AUGMENTING DYSPHONIA VOICE USING FOURIER-BASED SYNCHROSQUEEZING TRANSFORM FOR A CNN CLASSIFIER

Alice Rueda and Sridhar Krishnan

Ryerson University
Electrical, Computer, and Biomedical Engineering
350 Victoria Street, Toronto, Ontario, Canada M5B 2K3

ABSTRACT

The challenge of dysphonia voice studies is always the small dataset. It is difficult to apply more sophisticated deep learning techniques without overfitting or underfitting. Convolutional neural network (CNN) is a powerful classifier that requires a large amount of training data. Data augmentation techniques for voice are limited. Fourier-based synchrosqueezing transform (FSST) can be used as a data augmentation technique to increase the data size. The results indicated that not only can FSST increase the data size, the CNN can also learn better with FSST than with Short-Time Fourier Transform (STFT) power spectrum. The loss function for FSST converges, but not for STFT. FSST is also more stable and provides more accurate results.

Index Terms— Data augmentation, signal decomposition, pathological voice, FSST, reassignment method, CNN, regularization, overfitting

1. INTRODUCTION

Dysphonia is the disorder of voice which can make the voice hoarse, breathy and weak. A lot of signal processing techniques have been applied to discriminate dysphonia voice from healthy voice. In recent years, CNN can be found in pathological voice classification studies.

CNN was originally designed for image recognition [1] and is still mainly used in image and object detection. Voice is one-dimensional (1D) data and has to be transformed into two-dimensions (2D) before being input into CNN. The current standard practice is to use a spectrogram of the voice signal as the 2D representation. There are other well-known 2D representations, such as Wigner-Ville distribution (WVD), Cohen's class, and wavelet transform. All of these representations produce a trade-off between joint time and frequency (TF) resolution and cross-term reduction. It is a challenge to obtain a clear TF representation of the signal. Restricted by Heisenberg's uncertainty principle, the balance between frequency resolution and time localization is required to obtain a better result.

Compared to statistical models, neural networks require a relatively larger amount of training data due to model complexity. In pathological voice studies, datasets are generally small and clinical datasets typically have less than 100 samples. Pathological voice analysis can benefit from data augmentation. However, augmentation techniques used in image processing are not always appropriate for voice and have to be taken with care. Scaling techniques applied to the spectrogram of an unstable voice will potentially make it stable.

In this paper, an adaptive technique is used to decompose a signal into its components using a Fourier-based sychrosqueezing transform (FSST) as a mean for data augmentation and transformation. The resulting 2D TF representation becomes the input to CNN.

This paper is organized into the following sections. Section 2 summarizes prior CNN classifications on pathological voice studies and applied techniques. Section 3 describes the dataset and method used in this study. Section 4 details the implementation of the method. Section 5 presents the results, followed by conclusions.

2. LITERATURE REVIEW ON PRIOR WORK

CNN has been used in various object detection tasks, especially in image recognition. It is also becoming a popular technique for pathological voice classification. The problem about pathological voice, including dysphonia, is the small dataset. Transfer learning with a pre-trained network designed for a similar purpose will be the best option. With the lack of pre-trained pathological voice networks, researchers are looking into the suitability of using pre-trained image networks as in [2]. Alhussein and Muhammad retrained the VGG16 and CaffeNet for images to detect pathology voice recordings from the Saarbruecken Voice Database (SVD).

Arias-Vergara et al. used CNN for Parksinon's disease (PD) speech monitoring through mobile devices with 68 *people with Parksinson's disease* (PWP) and 50 control training

Thanks to NSERC Agency and Canadian Research Chair Program for funding and Intel for support. Author emails: arueda@ryerson.ca and krishnan@ryerson.ca

subjects, and 17 PWP and 7 control test subjects [3]. Short-time Fourier Transform (STFT) was used as the 2D transformation. Frid et al. fed raw voice signals directly into a CNN for Parkinson's speech diagnosis with 43 PWP and 9 control subjects reading "Rainbow passage" [4]. Peker et al. applied a complex-valued artificial neural network with one hidden layer on features extracted from PD phonation [5].

With a better understanding of how CNN learns, Wu et al. built a deep learning network with pre-trained weight from Convolutional Deep Belief Networks (CDBN) to classify pathological voices from the SVD. The voice recordings were transformed into spectrograms as input to the CNN. The network incorporated random dropout and L_2 regularization. However, the CDBN weight initialization did not improve the accuracy of the CNN.

