Modeling Subjective Vocal Attributes For Conservatory Singers Using Deep Learning

Freie Universität Berlin
Seminar Deep Learning for Biomedical Applications

Reuben Walker
Student Number: 5593708
reuben-scott.walker@charite.de

Date of Paper Submission: August 28, 2025

Supervisor:
Prof. Dr. Vitaly Belik[*], Freie Universität Berlin, Germany

Contents

[*] Institut für Veterinär-Epidemiologie und Biometrie

1 Summary

This study investigates whether subjective pedagogical labels, particularly timbre, can be predicted from acoustic features in a longitudinal dataset of classical singers. Using an institutional database of aligned student evaluations and sung samples, we trained CNN and MLP models on mel spectrograms and MFCCs to classify various vocal characteristics. While CNN models yielded limited significance, the MLP models achieved statistically significant results in predicting timbre (binary and three-level), gender, voice group, and nasality.

In dieser Studie wird untersucht, ob subjektive pädagogische Merkmale, insbesondere die Klangfarbe, aus akustischen Merkmalen in einem Längsschnittdatensatz von klassischen Sängern vorhergesagt werden können. Unter Verwendung einer institutionellen Datenbank mit abgeglichenen Bewertungen und Gesangsproben haben wir CNN- und MLP-Modelle auf Mel-Spektrogrammen und MFCCs trainiert, um verschiedene Stimmmerkmale zu klassifizieren. Während die CNN-Modelle eine begrenzte Aussagekraft hatten, erzielten die MLP-Modelle statistisch signifikante Ergebnisse bei der Vorhersage von Timbre (binär und dreistufig), Geschlecht, Stimmgruppe und Nasalität.

2 Introduction

Due to the existence of large-scale acoustic databases, deep learning models have been in use for the detection of human emotion and voice pathology for decades [1, 3, 4, 7]. The majority of papers in the literature analyze audio through conversion to mel-frequency cepstral coefficients (MFCCs) and subsequent training with a convolutional neural network (CNN) or a multilayer perceptron (MLP) [2]. A lack of large-scale pedagogically evaluated acoustic data has made such an approach impossible for classical singers up to this point.

The Hochschule für Musik Carl Maria von Weber Dresden has been collecting a database of subjective evaluations and acoustic recordings since 2002. At the beginning of studies students were evaluated by a pedagogue for timbre, resonance, nasality, and future aptitude for solo/choral singing, among other qualitative and quantitative tests. In that same year students recorded a set of seven vocal exercises, normally within a month of those subjective ratings. That set of seven exercises was repeated yearly until graduation. Although the ratings were not performed for the audio recordings themselves, one could expect a pedagogical evaluation to correlate with recordings made by that student within the first few months of study.

3 Methodology

Students recorded seven exercises yearly between 2002 and 2019 at the Hochschule für Musik Carl Maria von Weber Dresden. We calculated Mel spectrograms for the exercise "Manca sollecita" and used dynamic time warping (DTW) to align them to a representative sample, subsequently truncating/padding each to ensure identical length. A soprano singer rated as "certain soloist" at the beginning of studies whose recording differed less than a second from the median recording duration was chosen for the reference spectrogram. In order to find the optimal alignment, recordings were pitch shifted through all possible transpositions that could align them with the reference sample before the optimal time warping was extracted and applied to the original spectrogram. The process can be seen in Figure 1.

Three different neural network models were trained using the mel spectrograms from the initial year of study to predict labels for timbre, resonance, nasality, and solo/choir. As the resonance label was relatively balanced between "heady," "balanced," and "chesty," we retained the original labels. Label imbalances among the remaining variables led us to reduce timbre, nasality, and solo/choir to binary labels. As only a small subset of singers were identified as having a "dark" timbre at the beginning of studies, they were grouped with the "medium" timbre students resulting in either "medium/dark" or "bright" timbre. For nasality, all students identified as "somewhat nasal" and "strongly nasal" were grouped together for the resulting labels "nasal" and "neutral." For solo/choir, students had been labeled as "certain solo," "possible solo," "certain choir," and "possible, choir." Students labeled as "pedagogy" were not included in the analysis. The solo labels and choir labels were merged, resulting in either "solo" or "choir."

In order to decrease the importance of the six specific pitch levels present in the recordings, the data was augmented by including two variants of each spectrogram randomly pitch shifted by a semitone. All augmented and original spectrograms used in training data were time and frequency masked according to the SpecAugment method[6].

