Deep connected attention (DCA) ResNet for robust voice pathology detection and classification

Aayush Kucheria and Henna Kotilainen

Introduction - what is voice pathology?

Voice pathology =

- A change in how the voice sounds
- Vocal cysts, vocal cord nodules, dysphoria, laryngitis etc.
- Causes voice hoarseness, harshness and weakness resulting in a worse voice quality.

Voice pathology affects many people

- Estimated 7.6% of U.S. adults and 7.7% of U.S. children

Voice pathology causes inconveniences in daily life and results in social problems

Detection of voice pathology

Clinical pathological voice classification is divided into two categories

- subjective evaluation
  - Visual assessment & auditory–perceptual assessment
- objective evaluation
  - Computer-aided assessment based on speech signal analysis for pathological voice classification (CS-PVC)

Image source: https://www.medicalexpo.com/prod/jiangsu-mole-electronic-technology/product-301813-1020657.html

Traditional CS-PVC

Consists of two parts: feature extraction and classification

Features that are typically looked for include:

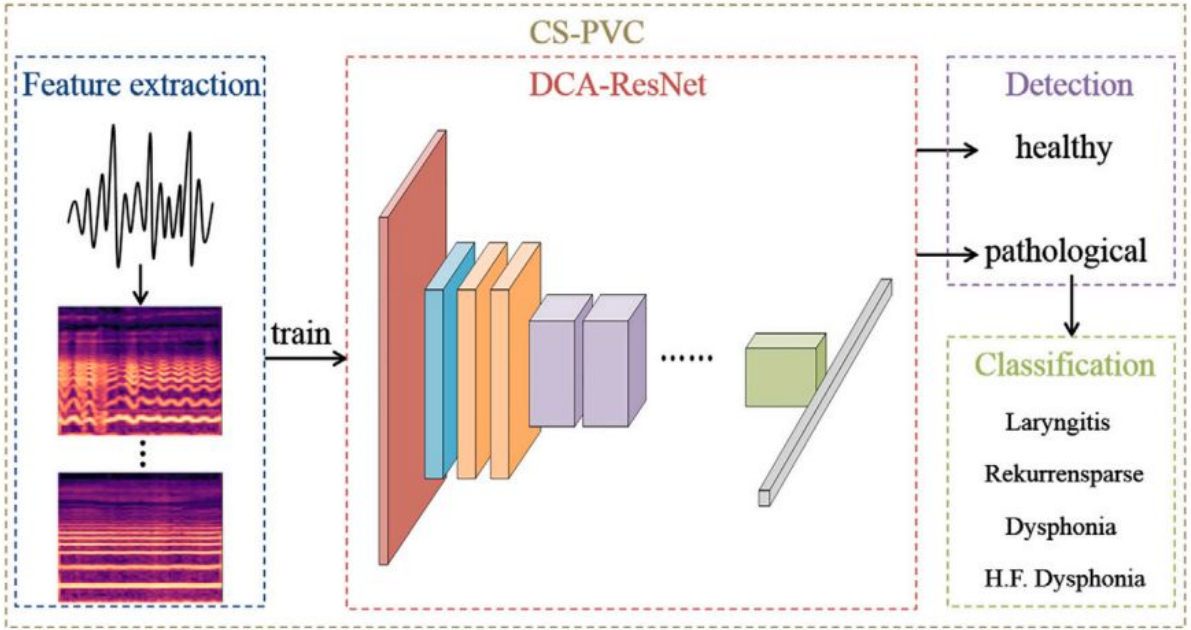
- multidimensional voice program (MDVP)
- parameters based on wavelet transform (WT)
- mel-frequency cepstral coefficients (MFCC)
- linear prediction cepstrum coefficient (LPCC)

After feature extraction, classification is applied. Traditionally lots of different classifiers have been used but all have their drawbacks.

- Gaussian mixture model (GMM), hidden Markov model (HMM), support vector, machines (SVM) and random forests (RF)
    - all used for small datasets, so they have a strong reliance on the dataset which leads to poor generalization and robustness