3. DATASET AND METHOD

Fig. 1 illustrates the conducted augmentation process. The process included converting voice signals into TF representation, performing global normalization on each type of recordings, and determining the region of interest to reduce matrix size with data augmentation before feeding to the CNN.

Fig. 1. Methodology block diagram.

3.1. Dataset

This study used the dysphonic and healthy recordings from the Saarbruecken Voice Database (SVD). Both groups contain recordings of mixed gender subjects age 25 and above. There are 255 control subjects and 94 dysphonic subjects. Down-sampling technique was used to balance the group size. The control subjects were randomly sampled to obtain 94 subjects for each type of recordings. Since there are only 349 subjects in total, global normalization was performed on each recording type before being fed into the CNN.

Recording types include short vowels of /a/, /i/ and /u/ voiced at 4 intonations each, high (_h), normal (_n), low (_l), and low-high-low (_lhl). Each recording is about $0.5-1.0$ s long, sampled at 50 kHz. Applying STFT on a 0.5 s recording with a 20 msec window and 50% overlapping will produce a small matrix of 50x513. After subject balancing, each set of voice has only 188 samples for training and testing.

3.2. Fourier-based Synchrosqueezing Transform

Synchrosqueezing transform (SST) is a type of Reassignment Method (RM) [6] that works on modified STFT. RM has been used to sharpen spectrograms by relocating elements to the nearby ridge, also known as the center of gravity of the energy distribution. This relocation process created sparse and sharpened TF representations (see Fig. 2), but increased the matrix size from 50x513 to $length_of_voice$x513. This allows separation of all stages of the voicing period and data augmentation. Ten equal distanced samples were extracted from each FSST power spectrum to augment data by 10 times.

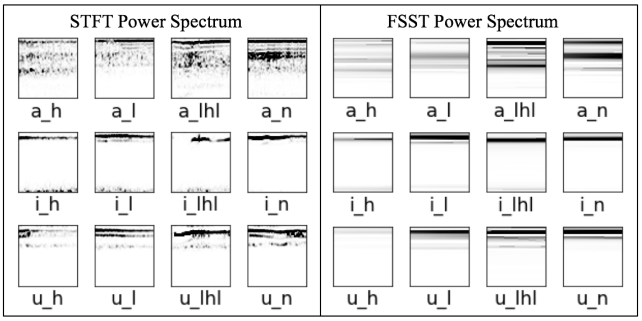


Fig. 2. STFT power spectrum on the left is not as clear as the corresponding FSST on the right.

SST, created by Daubechies and Mae for speaker identification [7], was originally based on Continuous Wavelet Transform (CWT). Fourier-based synchrosqueezing transform (FSST) [8, 9, 10, 11, 12, 13] evolved from CWT-based synchrosqueezing [14], is a phase-based technique that allows signal reconstruction and mode separation. Instead of concentrating energy distribution along both time and frequency axes as in RM, FSST operates on the modified STFT and only operates on the frequency axis.

3.2.1. FSST Model Assumptions

FSST defines signal $f(t)$ as a multi-component signal consisting of K oscillatory components, $f(t) = \sum_{k=1}^{K} A_k(t)e^{2\pi\phi_k(t)}$, where A_k is the instantaneous amplitude and ϕ'_k (derivative of the phase) is the instantaneous frequency of component k.

The components have weak frequency modulation. In other words, there exists a small $\epsilon \ll 1$, $\parallel A'_k \parallel \leq \epsilon \parallel \phi'_k \parallel$ and $\parallel \phi''_k \parallel \leq \epsilon \parallel \phi'_k \parallel$. This requires the amplitude to be differentiable and the phase twice differentiable.

The adjacent components are well-separated in frequency with a distance d, $\phi'_k - \phi'_{k-1} > d$. For a Gaussian window of size σ, it's frequency bandwidth $\Delta = \sqrt{2}\log(2)/\sigma$. The minimum distance between adjacent components is $d = 2\Delta$.