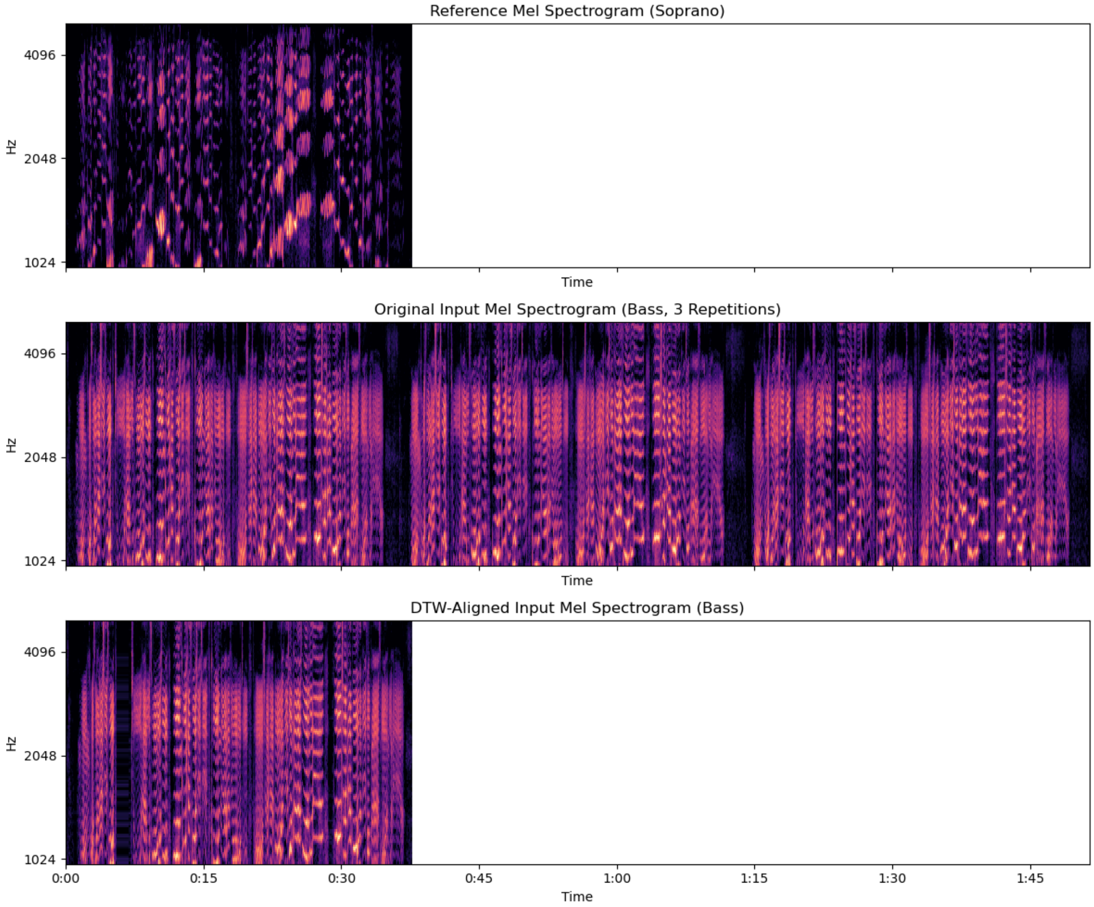

Figure 1: Upper panel: Reference mel spectrogram from soprano singer classified as certain soloist. Middle panel: Outlier recording of low-voiced male including three repetitions of sung task. Lower panel: Final dynamically time-warped and truncated version of low-voiced male recording used for analysis.

3.1 Initial CNN Model

The initial convolutional neural network (CNN) model was composed of the following layers using the PyTorch package:

1. The first convolutional layer used 16 filters of size 3 × 3 with padding to preserve input dimensions, enabling the model to extract local time–frequency features in the spectrogram. Each 3×3 filter computed a weighted sum over local time–frequency patches, enabling the model to detect recurring patterns regardless of their position in the input image. Zero-padding was used outside the image to detect behavior along the edges and retain the dimensions of the produced feature maps.
2. Following each convolution, a rectified linear unit (ReLU) activation function was applied element-wise to introduce non-linear transformations, allowing the network to model spectral–temporal relationships in the input image.
3. Each convolutional layer was followed by a 2 × 2 max pooling operation with stride 2, resulting in non-overlapping regions. This both reduced the spatial dimensions of the feature map and increased translational invariance due to time or pitch shifting.
4. A second convolutional layer with 32 filters of size 3 × 3, each spanning all 16 feature maps produced by the first layer (with a padding of 1), was applied to enable the network to learn higher-order spectral–temporal patterns from combinations of the previously extracted lower-level features.
5. Again, a ReLu function was applied element-wise to these 32 feature maps produced in the second convolution.
6. Another 2 × 2 max pooling operation was applied.
7. The convolutional feature maps were flattened into a 1D vector of 32 values.
8. Finally, a fully connected layer mapped the remaining features to the output classes.

