*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Residual + Capsule Networks (ResCap) for Simultaneous Single-Channel Overlapped Keyword Recognition

*Yan Xiong, Visar Berisha, Chaitali Chakrabarti*

School of Electrical, Computer and Energy Engineering
Arizona State University, Tempe, Arizona, USA
{yxiong35,visar,chaitali}@asu.edu

## Abstract

Overlapped speech poses a significant problem in a variety of applications in speech processing including speaker identification, speaker diarization, and speech recognition among others. To address it, existing systems combine source separation with algorithms for processing non-overlapped speech (e.g. source separation + follow-on speech recognition). In this paper we propose a modified network architecture to simultaneously recognize keywords from overlapped speech without explicitly having to perform source separation. We build our network by adding capsule layers to a ResNet architecture that has shown state-of-the-art performance on a traditional keyword recognition task. We evaluate the model on a series of 10-word overlapped keyword recognition experiments, using speaker dependent and speaker independent training. Results indicate that Residual + Capsule (ResCap) network shows marked improvement in recognizing overlapped speech, especially in experiments where there is a mismatch in the number of overlapped speakers between the training set and the test set.

**Index Terms**: speech recognition, keyword spotting, recognition, overlapped speech, capsule networks, residual networks, ResNet

## 1. Introduction

Overlapped speech processing is an important problem in several applications. For example, in speech recognition, the goal is to transcribe the speech from multiple speakers, even when they are speaking simultaneously. In speaker diarization, the aim is to determine which speakers are speaking at any given time, even if they speak over each other. In both of these examples, there is an application of interest (e.g. speech recognition, speaker diarization), that is significantly complicated by the presence of overlapped speech. To address these problems, the existing literature has focused on solutions that first separate the multiple sources, then process the individual channels using algorithms trained on non-overlapped speech [1]. We posit that application-specific solutions that consider overlapped speech during training can result in improved performance. In this paper, we focus on a specific application - overlapped keyword recognition - and propose a network architecture that simultaneously provides a prediction of all words spoken without first having to perform source separation.

Source separation has a long and storied history in speech signal processing. Initial attempts to solve the problem focused on traditional signal processing methods [2–5]; however, recent approaches formulate speech separation as a supervised learning problem, where the discriminative patterns of speech, speakers, and background noise are learned from training data [6]. In this context, new approaches based on deep learning [7, 8] and new permutation invariant training (PIT) schemes [9] have shown considerable improvement over traditional schemes. While these methods have moved the state-of-the-art forward, single-channel source separation is still considered a very difficult problem. As such, we expect that the difficulty of the source separation problem makes it a rate-limiting factor in improving the performance of systems that rely on it.

Existing systems for overlapped speech recognition rely on source separation. For example, [10] uses a multi-channel separation approach to first split streams, and then recognize overlapped speech from recorded conversations with multiple speakers. Similarly, [11] uses a location-based angle feature, instead of relying on blind separation. While these approaches show promise in the multi-channel case, they aren't readily applicable to the single-channel scenario. In fact, given the difficulty of the single-channel source separation, it seems ill-advised to first attempt to solve the much more difficult problem of source separation and then to perform recognition.

Our aim in this paper is to directly process the single-channel overlapped speech, without having to explicitly perform source separation. We begin with a state-of-the-art (SOTA) baseline in keyword recognition based on the residual network (ResNet) architecture [12]. We modify the architecture by adding a capsule network layer [13]. Capsule networks, with capsule layers connected by dynamic routing, have shown superior performance for overlapped digit recognition [13] and overlapped object segmentation in computer vision [14, 15]. This is because dynamic routing between capsules ensures that the output from lower-level capsules is sent to an appropriate higher-level capsule and allows higher-level capsules to ignore all but the most relevant features and preserve relative feature position. Here, we show the promise of capsule networks in overlapped speech processing as well. Our ResNet + Capsule network (ResCap) outperforms an architecture where the source separation and keyword recognition are considered as separate tasks. We also compare against a SOTA baseline in keyword recognition on a series of 10-word overlapped keyword examples, using speaker dependent and speaker independent training. The main highlights of our results include:

- ResCap has comparable performance to SOTA in non-overlapped keyword recognition.
- ResCap outperforms methods that first separate sources then recognize overlapped speech and a SOTA method for single keyword recognition that was modified to handle overlapped speech.
- ResCap also outperforms these methods when there is mismatch between the training set and testing set.



