

# Deep Learning Multi-task classification of vocal technique and gender

Sudipta Mondal

## 1 Introduction

The identification and knowledge of vocal technique is pivotal for singers to maintain style diversity in their singing; and also crucial for sound production in music and entertainment industry. The analysis of singing voice has been traditionally a challenging task, but with the data availability and deep learning advancement, this is now achievable [2]. In this paper, the aim is to study and compare multi-task classification by deep learning models: CNN and LSTM, for the classification of vocal techniques and gender of audio samples from VocalSet, that comprises of monophonic audio pieces recorded by male and female singers.

## 2 Data and pre-processing

VocalSet [1] is a singing voice dataset with 10.1 hours of recordings by 20 professional singers (11 male, 9 female) with 17 different vocal techniques, constituting 3105 wav files. On analysis of the waveform and when listened, the audio samples show a few similarity of vocal techniques in male voices for slow forte and slow piano; fast forte and fast piano; belt and vocal fry; messa, trill and trillo, as shown in the Figure 2.1.

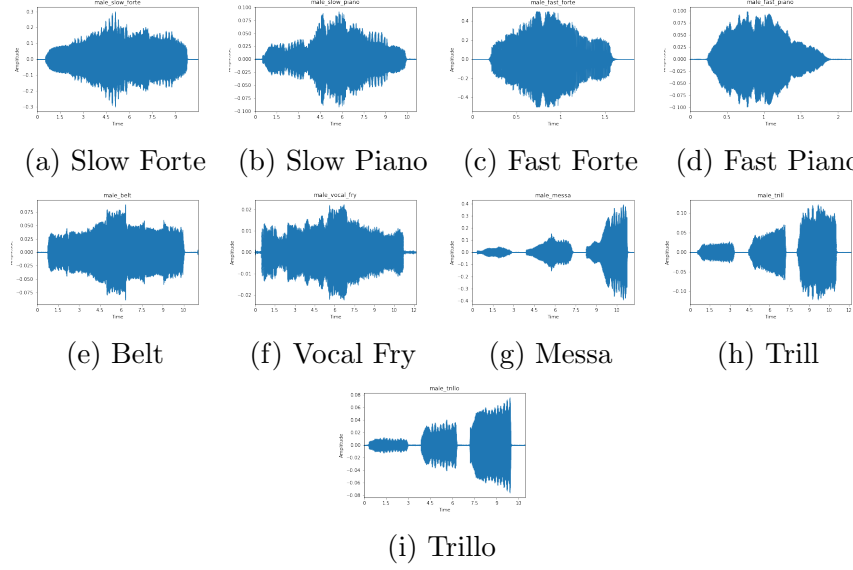


Figure 2.1: Waveform of male voice

Likewise, the amplitudes of audio samples for female voices also have some similarity in vocal techniques like belt, breathy and vocal fry; slow forte and slow piano; fast forte and fast piano, as shown in Figure 2.2.

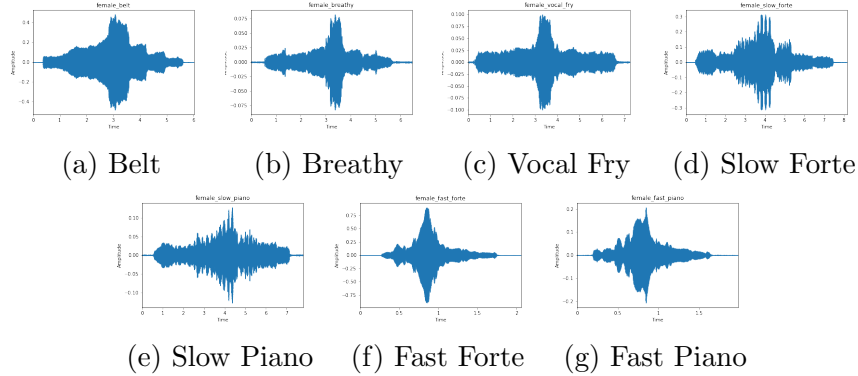


Figure 2.2: Waveform of female voice

Further exploration shows, the data is imbalanced on vocal techniques with fewer samples of female vibrato, messa, trill and male messa, forming minority classes. It is also notable that the number of audio samples for

vibrado is just 5 for female voices and none for the male voice. Therefore, the lack of data for vibrado might not be appropriate for the study of CNN and LSTM models. The focus will be on data with majority classes. Some of the samples have been hand picked from the dataset and removed from train and test set for inference later, making the total samples to 3096. The data distribution shown in Figure 2.3.