3.2 MFCCs and Multi-Layer Perceptron

As a second approach, we calculated MFCCs for the same recordings. They were similarly dynamic time warped and truncated or padded to a reference recording to be an identical length. The CNN model was trained and compared to simulated null distributions.

Finally, the MFCCs were trained with a multi-layer perceptron (MLP) using PyTorch and evaluated for accuracy using simulated null distributions.

0. MFCCs (13 coefficients per frame) were flattened across time to form one-dimensional input vectors for the fully connected layers.
1. The flattened MFCC input was projected into a 256-dimensional feature space using a fully connected layer, where each of the 256 units (neurons) of the layer is computed from a linear combination of all MFCCs across time, enabling the network to learn global combinations of cepstral features.
2. A ReLU activation was applied element-wise to introduce non-linear transformations.
3. Dropout with a rate of 0.3 was applied to reduce overfitting by randomly setting 30% of units to zero during training.
4. A second fully connected layer reduced the representation from 256 to 128 dimensions, refining the learned cepstral feature representation.
5. A second ReLU activation was applied to this reduced feature space.
6. A second dropout layer with a rate of 0.3 was used for additional regularization prior to classification.
7. Finally, a fully connected layer mapped the 128-dimensional representation to n_{classes} output units corresponding to the target classes.

Training and Simulated Null Distribution

For both models we used k-fold validation for training with five folds. Both were trained using the Adam (Adaptive Moment Estimation) optimizer which incorporates gradient ”memory” (momentum) and per-parameter learning rates to standard stochastic gradient descent to avoid local optima and increase efficiency. Loss was computed using the categorical cross-entropy loss function for multi-class classification.

In order to determine a null distribution, we simulated training each of the previous models for each variable 30 times with randomly shuffled labels. Any statistically significant MFCC results were used to identify variables of interest to investigate with a further CNN model trained on mel spectrograms that had been used successfully for music genre classification[5].

3.3 Bottom-up Broadcast Neural Network on Mel Spectrograms

The Bottom-Up Broadcast Neural Network architecture consisted of the following steps using the Keras package in TensorFlow:

1. A two-dimensional convolutional layer (Conv2D) was implemented with 32 filters, a 3×3 kernel. Padding (”same”) and default strides of one ensured feature maps with identical spatial dimensions to the input.

2. Batch normalization was performed where the activations from each layer were normalized to have mean zero and variance one in order to stabilize training and speed up convergence. The scale and shift parameters were also stored so the normalization was reversible.
3. ReLU activation was applied to introduce nonlinearity to the model.
4. Max pooling with a window of 4×1 was applied to reduce temporal resolution by a factor of four while preserving frequency resolution, thereby providing invariance to small shifts in timing for acoustic events.
5. Three dense blocks were applied by sequentially stacking multiple multi-scale blocks. Each dense block consisted of parallel convolutional branches with 1×1, 3×3, and 5×5 filters, along with a pooling branch. Outputs from all branches were concatenated, broadcasting low-level features from earlier layers upward through the network. This bottom-up broadcasting ensured efficient gradient flow, encouraged reuse of previously identified features, and allowed higher layers to refine predictions using both fine-grained spectral details and increasingly abstract representations.
6. A transition block with batch normalization, ReLU activation, a 1×1 convolution, and 2×2 average pooling was applied to reduce both the number of channels and the spatial resolution of feature maps.
7. Steps 2 and 3 were repeated, with batch normalization followed by ReLU activation for nonlinearity.
8. At the end of the network, global average pooling was used to condense each feature map into a single representative value, producing a compact feature vector for classification.
9. Finally, the model output was generated through a softmax layer, implemented as a fully connected layer in which each output unit computed a linear combination of the features in the final vector. The softmax function then transformed these values into class probabilities, with the predicted label corresponding to the class with the highest probability.
10. The model was compiled using the categorical cross-entropy loss function, accuracy as the primary metric to monitor model performance during training, and the Adam optimizer for weight updates.