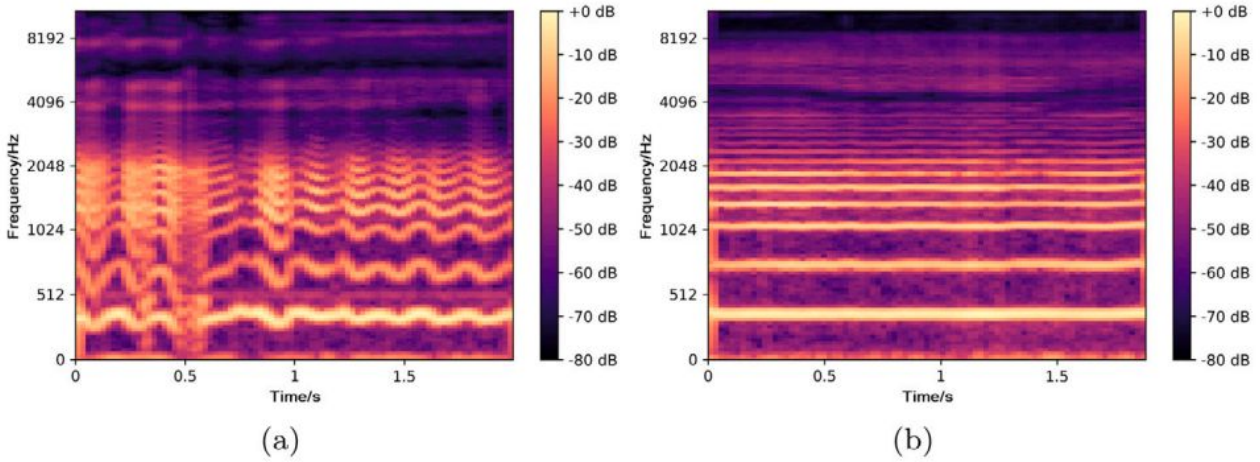
Some research has been done on using deep learning methods for pathological voice detection achieving good results.

Novel CS-PVC system



A novel CS-PVC system which can automatically detect pathological voices (pictured).

Feature extraction



Log-mel frequency spectrogram of (a) a pathological voice sample (b) a healthy voice sample for the vowel /a/.

Network structure (Overview)

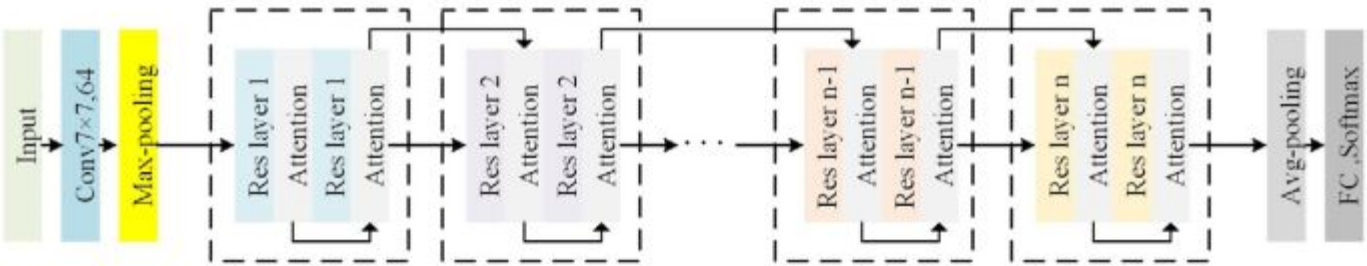


Fig. 4. An overview of DCA-ResNet.

Network structure (Residual)

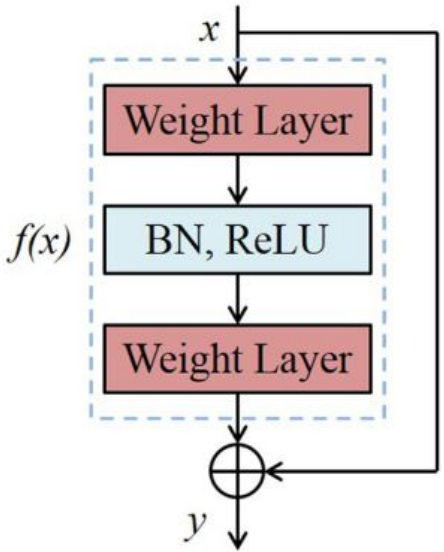


Fig. 5. The structure of the residual block.

Network structure (DCA Module)