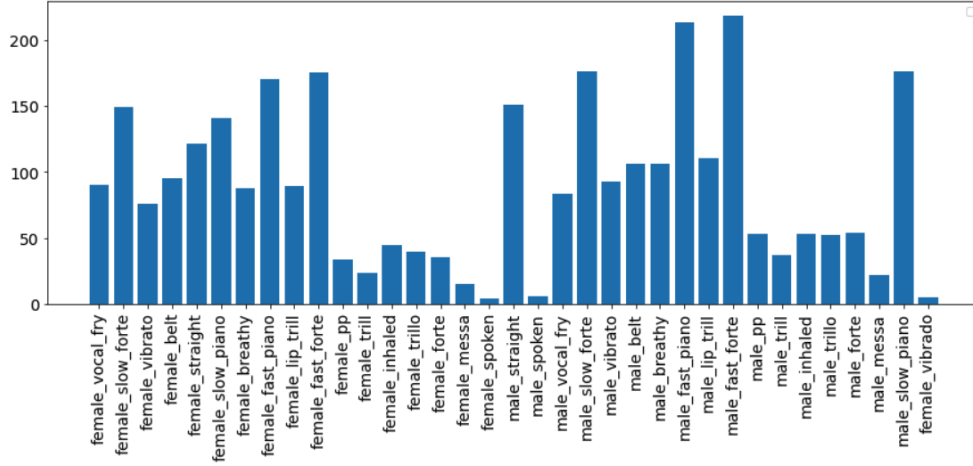


Figure 2.3: Data Distribution

The dataset is sorted in nested folders with a structure of gender, context (arpeggios, long tones, scales, and excerpts) and vocal techniques, shown in Figure 2.4.

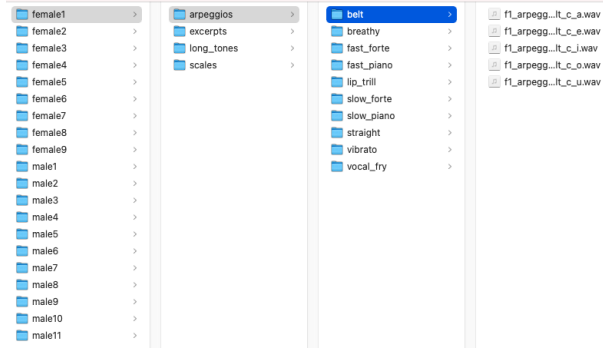


Figure 2.4: Folder Structure

Due to lack of annotated file, the gender and technique is captured from the folder names to form annotated target labels during the pre-processing. As part of pre-processing, along with getting the annotated labels (gender and vocal technique) into a dictionary, melspec of each entire audio piece is generated from the raw input audio using Librosa. See Figure 2.5 and Figure 2.6 for all combinations of techniques and gender represented as melspectrogram from VocalSet.