3.4 Feature Analysis

Feature analysis was performed with gradient-weighted class activation mapping (Grad-CAM) which identified which time-frequency regions most influenced the model's classification.

Table 1: MLP Classification Results

Variable	Accuracy	Sensitivity	Specificity	p-value (BH)
`Voice Group`	0.99	0.98	0.98	0.00
`Gender`	0.93	0.94	0.94	0.00
`Timbre (binary)`	0.72	0.71	0.71	3.49×10^{-13}
`Timbre (3 levels)`	0.68	0.47	0.80	6.91×10^{-5}
`Voice Subtype`	0.63	0.47	0.79	3.31×10^{-4}
`Vital Capacity (3 levels)`	0.63	0.44	0.75	0.01
`Nasal`	0.62	0.58	0.58	0.04

4 Results

4.1 CNN Trained on Mel Spectrograms

Directional improvements were observed for predicting the variables of interest above chance, but none of the results reached statistical significance.

4.2 CNN Trained on MFCCs

Directional improvements in prediction accuracy were again observed. Timbre classification achieved a mean accuracy of **0.64** over five folds, which was statistically significant (**p = 0.00003**). Other variables, including nasality, resonance, and solo/choir classification, showed no statistical significance.

4.3 MLP Model Results on MFCCs

Table 1 shows variables with statistically significant accuracies beyond the null distribution.

Timbre remains the most compelling result (See Figure 2). High accuracies for voice group, gender, and voice subtype are to be expected and serve to validate the model's baseline performance. Vital Capacity (below average/average/above average) and nasality were also predicted significantly above chance.

4.4 Bottom-up Broadcast Neural Network on Mel Spectrograms

This model trained on spectrograms showed and acuracy of 64.8%, a slight improvement compared to the simple CNN trained on MFCCs. The summary metrics for the five-fold cross-validation are shown in Table 2.

