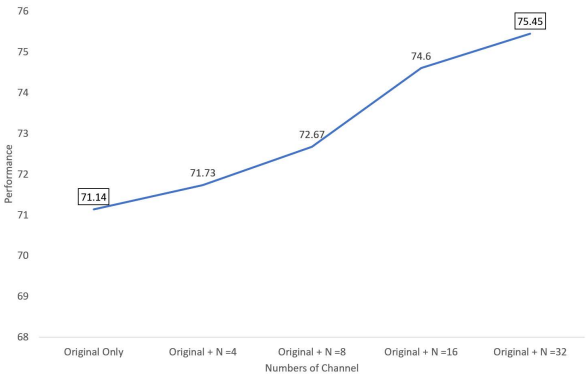

Figure 5. performance curve on noise corrupted audio speech at SNR=10 dB tested with the original data set and the augumented data sets. From let to right, "Original Only" means training with 100% original dataset, "Original + N =4" means trianing with augmented dataset (50% original dataset plus 50% 4-channel vocoded speech), and the same mixing strategy for N=8, N=16, and N=32.

V. CONCLUSIONS

In this paper, we demonstrated the possibility of using 16-channel and 32-channel vocoded speech to increase the size of training data in the KWS system. While the vocoded speech in the trained CNN performed similarly to human subjects, 32-channel vocoded speech with 92.96% accuracy has nearly the same performance as human subjects recognizing sentence in 16-channel (93%). Such finding suggests that vocoded speech has the potential for being used in training KWS system. On the other hand, the augmented data by mixing vocoded speech with the original dataset improves the KWS system in moderate noise (SNR = 10dB), resulting in an improvement of 4.3% by mixing the 32-channel vocoded speech with the original dataset. Training and testing with a larger set of keywords needs to be further studied to show how vocoded speech can add benefits to a real KWS system.

ACKNOWLEDGMENT

This study was supported by the MSEE graduate fund provided by the University of Washington-Bothell.

REFERENCES

[1] Chang, E. I., & Lippmann, R. P. (1995). Using voice transformations to create additional training talkers for word spotting. In Advances in Neural Information Processing Systems, pages 875–882.

[2] Quatieri, T. F., & McAulay, R. J. (1992). Shape invariant time-scale and pitch modification of speech. IEEE Transactions on Signal Processing, 40(3), 497–510. DOI: https://doi.org/10.1109/78.120793

[3] Jaitly, N., & Hinton, G. E. (2013). Vocal tract length perturbation (VTLP) improves speech recognition. In ICML Workshop on Deep Learning for Audio, Speech and Language, volume 117.

[4] Kanda, N., Takeda, R., & Obuchi, Y. (2013). Elastic spectral distortion for low resource speech recognition with deep neural networks. In IEEE Workshop on Automatic Speech Recognition and Understanding, pages 309–314. DOI: https://doi.org/10.1109/ASRU. 2013.6707748

[5] Ragni, A., Knill, K. M., Rath, S. P., & Gales, M. J. (2014). Data augmentation for low resource languages. In 15th Annual Conference of the Internationa

[6] Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(9), 1469–1477.

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in Proceedings of INTERSPEECH, 2015.

[8] R. Hsiao, J. Ma, W. Hartmann, M. Karafiat, F. Grezl, L. Burget, I. Szoke, J. H. Cernocky, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansky, S. Tsakalidis, and R. Schwartz, "Robust speech recognition in unknown reverberant and noisy conditions," in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 533–538.

[9] Anirudh Raju, Sankaran Panchapagesan, Xing Liu, Arindam Mandal, Nikko Strom, "Data Augmentation for Robust Keyword Spotting Under Playback Interference," arXiv preprint arXiv:1808.00563, 2018.

[10] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra, "Hello edge: Keyword spotting on microcontrollers," arXiv preprint arXiv:1711.07128, 2017.

[11] Michael F. Dorman & Phillip C. Loizou, "Speech Intelligibility as a Function of the Number of Channels of Stimulation for Normal-Hearing Listeners and Patients with Cochlear Implants," The American Journal of Otology, Vol 18, No.6(Suppl), 1997.

[12] Lendra M. Friesen, Robert V. Shannon, Deniz Baskent, Xiaosong Wang, "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," The Journal of the Acoustical Society of America 110, 1150 (2001); doi: 10.1121/1.1381538.

[13] Michael F. Dorman, Philipos C. Loizou, Dawne Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," The Journal of the Acoustical Society of America 102, 2403 (1997); doi: 10.1121/1.419603

[14] Philipos C. Loizou, Michael Dorman, Zhemin Tu, "On the number of channels needed to understand speech," The Journal of the Acoustical Society of America 106, 2097 (1999); doi: 10.1121/1.427954

[15] Charles T. M. Choi & Yi-Hsuan Lee (2012). A Review of Stimulating Strategies for Cochlear Implants, Cochlear Implant Research Updates, Dr. Cila Umat (Ed.), ISBN: 978-953-51-0582-4, InTech, Available from: http://www.intechopen.com/books/cochlear-implant-research-updates/stimulating-strategies-for-cochlearimplants

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

[17] J. Ba D. P. Kingma, "Adam: A method for stochastic optimization," in Conference on Learning Representations (ICLR), 2015.