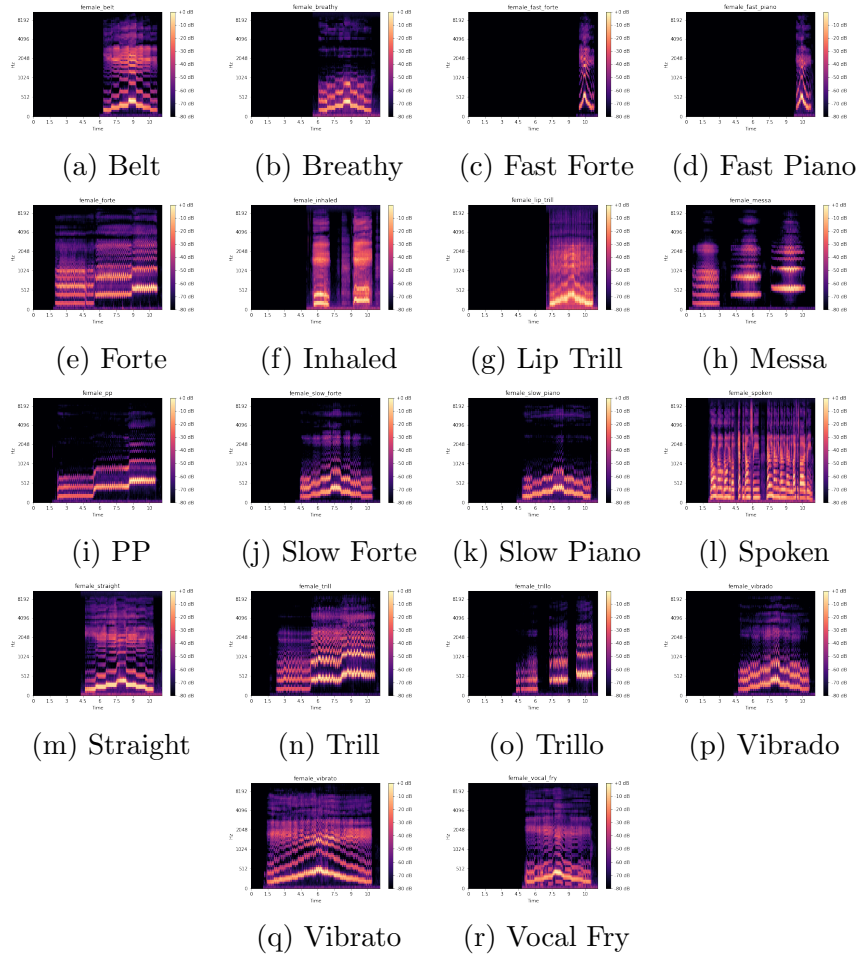


Figure 2.5: Melspectrogram for female voice techniques

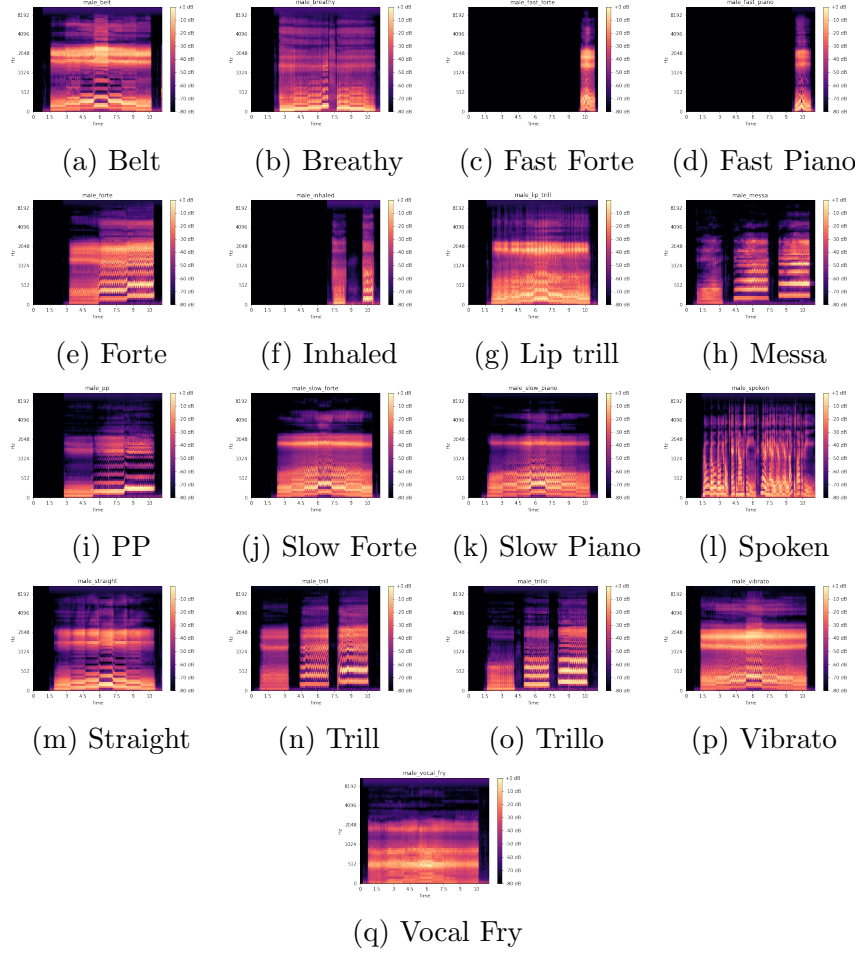


Figure 2.6: Melspectrogram for male voice techniques

The melspectrogram is extracted with a sample rate of 44.1 kHz, window length of 50 milliseconds, hop length of 25 milliseconds, 2205 number of frequencies and 64 mel filters, i.e., equally spaced frequencies to represent distance heard by human ear. The audio files vary in length, therefore the melspectrogram with smaller length is padded and each audio sample is ensured to have 500 melspec length.

The dataset is partitioned into train, validation and test set with the ratio of 70:20:10, resulting in 2228, 558 and 310 samples, respectively; ensuring that same recording is not present in any of the partition. Then the data is normalised using their mean and standard deviation to prevent the network

from using amplitude as a feature for classification [1]. Furthermore, to make the data compatible with CNN model, the 2D arrays of all datasets is converted to 3D, because CNN expects a third dimension of channel, which is 1 channel for audio. The shape of the inputs is batch, time length, melspec, with additional channel/depth for CNN. The models are trained with the training data and validation set is evaluated during training. Final evaluation of model prediction is performed on test data.

### 3 Deep Learning Methods

The audio classification problem is transformed into image-classification problem, by converting raw audio signals to melspectrogram, a visual representation of signal frequencies varying over time and converted into mel scale.

Two architectures are considered for multi-task learning by hard sharing of parameters through the layers. The first one being CNN that performs linear operation between the array of input data and two-dimensional array of weights (filter or kernel). The linear operation is element wise multiplication between filter and input. In convolutional layer, the filter acts as a feature detector that the network learns during the training with backpropagation of errors. The second model is LSTM that processes sequences of data. It comprises memory cell(s) to remember information across time intervals along with input, output and forget gate, to regulate the flow of information in and out of the memory cell. Here, the network learns by backpropagation through time (BPTT), i.e., errors calculated for each time step. See both the architecture in Figure 3.1.