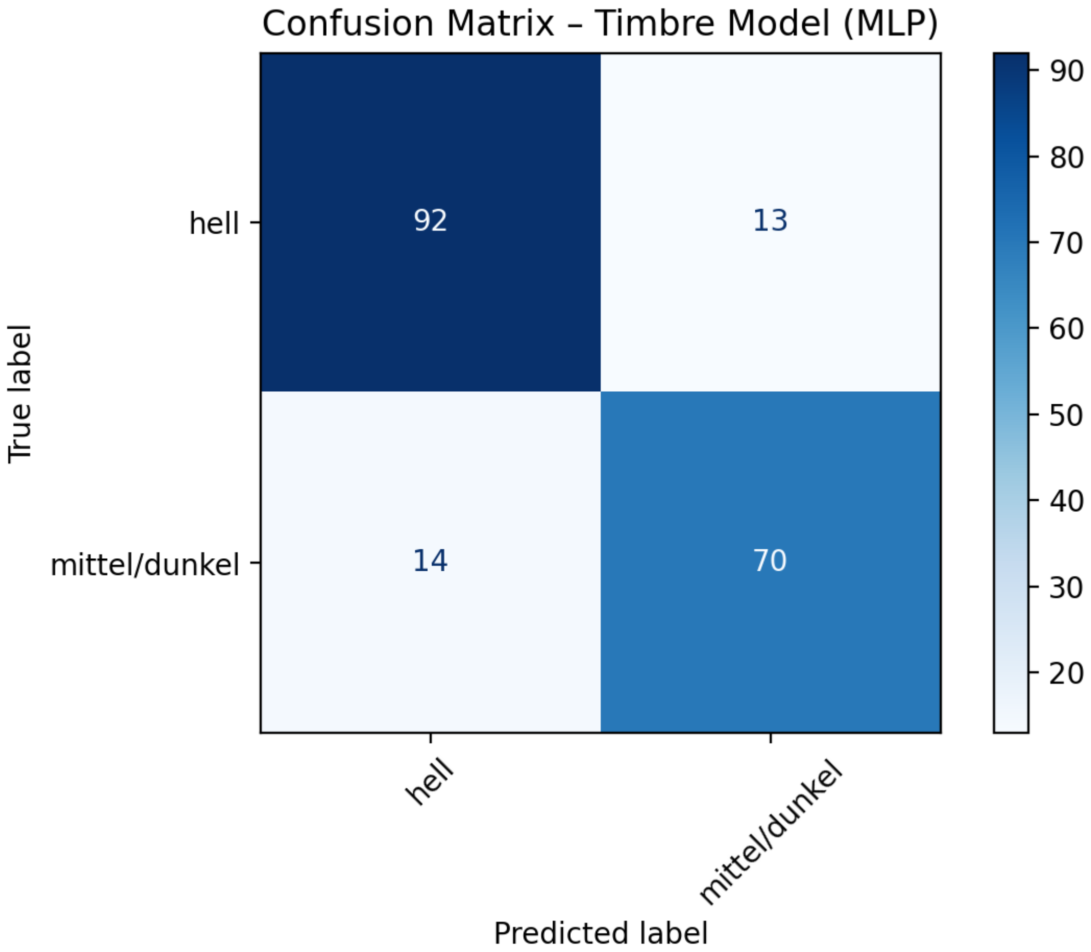

Figure 2: Confusion matrix for timbre as predicted by the MLP. Though 72% accuracy is lower than values reported in the literature for vocal pathology and emotion, it is encouraging given recordings that were only tangentially related to the subjective ratings from the pedagogue.

Metric	Accuracy	Loss	Macro F1	Weighted F1
Value	0.648 ± 0.013	0.640 ± 0.018	0.683	0.664

Table 2: Summary of k-fold cross-validation results for the bottom-up broadcast neural network. Accuracy and Loss are reported as mean ± standard deviation across folds; Macro and Weighted F1 scores are reported as single aggregated values.

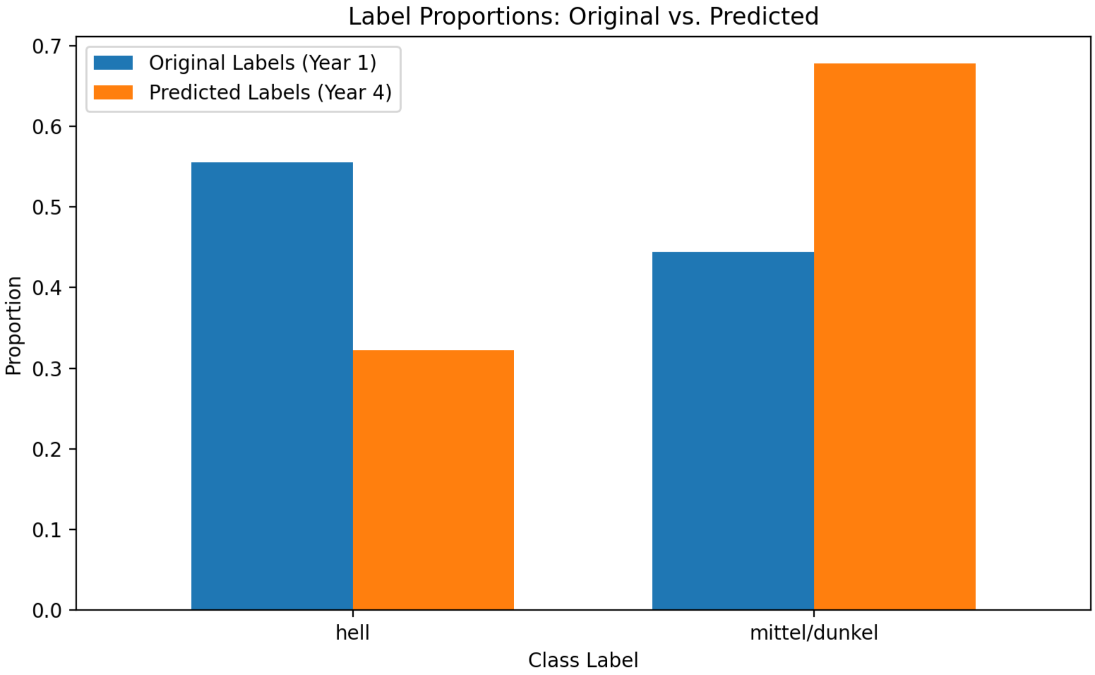

Figure 3: Application of the MLP model to predict timbre in year four recordings resulted in higher proportions of medium and dark voices. This suggests that singers' voices become more balanced or dark with training and maturity compared to the beginning of their studies.