http://dx.doi.org/10.21437/Interspeech.2019-2913

## 2. Keyword Recognition Network

### 2.1. Baseline Network: ResNet15

For our baseline keyword spotting architecture, we chose the ResNet architecture from [16]. As in the standard ResNet architectures, the network consists of a bias-free convolution layer followed by 6 residual blocks. Each residual block has two convolution layers with each convolution layer followed by a batch normalization layer. To increase the receptive field of the network, an exponential sized dilation is applied to each convolution layer. The six residual blocks are followed by another convolution layer with batch normalization. The output is average-pooled and fed into a fully connected layer for classification. Details of the ResNet15 network are given in Table 1. We also studied the performance of ResNet with additional convolutional layers to make it more competitive with ResCap. Unfortunately such a network increased the number of parameters but did not increase the accuracy, and so was not considered.

Table 1: *ResNet15 Configuration*

| layer | kernel | channel | dilation |
|---|---|---|---|
| conv1 | 3 | 45 | none |
| res$[i]$ $i \in [1:6]$ | 3 | 45 | $2^{i/3}$ |
| conv2 | 3 | 45 | 16 |
| avg-pool | none | 45 | none |
| FC softmax | none | 10 | none |

### 2.2. Speech Separation Baseline Based on TasNet

Conventional methods for overlapped speech recognition are based on speech separation. Among different speech separation methods, we choose the Convolutional Time-domain audio separation network (Conv-TasNet) to separate overlapped keywords. Conv-TasNet is an encoder-separation-decoder neural network that uses trainable 1D-convolution layers to learn a separation mask that separates speakers [8]. We refer the reader to [8] for a detailed description of the network.

The configuration in Table 2 achieves optimal scale-invariant source-to-noise ratio (SI-SNR) and source to distortion ratio (SDR) values on the WJS0 corpus [8], therefore we use this network in our evaluations. A follow-on ResNet architecture is used for keyword recognition. To adapt the ResNet to separated speech from TasNet, we separate the overlapped keywords in the training set using TasNet and adapt the ResNet with the separated keywords. To compare the performance of a separation+recognition method, namely, TasNet+ResNet, with our recognition method that implicitly considers multiple sources, we trained TasNet and ResNeT separately.

Table 2: *Conv-TasNet Configuration*

| Description | Value |
|---|---|
| Number of filters in autoencoder | 256 |
| Length of filters (in sample) | 20 |
| Number of channels in bottleneck conv block | 256 |
| Number of channels in convolutional blocks | 512 |
| Kernel size in convolutional blocks | 3 |
| Number of convolutional blocks in each repeat | 8 |
| Number of repeats | 4 |

### 2.3. Residual + Capsule Networks for Overlapped Speech

ResCap is built by adding a capsule network to ResNet. Specifically, the average pooling and fully connected layers after the residual blocks is replaced with capsule layers. Figure 1 shows the architecture. The output of the convolutional layers is processed by the primary capsules and class capsules in an iterative manner. In our setup, there are 10 class capsules corresponding to 10 keyword classes. The output of the class capsules is fed into a three-layer multi-layer perceptron (MLP) to reconstruct the features. The information flow between the two capsule layers is done through the dynamic routing algorithm [13]. In each iteration, the weights between primary capsules and class capsules are adjusted based on whether the class capsule 'agrees' more with a primary capsule or not. An increase in the weights implies that the class capsule will pay more attention to the specific primary capsule in the following iteration.

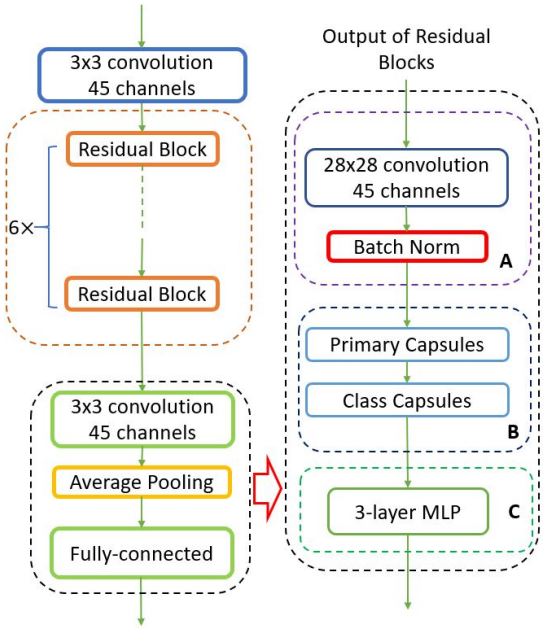


Figure 1: *ResCap: Replace pooling and fully-connected layer of ResNet15 with capsule layers. Block A:Convolutional layer for feature dimension reduction. Block B: Capsule layers. Block C: MLP for reconstruction*

