Deep Learning Approaches for Sung Vowel Classification

Parker Carlson[1,2(✉)] and Patrick J. Donnelly[2]

[1] UC Santa Barbara, Santa Barbara, USA
parker_carlson@ucsb.edu
[2] Oregon State University, Corvallis, USA
patrick.donnelly@oregonstate.edu

Abstract. Phoneme classification is an important part of automatic speech recognition systems. However, attempting to classify phonemes during singing has been significantly less studied. In this work, we investigate sung vowel classification, a subset of the phoneme classification problem. Many prior approaches that attempt to classify spoken or sung vowels rely upon spectral feature extraction, such as formants or Mel-frequency cepstral coefficients. We explore classifying sung vowels with deep neural networks trained directly on raw audio. Using VocalSet, a singing voice dataset performed by professional singers, we compare three neural models and two spectral models for classifying five sung Italian vowels performed in a variety of vocal techniques. We find that our neural models achieved accuracies between 68.4% and 79.6%, whereas our spectral models failed to discern vowels. Of the neural models, we find that a fine-tuned transformer performed the strongest; however, a convolutional or recurrent model may provide satisfactory results in resource-limited scenarios. This result implies that neural approaches trained directly on raw audio, without extracting spectral features, are viable approaches for singing phoneme classification and deserve further exploration.

Keywords: Sung Vowels · Phoneme Classification · Raw Audio · Automatic Speech Recognition · CNN · LSTM · Transformer · VocalSet

1 Introduction

The ability to discriminate between different phonemes is a critical and often studied aspect of automatic speech recognition (ASR) systems. While most modern ASR systems perform strongly for well-resourced natural languages, performance tends to degrade with lower-resource languages. Further, many ASR systems tend to struggle with variations in pronunciation and the rate of speech [16]. These variations in delivery are further exaggerated during singing. Moreover, the numerous fundamental differences between speech and song present many challenges when applying conventional ASR systems to singing.

Parker Carlson–Work done while at Oregon State University.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
C. Johnson et al. (Eds.): EvoMUSART 2024, LNCS 14633, pp. 67–83, 2024.
https://doi.org/10.1007/978-3-031-56992-0_5

A typical ASR pipeline often contains an acoustic front-end which extracts phonemes or other acoustic features from the audio signal, an acoustic model which proposes likely words based on the observed phonemes, and a language model which limits word choice based on the grammatical structure of the language. Before deep learning approaches became widespread, many ASR systems relied on a set of spectral features, such as formants or Mel-Frequency Cepstral Coefficients (MFCCs), which often were modeled with hidden Markov models [8]. Researchers have also trained deep neural networks on MFCC features or raw audio, including bidirectional long short-term memory networks, convolutional neural networks, and more recently, transformer models [2,10,16,28].

While end-to-end neural methods for ASR usually outperform phoneme-based ASR systems for high-resource languages [12], they often struggle to extend to low-resource languages given the absence of sufficient training data [17]. Likewise, researchers investigating singing, such as the task of automatic lyric transcription, also suffer from a lack of annotated training data. When sung, vowels take importance over consonants and are sustained much longer than their spoken counterparts [3]. Moreover, sung vowels inherently have different acoustic properties than spoken vowels [31]. For example, the formant frequencies of sung vowels varies, with the direction and magnitude of the change depending on the vowel and formant [32]. In addition, differences between male and female voices are exaggerated during singing. For instance, the "singer's formant", a clear formant around 3000 Hz that produces a ringing timbre is more common in trained male singers than female singers [33,37]. Furthermore, singers may alter the pronunciation of words or elide certain phonemes, exacerbating the challenges of applying spoken ASR techniques to vocal lines. These differences can be clearly observed in Fig. 1, a comparison of a sung and spoken vowel.

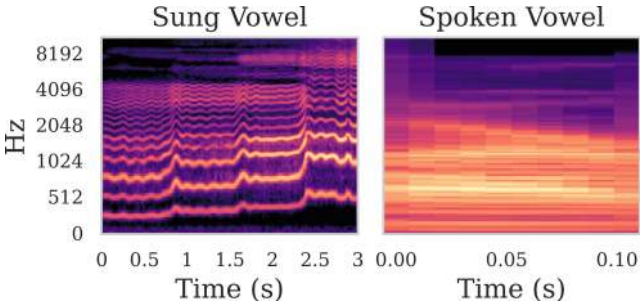


Fig. 1. Sung and sung vowels are significantly different in duration, pitch, and harmonics. Mel-spectrograms visually demonstrate the difference between a spoken and sung vowel `a`. The spoken vowel is taken from the TIMIT corpus [9]; the sung vowel is sung by `female1` with vocal technique `belt`.