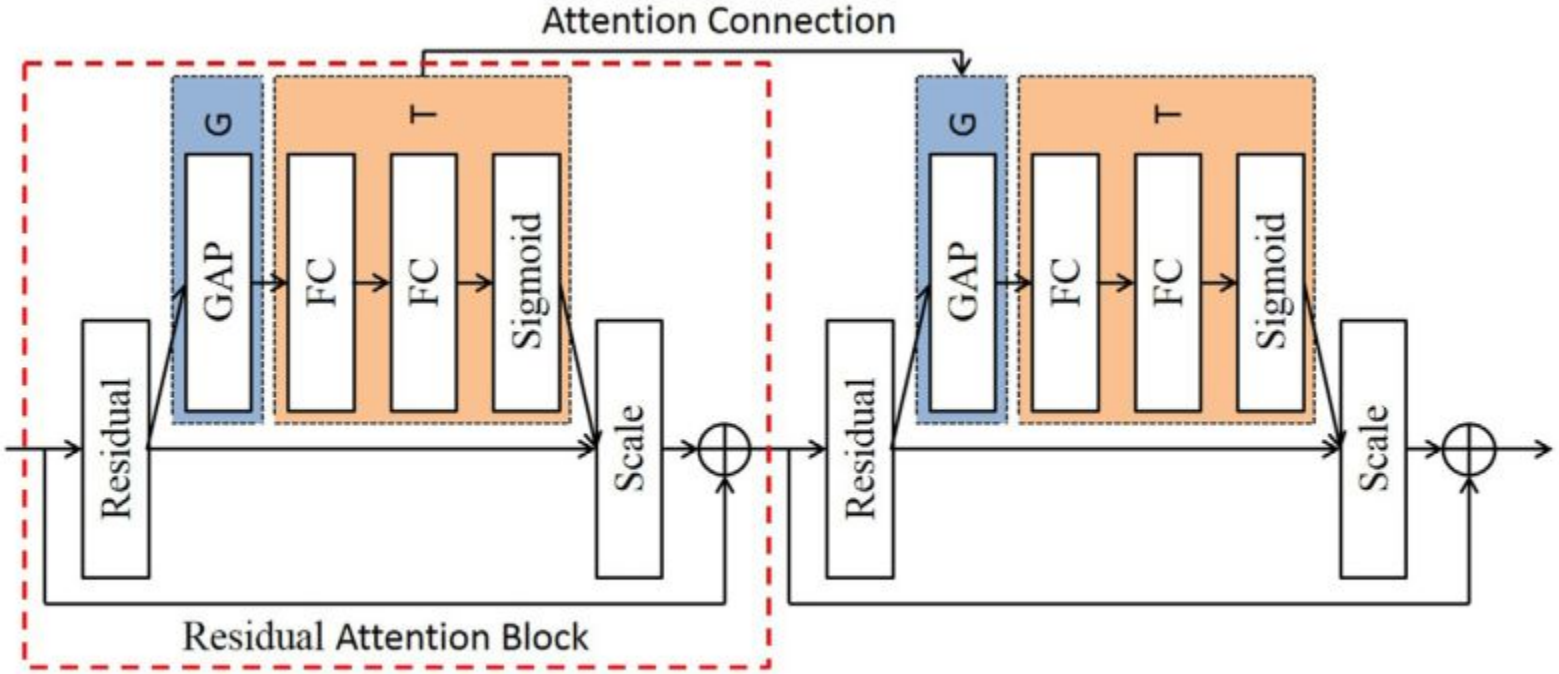


Fig. 6. The structure of the proposed DCA module.

The experiment

The dataset

Saarbruecken voice database (SVD), 2041 samples from

- 687 healthy persons (428 female, 259 male)
- 1356 patients with 71 different voice pathologies (727 female, 629 male)

Each speaker was record with

- sustained vowels /a, i, u/ produced at normal, high, low, low–high–low pitch
- The German sentence ‘‘Guten Morgen, wie geht es Ihnen?’’ (‘‘Good morning, how are you?’’).

1685 sustained normal pitch vowels /a/ were used in the experiment including 595 healthy and 1090 pathological recordings.

Self-built dataset SZUPD

- contains recordings of the vowels /a/ of 40 healthy persons and 67 patients with different voice pathologies

Evaluation Metrics

$$accuracy = tp + tn/tp + fp + tn + fn \tag{9}$$

$$precision = tp/tp + fp \tag{10}$$

$$recall = tp/tp + fn \tag{11}$$

$$F1 = 2 \cdot precision \cdot recall/precision + recall \tag{12}$$

Results

Evaluating across input features (MFSC, MFCC) and depths of ResNets

Table 2
Evaluation results using different input features, including MFSC and MFCC, and different depth of the ResNet networks.

Methods	MFSC				MFCC			
	accuracy	precision	recall	F1 score	accuracy	precision	recall	F1 score
ResNet18	**0.786**	**0.838**	0.830	**0.834**	0.751	0.768	**0.881**	0.821
ResNet34	0.766	0.806	0.839	0.822	0.762	0.811	0.826	0.818
ResNet50	0.763	0.792	0.858	0.824	0.745	0.777	0.849	0.811

Results

Evaluating across models

Table 3
The comparison of evaluation results using different models.