5 Discussion

5.1 Interpretation of MFCC Timbre Results

Although the classification accuracy for timbre using MFCCs was moderate, the results suggest that meaningful signal exists in MFCC features reflecting perceptual timbre judgments by pedagogues. Unfortunately, the use of MFCCs limits interpretability, as these features already serve as an abstraction for the acoustic characteristics that underlie perceptual timbre.

5.2 Longitudinal Analysis: Year 4 Predictions

Applying the trained model to recordings made during the final year of studies revealed a higher proportion of predicted "medium/dark" voices (See Figure 3). This distributional shift suggests a perceptual darkening or balancing of timbre over the course of training, aligning with descriptive findings from an independent study (in press) that suggested a similar pattern based on divergent spectral energy metrics over time.

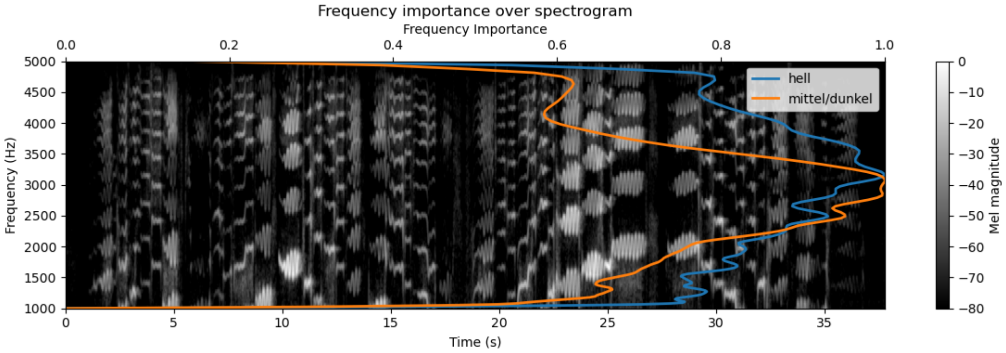

Figure 4: Peak importance in the frequency spectrum was assigned in the region between 2.8 and 3.2 kHz with a local maximum around 2.5 kHz. Treble voices also had a local maximum around 3.5 kHz.

5.3 Interpretation of Bottom-up Broadcast Neural Network Timbre Results

For the bottom-up broadcast neural network trained on mel spectrograms, modest accuracy further emphasizes that predicting pedagogue's timbre labels from secondary recordings is challenging. Nevertheless, the model demonstrates that meaningful signal exists in mel spectrograms that can reflect perceptual timbre judgments.

Frequency importance showed local and global maxima between 2.5-3.5 kHz for both classes (see Figure 4), which is in keeping with the pedagogical literature.

Time importance identified phonemes of interest as ”bright” (/i/, /e/) and ”dark” (/u/, /o/) vowels in the low-medium range (see Figure 5). The highest differences between class attributions seem to appear in the 2.5-4 kHz region for the ”bright” vowels and between 1-3 kHz for the ”dark” vowels (see Figure 6).

5.4 Class Imbalance

Even for the more balanced variables, imbalances in subgroupings may have influenced model accuracy. For example, ”bright” and ”medium/dark” ratings are represented disproportionately among treble and non-treble singers as can be seen in Table 3. The features learned for the models are likely influenced by these imbalances in voice group which is much easier to detect due to the differing octaves of the sung exercises and resulting harmonic spacing in the 1-5 kHz window of our trained spectrograms.