In this work, we investigate the use of deep neural networks to classify sung vowels directly from raw audio. We replicate experiments from VocalSet [38] and

extend them to investigate vowel classification, evaluate the efficacy of three neural models for sung vowel classification, and examine their performance relative to two spectral models. To our knowledge, this is the first study to to explore neural methods trained on raw audio to the sung vowel classification problem.

2 Related Work

Although most research in the area of automatic phoneme recognition has focused on speech, a few researchers have applied these methods to sung vocal lines. In this section, we review relevant literature in the areas of automatic vowel classification, automatic lyric transcription, and pre-trained musical transformers.

2.1 Automatic Vowel Classification

The task of automatic vowel classification can be viewed as a subset of the phoneme classification problem. Unfortunately, there do not exist many datasets of singing examples that contain annotations at the phonemic level. Therefore, many phoneme classification systems are trained on spoken speech data that may be synthetically altered or augmented to closer resemble aspects of singing [22,23,25]. As more datasets, such as *VocalSet* [38] and *DAMP Sing!* [30], are released, researchers will be less reliant on augmented speech alone.

Vowel classification has been studied in numerous context across many different languages [10,15,20]. When investigating this task, researchers may examine spectral features, such as spectrograms or analysis of vocal formants, either labeled by experts or automatically generated [5,13,21]. Other times, researchers rely upon acoustic features, such as MFCCs, zero-crossing rate, and energy. And these features have been explored through a variety of models, such as k-nearest neighbor, linear and quadratic discriminant analysis, and support vector machines [13,20,21]. However, there remain many unsolved problems to explore. For example, a study by Jha and Rao reported a 10% difference in accuracy between male and female voices when classifying five sung Hindi vowels [15]. Furthermore, vowel classification varies significantly by language, on account of the differing number of vowels and language-specific pronunciation.

2.2 Automatic Lyric Transcription

Automatic lyric transcription (ALT) is an emerging field that intertwines approaches from the fields of speech recognition and music information retrieval. The two primary tasks of ALT are lyric transcription and lyric alignment. In the first task, researchers seek to transcribe sung vocal lines to text. In the second task, researchers attempt to align this transcription to precise time-stamps, in order to determine what is being sung when. Although the field has not received the same attention and resources as ASR, recent work has advanced understanding of both lyric transcription [4,6] and lyric alignment [11,27,34].

There have been many different approaches to ALT, with some systems focusing on classifying sung speech in a monophonic setting [6] and others tackling the more challenging polyphonic setting [11]. In one such study, Mesaros and Virtanen found that their hidden Markov model adapted to gender showed improvements over the baseline model for the lyric transcription task [25].

While much of the previous work in ALT has focused on word-level classification and alignment, there has been some work on phoneme classification and alignment. Spectrograms have been used to aid in phoneme alignment with sung speech [34], as well as classifying words sung in Thai [18]. Additionally, MFCCs have been used to classify sung Hindi vowels with Gaussian mixture models [15].

2.3 Pre-trained Musical Transformers

For many text classification problems, the dominant paradigm is a two-stage process that begins with pre-training a large transformer that captures a general understanding of a language, followed by additional training to fine-tune it for a domain-specific problem [7]. Due to the success of pre-training and fine-tuning transformers for text-based tasks, there has been many recent approaches on developing a similarly capable model for acoustic, speech, and musical understanding [14,36]. One such state-of-the-art model, MERT [24], introduced musical pre-training objectives and applied them to HuBERT [14], a popular model for speech representation. The authors found their model generalized well to fourteen tasks that required a broad understanding of musical features. In this study, we fine-tune MERT to investigate the task of sung vowel classification.

3 Methodology

In this section, we discuss the dataset used in this study and provide a detailed description of our model architectures for our vowel classification experiment.

3.1 Dataset

To explore this task, we utilize VocalSet [38], a dataset of 10.1 h of recordings of twenty professional singers performing a variety of vocal techniques on five different Italian vowels (`/a/`, `/e/`, `/i/`, `/o/`, `/u/`). Henceforth, we refer to these vowels as `a`, `e`, `i`, `o`, and `u` respectively, following convention established in VocalSet. The dataset contains examples of eleven male and nine female singers, comprising the seven different common voice types (bass through soprano).