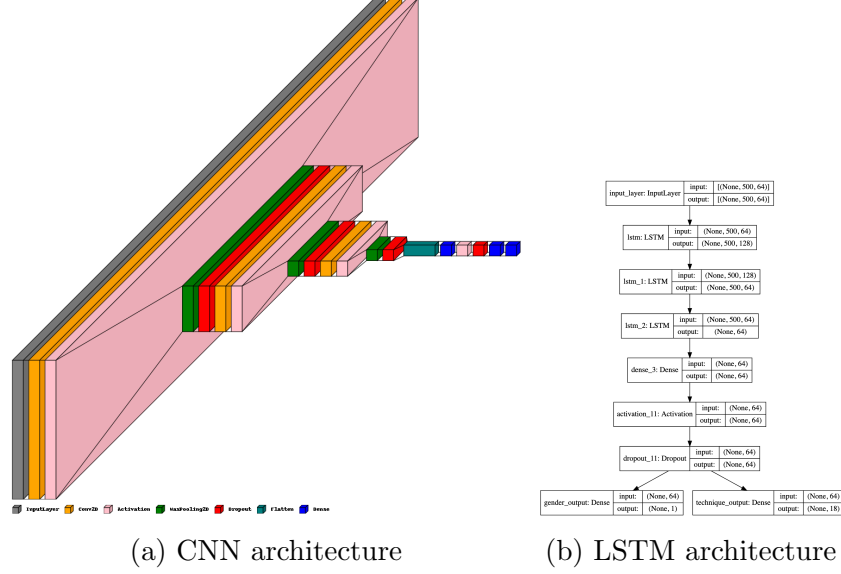


Figure 3.1: Architectures

Each model is learned using RMSProp optimizer with a learning rate of 0.0001 and trained with a batch size of 64 for 200 epochs, and best model is saved by using ModelCheckpoint. Binary categorical and sparse categorical cross entropy is employed as loss function for gender and vocal technique, respectively. Accuracy and loss for train and validation data is plotted against the epochs for understanding the early stopping criteria. Finally, a dense layer with sigmoid activation is used for the gender classification task because it can be either male or female; and a dense layer with softmax for vocal technique classification task due to number of output classes being 18. Softmax provides the probabilities of belonging to each class. The below sections cover each model in detail.

### 3.1 Convolutional Neural network (CNN)

The CNN model is built with 3 layers. The first layer consists of 64 convolutional 3x3 filters with ReLU non-linearity, followed by 3x3 max pool, with a drop out rate of 40%. The second layer has 32 convolutional 3x3 filters with ReLU non-linearity, followed by 3x3 max pool, with a drop out rate of 40%. The third layer has 16 convolutional 3x3 filters with ReLU non-linearity, followed by 3x3 max pool, with a drop out rate of 40%. The default stride of 1

has been used to capture the temporal sequences of the audio data for every time step. The final dropout layer is flattened to feed into the dense layer with 64 hidden units, followed by a final dropout layer after the dense layer. The dropout layer helps to reduce overfitting and max pool layers aid in down sampling the feature maps capturing important and relevant features only.

### 3.2 Long-Short term memory (LSTM)

LSTM is a type of recurrent neural network that stores and uses the previous data as information to make predictions and is suitable for time series related problems. The LSTM model is built with 3 layers as well for comparison with CNN. The first layer consists of 128 memory units implemented with return sequences as True, so that the hidden state for each input time step is captured. The second and third LSTM layer have 64 units. The final LSTM layer is fed to a dense layer, followed by ReLU activation for non-linearity and a dropout layer.

## 4 Experimentation and Evaluation

### 4.1 Model loss and accuracy

In CNN model, the training versus validation loss and accuracy for both gender and technique appear to be improving equally with the number of epochs as shown in Figure 4.1 and Figure 4.2.

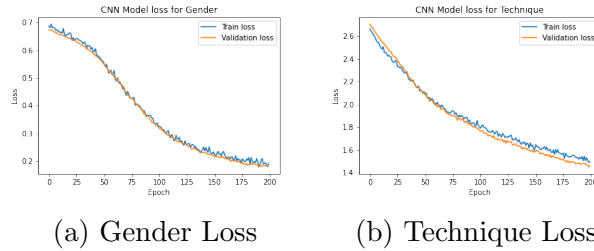


Figure 4.1: Plot for training and validation loss



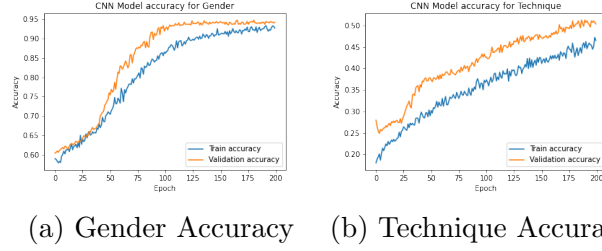


Figure 4.2: Plot for training and validation accuracy

In LSTM, the training versus validation loss and accuracy follows the similar trend after almost 25 epochs, but the validation loss remains higher than the training loss as shown 4.3 and Figure 4.4.