While the dynamic routing algorithm helps improve performance, it introduces extra parameters during training. To limit the computation cost, we add a convolutional layer with a stride of 2 to reduce the number of capsules. After extensive experimental evaluation, we found that a larger convolutional kernel leads to better performance. Since the network performance stops improving after kernel size $28 \times 28$, we set the kernel size to be $28 \times 28$. We keep the number of channels to be 45, which is the same as the number of channels in the residual blocks.

We group the batch-normalized output of the convolution layer into 45-dimensional primary capsules, where each primary capsule includes all the values related to one receptive area of 45 channels. Thus the number of primary capsules is the same as the number of features in a single feature map. We use 16-dimensional class capsules, where each dimension models the characteristics of the keyword and the length of the vector demotes the probability of existence of the keyword. We choose the dimensions of the primary and class capsules based on extensive empirical experimentation.

The loss function in a capsule network consists of a cross-entropy term and a reconstruction term [13]. A separate marginal loss allows multiple classes to be detected simultane-

ously. The loss function for an object of class $k$, $L_k$, is

$$\begin{aligned} L_k =& T_k \max(0, m^+ - ||V_k||)^2 \\ &+ \lambda(1 - T_k) \max(0, ||V_k|| - m^-)^2, \end{aligned} \tag{1}$$

where $V_k$ is the length of the vector in class capsule $k$, $T_k = 1$ if class $k$ is present, $m^- = 0.1$ and $m^+ = 0.9$, $\lambda$ is for down-weighting the loss for absent classes, and $\lambda = 0.5$. To regularize the model, the network reconstructs the input array using three fully connected layers [13]. Since filter bank features are more complex than images, we increase the number of nodes in hidden layers (from 512 and 1024 to 1024 and 2048) to better handle the reconstruction task. The reconstruction error is scaled mean square error (MSE) between the reconstructed array and the input array. This error is added to the marginal loss to form the loss function. Only the outputs of capsules with the correct labels take part in the reconstruction. Using reconstruction as a regularizer turns out to be an important factor for the improved performance in overlapped speech recognition using capsule networks.

# 3. Experiments and Results

## 3.1. Experimental Setup

### *3.1.1. Dataset*

We use keywords from Google's Speech Commands Dataset [17] to evaluate the networks. We select the following ten classes of keywords: "backward", "bed", "follow", "forward", "marvin", "nine", "sheila", "six", "visual" and "wow". Every class of keywords is split into three parts: 80% for training, 10% for validation, and 10% for testing. We built both speaker-dependent and speaker-independent testing sets, where samples for the speaker-dependent testing set are derived from speakers in the training set, and samples for the speaker-independent sets are derived from speakers not in the training set.

We choose the frame size to be 25ms with a frame stride of 10ms, resulting in a 15ms overlap between frames. From each frame, we extrtact filterbank features (60 triangular filters). The sampling frequency is 8000 Hz. We apply a 0.97 pre-emphasis filter to balance the spectrum.

### *3.1.2. Model Training*

We use the same training method as [16] to train ResNet. Specifically, we use the cross entropy loss with stochastic gradient descent using a momentum of 0.9 and an initial learning rate of 0.1. We use the Adam optimizer described in [13] to train the ResCap network. We use a batch size of 50 and train the network for 40 epochs. We use the PyTorch [18] 0.4.1 framework on Linux Ubuntu. The machine configuration is i7-6700K @4.00GHz, 32GB RAM and single Nvidia GeForce GTX 1080 Ti 11GB card.