3.2.2. FSST Method

FSST is based on modified STFT, $V_g f(t,\eta)$, to reduce smearing. The additional linear phase shift, $e^{j2\pi\eta t}$, makes equation (1) a modified STFT.

$$V_g f(t,\eta) = \int f(\tau) g * (\tau - t) e^{-j2\pi\eta(\tau - t)} d\tau \qquad (1)$$

Phase Transform: When $|V_g f(t,\eta)| > 0$, the instantaneous frequency (IF) can be approximated by $\tilde{\omega}_f(t,\eta)$,

$$\tilde{\omega}_f(t,\eta) = \frac{\frac{\partial}{\partial t} V_g f(t,\eta)}{j2\pi V_g f(t,\eta)} \quad (2)$$

Synchrosqueezing operator: Squeezes the content $V_g f(t,\eta)$ to the IF curves to form the operator $T_g f(f,\eta)$,

$$T_g f(t,\omega) = \frac{1}{2\pi g(0)} \int_R V_g f(t,\eta)\delta(\omega - \tilde{\omega}(t,\eta))d\eta \quad (3)$$

A mode can be reconstructed by integrating the synchrosqueezing operator within the vicinity of its IF ridge.

3.3. Dysphonia Spectrogram Representation

In dysphonia studies, the acoustic characteristics of voice play an important role. Some of these characteristics are visible from the spectrogram of sustained vowel /a/ shown in Fig. 3. Breathy, low resonance, and hoarse voices have the energy less concentrated than the healthy voice. An unstable voice fluctuates across frequency over time.

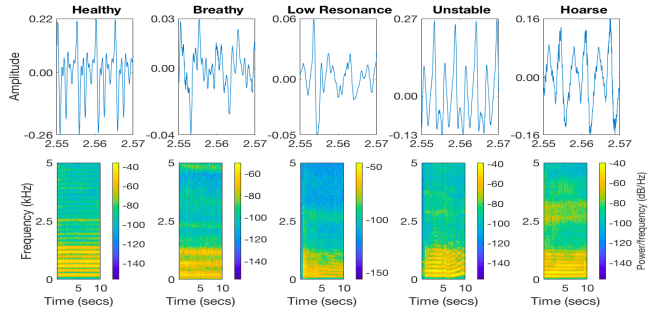


Fig. 3. Voice samples for sustained /a/ from left to right illustrating healthy, breathy, low resonance, unstable, and hoarse voices. Waveform on top of the corresponding spectorgram.

3.4. Augmentation and Normalization of CNN

Normalization of voice can change the characteristic of voice and should be taken with care since the power spectrum also reflects the intensity of the voice. A dysphonic voice can have reduced intensity. It is necessary to maintain a global relative intensity. Global normalization was used in this study.

Augmentation is a great way of increasing the data size. Unlike image objects that can be viewed from different angles, voice's TF representations are orthogonal in time and frequency domains. Rotation does not apply to 2D representation. Data augmentation techniques like scaling, rotation, and random cropping are not appropriate. A lot of the techniques normally applied to images cannot be applied to a voice's TF representation since it can change the voice's characteristics.

There are only a few papers that have studied the possible data augmentation techniques for sound [15] and voice [16]. Jiao et al. applied adversarial training to create the synthetic pathological voice. Salamon et al. used more standard image augmentation techniques adjusted for environmental sound. Augmentation techniques such as pitch shifting seem proven to be beneficial on environmental sound.

4. IMPLEMENTATION

Signal transformation into 2D was performed in MATLAB and the CNN was built in tensorflow with data pipelining in tfrecords. The 94 subjects from each group were split into 10-fold for cross-validation.

The sampling frequency of the SVD recordings is 50 kHz. Using a 1024-datapoint (21.2 msec) Gaussian window, the voice signal was transformed and synchrosqueezed into a TF representation matrix. The matrix has the time dimension as the original signal, but consists of 513 frequency bins. Each frequency bin is around 48 Hz. For a 0.5 s recording, the resulting FSST matrix is 25000x513 that can be cropped. The FSST also produced sparsity in the TF representation with virtually no information left in the high-frequency bins.

A TF power spectrum matrix dimension of 50x50 for FSST and 45x50 for STFT were used to train the CNN. Our region of interest is at the lower frequency bins where the energy is concentrated. Fig. 4 shows the box plot of the cumulatve energy distribution. The first 50 bins with frequency range 0-2.4 kHz cover 98% of the energy. Pouchoulin et al. has also indicated that the lower frequency range up to 3 kHz is more relevant for dysphonia discrimination [17].

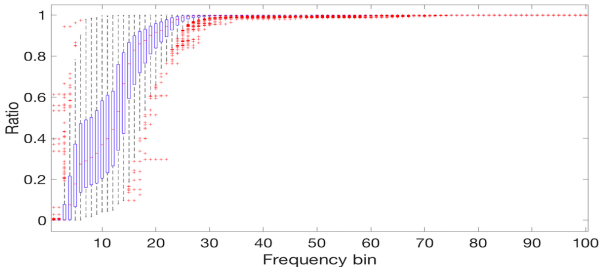


Fig. 4. Cumulative energy distribution showing 98% of the power is contained in the first 50 frequency bins, 0-2.4 kHz.