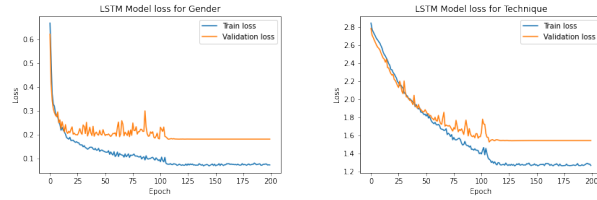


Figure 4.3: Plot for training and validation loss

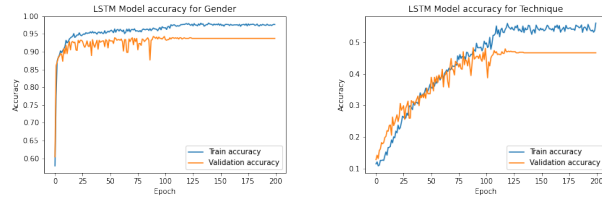


Figure 4.4: Plot for training and validation accuracy

The test accuracy for CNN and LSTM model was 94.19% and 95.48% for gender, respectively; and accuracy of 52.90% and 44.52% for vocal techniques, respectively. The classification of the vocal techniques is more complex as compared to gender classification, which might be the reason for lower accuracy.

## 4.2 Confusion matrix

The confusion matrix for CNN in Figure 4.5 reveals fast forte, fast piano, slow piano, slow forte, straight, belt and vocal fry are identified correctly in most instances. Trill, lip trill vibrato and breathy are not correctly identified in most instances due to the lesser samples presumably. Interestingly, the confusion I had when I heard fast forte and fast piano, slow forte and slow piano was appropriately captured by the model as well and got confused.



Figure 4.5: Confusion matrix of vocal technique for CNN

The confusion matrix for LSTM in Figure 4.6 reveals a similar trend of confusion with fast forte, fast piano, slow piano and slow forte and at par with the CNN model. Lip trill is identified better by LSTM, while trill,

vibrato and breathy follow similar trend as CNN.



Figure 4.6: Confusion matrix of vocal technique for LSTM

Overall, CNN performs better than LSTM for vocal technique classification, though gender classification is nearly same.

### 4.3 Inference of the models on test samples

For the inference and case study I had recorded my voice with 4 vocal techniques-fast forte, slow forte, belt and vocal fry, along with some hand picked male and female samples from VocalSet of the majority classes. The wav file was processed in similar manner as training data to get melspectrogram and normalized with scalar fit on the trained data. I captured the

gender and top 3 vocal technique for the combined prediction. CNN model captured gender correctly and, vocal techniques are predicted correctly in one of the top 3 predictions as well in most instances. LSTM incorrectly classifies gender in more instances as compared to CNN, and predicts vocal technique correctly only in some instances. Below are the results:

Model	Audio file	Gender predicted	Vocal technique		
			Top 1	Top 2	Top 3
CNN	Sudipta_belt_now.wav	female	breathy	lip_trill	vocal_fry
	Sudipta_fast_forte.wav	female	fast_forte	forte	trillo
	Sudipta_slow_forte.wav	female	belt	vocal_fry	breathy
	m4_arpeggios_c_slow_forte_a.wav	female	belt	vocal_fry	straight
	f2_arpeggios_c_fast_piano_e.wav	female	fast_piano	fast_forte	inhaled
	m5_arpeggios_c_fast_forte_a.wav	female	fast_piano	fast_forte	inhaled
	m2_scales_c_fast_piano_a.wav	male	fast_piano	fast_forte	inhaled
	m4_arpeggios_straight_e.wav	female	belt	vocal_fry	straight
	f2_arpeggios_belt_a.wav	female	belt	vocal_fry	fast_forte
	m2_scales_breathy_u.wav	male	slow_piano	fast_piano	breathy
	Sudipta_vocalfry.wav	female	belt	fast_forte	vocal_fry
	m3_scales_belt_a.wav	female	belt	slow_forte	vocal_fry
LSTM	Sudipta_belt_now.wav	female	lip_trill	slow_piano	breathy
	Sudipta_fast_forte.wav	female	lip_trill	slow_piano	breathy
	Sudipta_slow_forte.wav	male	fast_piano	slow_piano	vocal_fry
	m4_arpeggios_c_slow_forte_a.wav	female	belt	slow_forte	vibrato
	f2_arpeggios_c_fast_piano_e.wav	female	fast_forte	fast_piano	slow_forte
	m5_arpeggios_c_fast_forte_a.wav	male	fast_piano	fast_forte	slow_piano
	m2_scales_c_fast_piano_a.wav	male	fast_forte	fast_piano	slow_forte
	m4_arpeggios_straight_e.wav	female	belt	slow_forte	vibrato
	f2_arpeggios_belt_a.wav	female	inhaled	straight	vocal_fry
	m2_scales_breathy_u.wav	male	slow_piano	fast_piano	slow_forte
	Sudipta_vocalfry.wav	female	lip_trill	slow_piano	breathy
	m3_scales_belt_a.wav	female	trillo	forte	trill