## 3.2. Experiments and Evaluation

### *3.2.1. Non-overlapped Keyword Recognition*

In this experiment, we use single keywords from ten classes to evaluate the performance of ResNet and ResCap. Both networks are trained on single keywords and both output predictions for each of the ten classes. The network decision for ResNet is the class with the largest activation value while the decision for ResCap is the class capsule with the largest Euclidean norm. Table 3 lists the test accuracies of ResNet and ResCap. We see that both networks achieve high accuracy for both the speaker independent and speaker dependent cases.

Table 3: *Test accuracy for recognizing single keywords for the speaker independent (SI) and speaker dependent (SD) cases*

| Model | SI | SD |
|---|---|---|
| ResNet | 94.66% | **97.67%** |
| ResCap | **96.03%** | 96.88% |

### *3.2.2. Overlapped Two-Keyword Recognition*

We build a two-keyword overlapped dataset by overlapping two different keywords from ten classes of keywords. There are in total 45 different combinations. The network decision for ResNet is formed by the two largest activation values while the decision for ResCap is based on the two class capsules with the largest Euclidean norms.

To evaluate the speech separation+recognition method, we train TasNet for speech separation using the overlapped keyword data set. TasNet is trained for 80 epochs with the average SI-SNR on test set being 13.84. Next, we use single keywords from the training set to train ResNet for recognizing the separated keywords. ResNet is trained for 40 epochs. For testing, the overlapped keywords are given as input to the trained TasNet and the two separated keywords from the output of TasNet are fed to the trained ResNets. The recognition accuracies of Tasnet+ResNet, ResNet and ResCap architectures are shown in Table 4.

Table 4: *Test accuracy of recognizing two overlapped keywords from 10 classes*

| Model | SI | SD |
|---|---|---|
| TasNet+ResNet | 73.09% | 73.08% |
| ResNet | 88.89% | 88.04% |
| ResCap | **89.33 %** | **90.44 %** |

The accuracy results show that both ResNet and ResCap achieve much higher accuracies than the separation+recognition method implemented using TasNet+ResNet. While the separated TasNet samples are intelligible to humans, there are also artifacts that have a negative impact on the follow-on keyword recognition algorithm. In our experiments, we initially trained ResNet with clean speech samples and then adapted it with the training and validation data that was first processed by TasNet. This allowed the model to adapt to the artifacts; however, as is clear from the results, this was not sufficient to improve performance to the level of ResNet or ResCap. It is possible that the TasNet performance could be further improved if we had a larger training set.

While both networks show high recognition accuracies, ResCap outperforms ResNet in both speaker independent (0.44% higher) and speaker dependent (2.4% higher) cases. In this experiment, we also analyze the confusion matrices for both networks. ResCap outperforms ResNet for every keyword class, with 6 of the 10 classes having error rates reduced by more than 1%. There are several cases of potential confusion where ResCap performs better than ResNet. The capsule network can better deal with keywords that have similar features, for example, 'forward' and 'follow'. In the speaker independent case, while both models make wrong predictions for "follow" and "forward", ResCap's error rate is lower (Label = follow, Prediction = forward: 1.72% compared to 3.08%).

#### 3.2.3. Train-test Data Mismatch

In overlapped keyword recognition, determining the number of overlapped keywords a priori, is difficult. Moreover, training data for different number of overlapped keywords may not be available. To verify the robustness of ResCap network in recognizing overlapped keywords, we test both networks that have been trained on two overlapping keywords with samples that have different number of overlapped keywords. Specifically, we evaluate the accuracy performance on test sets built with non-overlapped keywords and three overlapped keywords. We only judge the decision to be correct when all the predictions are correct. The results are shown in Table 5.

Table 5: *Test accuracy of model trained with two overlapping keywords in recognizing single keyword and three overlapping keywords*

| Data Set | Model | SI | SD |
|---|---|---|---|
| Non-Overlapped | ResNet | 94.92% | 95.86% |
| Keywords | ResCap | **96.98%** | **97.43%** |
| Three Overlapped | ResNet | 55.63% | 55.46% |
| Keywords | ResCap | **60.68%** | **57.95%** |

The result shows that, while both ResNet and ResCap achieve high accuracy in recognizing non-overlapped keywords, ResCap's performance is better than ResNet for both speaker independent and speaker dependent cases. It is worth mentioning that ResCap's recognition accuracies are improved compared to the case when the network was trained for single keywords (Table 3). When recognizing three overlapped keywords, ResCap has significantly higher recognition accuracies, namely, 5.05% higher in the speaker independent case and 2.49% higher in the speaker dependent case. This result validates our claim that ResCap is able to better recognize overlapped keywords when there is a mismatch between training and testing datasets.