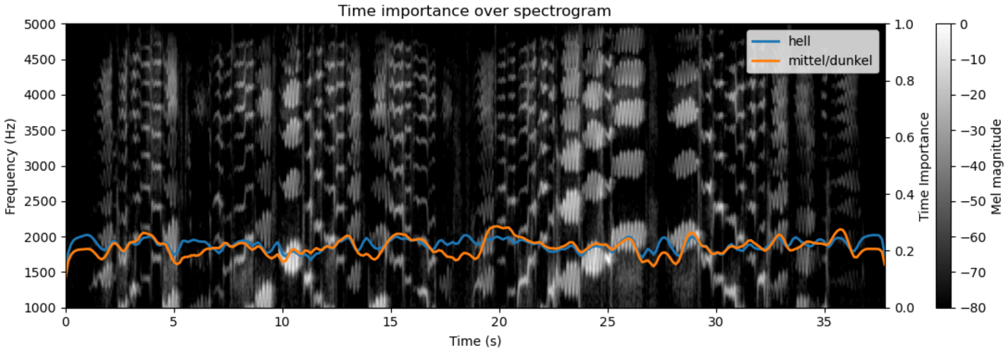

Figure 5: Time importance was assigned local maximums for both classes around 3.6 s in the middle of the first sung phrase, around 13 s at the end of a descending phrase on two /i:/ vowels, around 15.5 s during the phoneme /lje:ve:/ in medium range, and then for the highest sung note of the exercise at 26 s and 28.8 s for an /i:/ vowel and /a:/ vowel respectively. Peak importance occurred for medium/dark voices around 20 seconds, when the singers were singing the vowels /a: e:/ for the lowest notes of the sung exercise. Bright voices had localized importance around 5.7 s and 17.7 s on /u:/ and /o:/ vowels on the lowest note of the sung exercise.

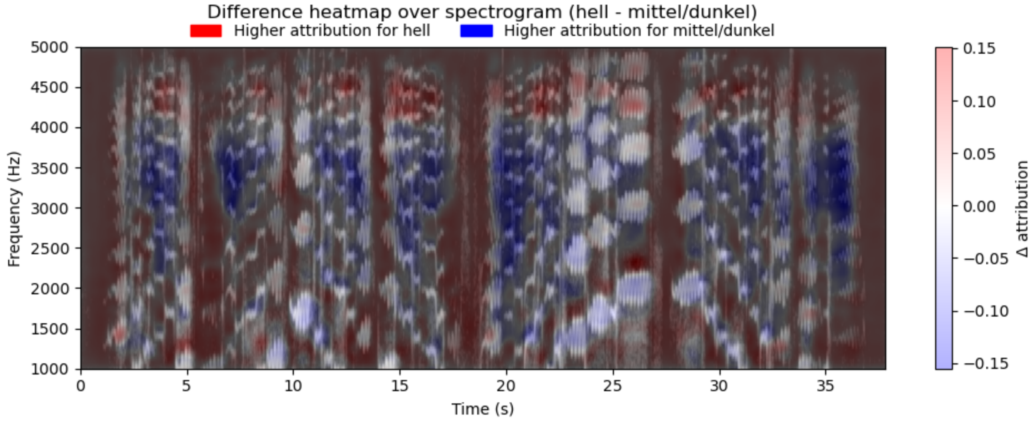

Figure 6: Red represents regions more important for distinguishing bright voices while blue represents regions more important for distinguishing medium/dark voices. Largest sections of blue occur between 2.5-4 kHz for /i:/ and /e:/ vowels in low-medium range. Largest sections of red occur between 1-3 kHz for /u:/ and /o:/ vowels in low-medium range.

Voice Group	**Timbre**	**Count**
Total	hell	309
	mittel/dunkel	250
Sop/Mezzo/Alt	hell	231
	mittel/dunkel	121
Ten/Bar/Bass	hell	78
	mittel/dunkel	129

Table 3: Distribution of timbre with subdivision by voice group.

6 Conclusion

Three different neural networks showed varying levels of accuracy predicting subjective pedagogical labels from mel spectrograms and MFCCs. Feature analysis suggested that certain vowel frequency regions may be instructive in classifying voices based on timbre. Future work will apply these trained models to a post-COVID dataset to assess generalizability across student populations and conservatory eras.

References

[1] A. A. Alnuaim, M. Zakariah, P. K. Shukla, A. Alhadlaq, W. A. Hatamleh, H. Tarazi, R. Sureshbabu, and R. Ratna. Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. *Journal of Healthcare Engineering*, 2022(1):6005446, 2022.

[2] J. Barlow, Z. Sragi, G. Rivera-Rivera, A. Al-Awady, Ü. Daşdöğen, M. S. Courey, and D. N. Kirke. The use of deep learning software in the detection of voice disorders: a systematic review. *Otolaryngology–Head and Neck Surgery*, 170(6):1531–1543, 2024.

[3] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda. Automatic detection of voice impairments from text-dependent running speech. *Biomedical Signal Processing and Control*, 4(3):176–182, 2009.

[4] J. I. Godino-Llorente and P. Gómez-Vilda. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384, 2004.

[5] C. Liu, L. Feng, G. Liu, H. Wang, and S. Liu. Bottom-up broadcast neural network for music genre classification. *Multimedia Tools and Applications*, 80(5):7313–7331, 2021.

[6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[7] G. Schlotthauer, M. E. Torres, and M. C. Jackson-Menaldi. A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification. *Journal of voice*, 24(3):346–353, 2010.