To reduce the number of training parameters, a simple CNN was used in this study. Fig. 5 outlines the CNN's layers. The CNN has 2 convolutional layers, 2 pooling layers, a fully connected (dense) layer, and an output (dense) layer. There is a flatten layer connecting the convolution layer to the dense layer. Random dropout with rate 0.5 was placed into and out of the dense layer. Each convolutional layer has 8 kernels with a small receptive field of 3x3. A small pooling size of 2x2 is used in both pooling layers. The fully connected layer has 64 neurons and the output layer has 2.

ReLU activation function was used across the network and softmax at the output dense layer. This simple network has over 62 thousand parameters to be trained. The CNN was trained over 200 epochs using binary cross-entropy loss function and Adagrad optimizer for its speed with learning rate 0.01. The dense layers also use L_2 regularization with penalty rate of 0.001.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 48, 48, 8)	80
max_pooling2d (MaxPooling2D)	(None, 24, 24, 8)	0
conv2d_1 (Conv2D)	(None, 22, 22, 8)	584
max_pooling2d_1 (MaxPooling2	(None, 11, 11, 8)	0
flatten (Flatten)	(None, 968)	0
dropout (Dropout)	(None, 968)	0
dense (Dense)	(None, 64)	62016
dropout_1 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 2)	130

Total params: 62,810
Trainable params: 62,810
Non-trainable params: 0

Fig. 5. CNN model with 2 convolutional layers, 2 pooling layers, 2 dropout layer, and 2 dense layers.

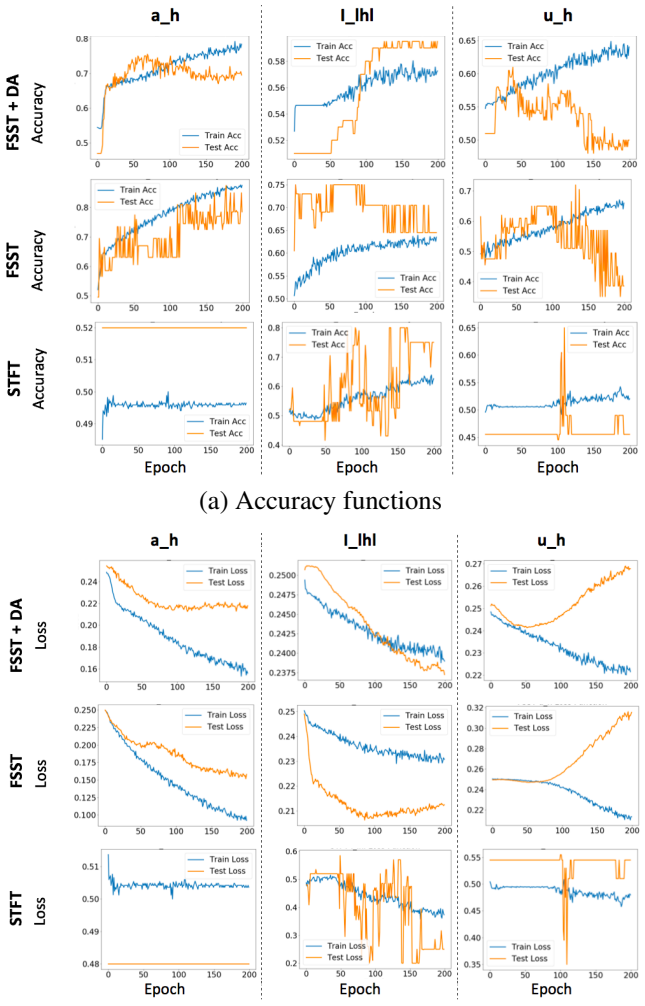
5. RESULTS

The CNN was trained with three sets of TF representation, STFT power spectrum, FSST power spectrum, and FSST augmented with 10 samples from each recording (FSST+DA).

Figs. 6 shows the CNN results using 10-fold validation. Both sub-figures show that the CNN was able to learn better with FSST representation ($> 70\%$ accuracy for a_h) than STFT representation (52% accuracy). The predictive results are more stable with FSST. It is also easier to tell when overfitting starts with FSST+DA. After 150 epochs for a_h, the performance for FSST+DA starts to drop. This is when the CNN started to learn the training data too well which produced less satisfactory validation result.