### 3.3. Discussion

Our experiments on overlapped keyword recognition showed that ResCap shows improved performance over the baselines. While the results presented in Section 3.2 were for the case when the number of classes is 10, ResCap showed even better performance when the number of classes is 5. For instance, in train-test data mismatch experiment carried with 5 classes speaker-dependent testing set, ResCap achieved an accuracy of 79.58%, compared to 65.74% for ResNet.

Our experiments also help explain why ResCap performs better in overlapped keywords recognition. Conventional CNNs use pooling operation to make decisions on feature detection. This detection strategy makes little use of more holistic feature properties like position, size, orientation, etc. So while the pooling strategy works well in image processing, it does not work well on filterbank features in speech. For example, if we plot the filterbank features of 'god' and 'dog', local features may be similar, but the relative position of these features is different. If the network cannot make use of relative position, it is more likely to make an incorrect decision. ResCap solves this problem with capsule layers which take full advantage of the fact that spatial relationships can be modelled by vector-matrix representation [13]. Using a multi-dimensional vector instead of one value to encode a feature enables the network to make more reasonable decisions.

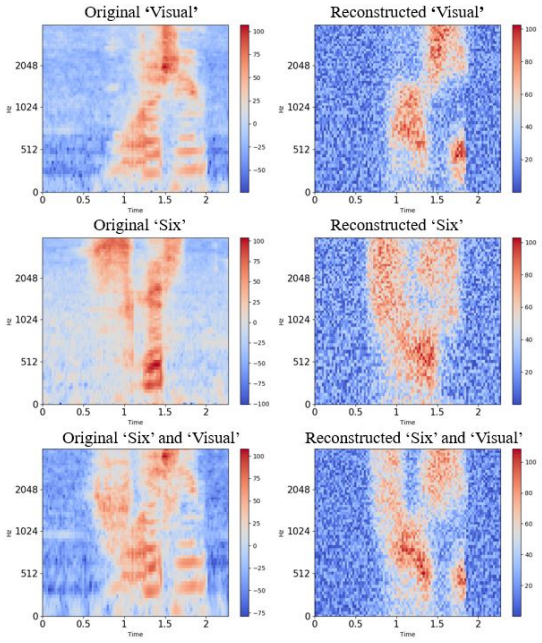


Figure 2: *Original and reconstructed features for three words*

Another advantage of ResCap in dealing with overlapped keywords is the dynamic routing algorithm between capsules. We can see the effect of the dynamic routing algorithm by analyzing the reconstruction result. After ResCap makes the decision, we can use the predictions to reconstruct the features of the input sample. When reconstructing overlapped features, we use the values in class capsules predicted by ResCap as the input of a reconstruction MLP. One set of original features and reconstructed features is shown in Figure 2. We can see that the reconstruction result from a single capsule matches the features of the corresponding single keyword. We posit that the reconstruction criteria built into the training process makes this model more robust to the overlapped keyword case when compared against the baselines.

## 4. Conclusion

This work aims to use a convolutional neural network based on a deep residual network and a capsule network to solve the single-channel overlapped keyword recognition problem without first having to carry out speech separation. We introduce the capsule layers with dynamic routing to a SOTA ResNet architecture. The results show that the ResCap network yields similar performance in non-overlapped keyword recognition when compared to the ResNet architecture, and outperforms ResNet in overlapped keyword recognition, especially when there is a mismatch between the number of overlapped speakers in the training and testing sets. Future work in this area will extend beyond overlapped keyword recognition to address overlapped speech recognition more broadly.

## 5. Acknowledgement

This research was supported by the National Institutes of Health Grant R01DC006859.

# 6. References


[1] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 558–565.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[5] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1562–1566.

[8] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700.

[9] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245.

[10] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4353–4356.

[11] T. Yoshioka, , Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks." International Symposium on Computer Architecture (ISCA), September 2018.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, 2017, pp. 3856–3866.

[14] R. LaLonde and U. Bagci, "Capsules for object segmentation," *arXiv preprint arXiv:1804.04241*, 2018.

[15] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3129–3133.

[16] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5484–5488.

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

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.