Figure 4.7: Inference results













Audio file	Melspec	Audio file	Melspec
Sudipta_belt_now.wav		m4_arpeggios_straight_e.wav	
Sudipta_fast_forte.wav		f2_arpeggios_belt_a.wav	
Sudipta_slow_forte.wav		m2_scales_breathy_u.wav	
m4_arpeggios_c_slow_forte_a.wav		Sudipta_vocalfry.wav	
f2_arpeggios_c_fast_piano_e.wav		m2_scales_c_fast_piano_a.wav	
m5_arpeggios_c_fast_forte_a.wav		m3_scales_belt_a.wav	

Figure 4.8: Sample melspectrogram

## 4.4 Improvements on base model

A few more experiments were carried out with variations in base CNN and LSTM models. With additional dense layer in CNN and updating the hidden units of first and second dense layer to 256 and 128, respectively, boosts the accuracy of gender and vocal technique classification by 3%. Some of the vocal techniques like belt, breathy, pp and vocal fry saw increase in accuracy. A single Bidirectional LSTM model with 64 memory units, return sequences true, a dense layer and dropout was compared against LSTM. The difference being a unidirectional LSTM only preserves information of the past, however bidirectional preserves information from both past and future at a given current time step. BiLSTM achieved highest test accuracy amongst all the models, however the inference seems it is overfitting due to large number of parameters and the predictions are not very satisfactory. See the table 4.9 for test accuracy results of base model and improvements.

Models	Trainable parameters	Gender accuracy	Vocal technique accuracy
CNN-base model	61,891	94.19%	52.90%
LSTM-base model	186,643	95.48%	44.52%
Improvements	Trainable parameters	Gender accuracy	Vocal technique accuracy
CNN with additional dense layer and updating the hidden units of the first and second dense layer to 256 and 128, respectively	206,787	96.13%	55.48%
Single BiLSTM layer with 64 memory units, return sequences true, a dense layer and dropout	4,163,347	98.71%	61.61%

Figure 4.9: Base model and Variations

## 5 Conclusion and future works

The classification of vocal technique and gender is vital for sound production, music composition and incorporating emotions in movies, and can be useful for recommendation systems. This multi-task learning for classification is implemented with two models. Inference on the base models suggest CNN performs quite well and can be further worked upon for improvement. The experiment with different variations show that Bidirectional LSTM achieves high test accuracy for both the tasks, but seems to be overfitting. However, both models suffer with lower accuracy for vocal technique classification due to the complexity of the task.

As future work, to deal with the data imbalance for vocal techniques and to improve the models, audio data can be synthesized for minority classes and deep embeddings can be learnt in the models similar to the one mentioned in paper [3]. Furthermore, MFCC feature extraction could be studied as compared to melspectrogram and chunking the audio samples for better feature capturing might be effective. Deeper and wider CNN model and training with more epochs might improve the performance of the models.

## References

- [1] Wilkins, J., Seetharaman, P., Wahl, A. and Pardo, B., 2018, January. VocalSet: A Singing Voice Dataset. In ISMIR.
- [2] Gómez, E., Blaauw, M., Bonada, J., Chandna, P. and Cuesta, H., 2018. Deep learning for singing processing: Achievements, challenges and impact on singers and listeners.
- [3] Arora, V., Sun, M. and Wang, C., 2019, May. Deep embeddings for rare audio event detection with imbalanced data. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).