The study used all dysphonia recordings available from SVD. Fig. 6 illustrates that the loss functions for both the FSST and FSST+DA converge, but vary depending on the vowel. There is some clear indication of when the CNN started to overfit. The loss function for the test set starts to increase and the accuracy starts to decrease. The STFT power spectrum power shows difficulty in learning. The data size is simply too small for the CNN to learn.

The overall performance for all three cases is mediocre. Both sensitivities and specificities are around 60%. STFT provided a slightly higher specificity. FSST without data augmentation scores better than with augmentation. However,



(a) Accuracy functions

(b) Loss functions

Fig. 6. Comparing accuracies and loss functions between FSST with data augmentation, FSST sampled at the middle of the recording, and STFT power spectrum.

these numbers are meaningless since STFT has the highest loss function. The predicted outcomes for STFT are unstable.

6. CONCLUSIONS

Although the FSST alone performs better with FSST+DA, the sharpened FSST representation is still a better alternative than STFT for CNN. A dynamic stopping algorithm might be needed to make the CNN for various vowels. The results might be improved if the CNN was first pre-trained on a smaller handpicked dataset. Increasing the number of convolutional layers might also help with performance, but will require a larger dataset to compensate overfitting. A larger dataset can provide more meaningful and reliable results.

7. REFERENCES

[1] Y. Le Cun, O. Matan, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jacket, and H.S. Baird, "Handwritten zip code recognition with multilayer networks," in *Proceedings of 10th International Conference on Pattern Recognition*, 1990, vol. ii, pp. 35–40.

[2] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[3] T. Arias-Vergara, J.C. Vasquez-Correa, J.R. Orozco-Arroyave, P. Klumpp, and E. Noth, "Unobtrusive Monitoring of Speech Impairments of Parkinson's Disease Patients Through Mobile Devices," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 6004–6008.

[4] A. Frid, A. Kantor, D. Svechin, and L.M. Manevitz, "Diagnosis of Parkinson's disease from continuous speech using deep convolutional networks without manual selection of features," in *2016 IEEE International Conference on the Science of Electrical Engineering, ICSEE 2016*, 2017, pp. 1–4.

[5] M. Peker, B. Sen, and D. Delen, "Computer-Aided Diagnosis of Parkinson's Disease Using Complex-Valued Neural Networks and mRMR Feature Selection Algorithm," *Journal of Healthcare Engineering*, vol. 6, no. 3, pp. 281–302, 2015.

[6] F. Auger and P. Flandrin, "Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[7] I. Daubechies and S. Maes, "A nonlinear squeezing of the continuous Wavelet transform based on auditory nerve models," pp. 527–546, 1996.

[8] G. Thakur and H.T. Wu, "Synchrosqueezing-Based Recovery of Instantaneous Frequency from Nonuniform Samples," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 5, pp. 2078–2095, 2011.

[9] G. Thakur, E. Brevdo, N.S. Fučkar, and H.T. Wu, "The Synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications," *Signal Processing*, vol. 93, no. 5, pp. 1079–1094, 2013.

[10] F. Auger, P. Flandrin, Y.T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.T. Wu, "Time-frequency reassignment and synchrosqueezing: An overview," *IEEE Signal Processing Magazine*, vol. 30, no. 6, pp. 32–41, 2013.

[11] T. Oberlin, S. Meignen, and V. Perrier, "The Fourier-based synchrosqueezing transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, vol. 2, pp. 315–319.

[12] G. Thakur, "The Synchrosqueezing transform for instantaneous spectral analysis," *Applied and Numerical Harmonic Analysis*, pp. 397–406, 2015.

[13] I. Daubechies, Y Wang, and H.T. Wu, "ConceFT: Concentration of frequency and time via a multitapered synchrosqueezed transform," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, pp. 1–44, 2016.

[14] I. Daubechies, J. Lu, and H.T. Wu, "Synchrosqueezed wavelet transforms : An empirical mode decomposition-like tool," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 243–261, 2011.

[15] J. Salamon and J.P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[16] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 6009–6013.

[17] G. Pouchoulin, C. Fredouille, J-F. Bonastre, A. Ghio, and A. Giovanni, "Frequency Study for the Characterization of the Dysphonic Voices," in *INTERSPEECH*, 2007, pp. 1198–1201.