Meathod	Accuracy	Precision	Recall	F1 score
AlexNet [49][51]	0.721	0.772	0.807	0.789
VGG16 [36][50]	0.766	0.814	0.826	0.820
ResNet [47]	0.786	0.838	0.830	0.834
DA-ResNet	0.804	0.862	0.830	0.846
DCA-ResNet	**0.816**	**0.875**	**0.835**	**0.855**

Results

Evaluating across models

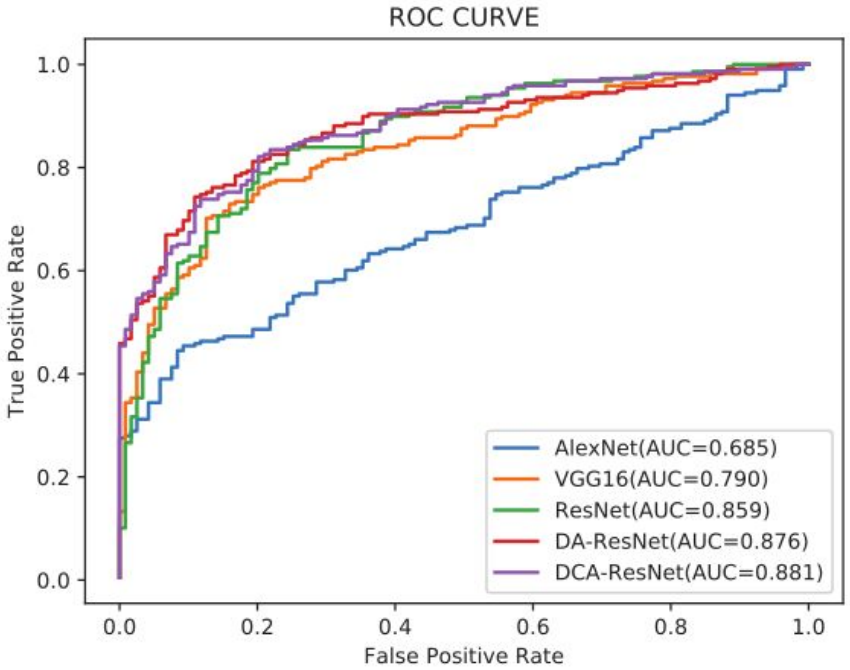


Fig. 7. ROC curves of different models.

Results

Evaluating across datasets (generalization?)

Table 4
The comparison of evaluation results using different database.

	Accuracy	Precision	Recall	F1 score
Train:SVD Test:SZUPD	0.822	0.980	0.731	0.837
Train:SVD Test:SVD	0.816	0.875	0.835	0.855

Conclusion

Key Findings

- Achieved pathological voice detection accuracies of 81.6% and 82.2% on SVD and SZUPD databases, respectively.
- Demonstrated superior performance of MFSC features over MFCC in the context of deep learning models for voice pathology detection.
- DCA-ResNet model showed better generalization across different datasets and outperformed traditional and some deep learning models.

Future Work

- Exploration of new acoustic features to enhance the distinction between pathological and healthy voices.
- Extension to continuous speech analysis for detecting voice diseases not evident in vowel pronunciation.
- Optimization of network structure for faster computation, aiming for clinical application.

As a reminder, the big picture again

- **Objective**: To develop a novel CS-PVC system for the automatic diagnosis of pathological voices using speech signal analysis, focusing on detecting significant differences between pathological and healthy voices while minimizing the impact of irrelevant information.

As a reminder, the big picture again

- **Objective**: To develop a novel CS-PVC system for the automatic diagnosis of pathological voices using speech signal analysis, focusing on detecting significant differences between pathological and healthy voices while minimizing the impact of irrelevant information.
- **Methodology**: Utilizes a deep connected attention mechanism within a Residual Network (DCA-ResNet) framework, with experiments conducted on two databases: the Saarbruecken Voice Database (SVD) and a self-built database from Shenzhen People's Hospital (SZUPD).

As a reminder, the big picture again

- **Objective**: To develop a novel CS-PVC system for the automatic diagnosis of pathological voices using speech signal analysis, focusing on detecting significant differences between pathological and healthy voices while minimizing the impact of irrelevant information.
- **Methodology**: Utilizes a deep connected attention mechanism within a Residual Network (DCA-ResNet) framework, with experiments conducted on two databases: the Saarbruecken Voice Database (SVD) and a self-built database from Shenzhen People's Hospital (SZUPD).
- **Conclusion**: The DCA-ResNet model provides a significant improvement in the automatic detection and classification of pathological voices, offering potential for universal applicability in medical diagnostics.

Home Assignment

1. How does the DCA-ResNet model proposed in the paper improve upon traditional methods for pathological voice detection?
2. What were the main findings regarding the comparison between MFSC and MFCC features in pathological voice detection using different ResNet models?
3. What challenges did the DCA-ResNet model face in accurately predicting mild voice diseases?