We preprocess our dataset following the approach of Wilkins et al. [38], making informed choices for any missing details. For each recording, we automatically remove silence at the beginning, middle, and end of recordings using Sox.[1] Next, we split the recordings into three second examples sampled at 44.1 kHz. Because of the presence of short (less than one second) examples, we opt to discard any

Table 1. Examples are nearly uniformly distributed by vowel in our dataset.

Vowel	Examples
`a`	1148
`e`	1147
`i`	1149
`o`	1190
`u`	1187

examples with a duration less than three seconds of audio. Lastly, we standardize each example to have a zero mean and unit variance.

Although VocalSet provides suggested train and test splits [38], we choose to evaluate our models using k-fold cross-validation in order to test our models on every example. We randomly partition the singers into four folds of five singers, while ensuring that each fold has an approximately equal number of male and female singers.[2] We exclude spoken excerpts in order to focus on sung vowels, as well as the `trillo` technique because each example is shorter than three seconds in duration. These exclusions represent less than 2.5% of all available data. After preprocessing our dataset we have 5821 examples, totaling 4.85 h. The distribution of examples per vowel is shown in Table 1.

VocalSet Replication. In their original paper, Wilkins et al. explored VocalSet in the context of vocal technique classification and singer identification [38], but curiously, they did not explore the sung vowel classification task. To contextualize our results and provide a vowel classification baseline for VocalSet, we replicate their convolutional model and evaluate on the tasks of vocal technique classification and singer identification explored in their initial publication.

3.2 Neural Models

We select three neural models of varying architectures to evaluate in the task of sung vowel classification. We evaluate a convolutional model, a recurrent model, and a pre-trained transformer. Each of these model architectures have been successfully applied to raw audio analysis in various contexts [1,24,26].

Convolutional. For our convolutional model, we select a three layer convolutional neural network (CNN) that is identical to Table 1 of Wilkins et al. [38] in order to provide an identical baseline for our vowel classification results. We direct readers to the original VocalSet paper [38] for a detailed model description.

[1] https://sox.sourceforge.net/sox.html.
[2] Each split of singers is either (two male, three female) or (three male, two female).

Recurrent. To consider a recurrent neural model of similar depth to our convolution model, we stack three Bidirectional Long Short-Term Memory (BiLSTM) layers with eight units each. After each of the first two BiLSTM layers, we apply a 20% dropout during training. The last BiLSTM layer is followed by a dense layer with 32 units and a ReLU activation function, then a final dense layer with five units and a softmax activation. Both dense layers have a L2 regularization of `1e-3`. We train this model for 200,000 batches of size 64 using RMSProp [35] with a learning rate of `1e−4` and momentum of 0.99 using TensorFlow.[3]

Transformer. As our transformer model, we consider MERT, a state-of-the-art pre-trained acoustic transformer that has been shown to generalize well to a variety of musical tasks [24]. We begin from `MERT_v0_public`, a 95M parameter version of MERT pre-trained on `music4all`, a dataset of 910 h of publicly available music [29]. MERT is pre-trained with acoustic and musical masked language model objectives, which predict k-means clusters from log-Mel spectral features and reconstruct masked audio from a constant-Q transform respectively. For each of the transformer layers, we take the average of all embeddings across the time dimension, creating thirteen total embeddings that capture the musical information of the entire example. We empirically consider each embedding as well as combinations of them, following the authors' suggestion,[4] finding that a sum of the first seven embeddings yields the best results. This combined embedding is used as input to a dense layer with 256 units and ReLU activation, followed by a dense layer with five units and a softmax activation function. We fine tune the entire model for five epochs using the Adam optimizer [19], cross-entropy loss, and a learning rate of `3e−5` using PyTorch[5] and Transformers.[6]

3.3 Spectral Models

We first explore two existing spectral models that have been successfully applied to vowel classification to serve as a baseline for our raw audio neural models.

Formant Model. We follow recent work from Korkmaz et al. to classify spoken Turkish vowels using a formant-based classifier [20]. We divide each example into frames of 512 samples (11.6 ms) with a 256-sample overlap (5.8 ms) and apply a Hamming window to each frame. Next, we apply linear predictive coding with eight filters to extract three formants per frame. Frames in which no formants were found are discarded (21.2% of data), and frames missing the second or third formant have these values imputed as zero. We use these features to train a k-nearest neighbor classifier, with $k = 1$, using the Manhattan distance. We do not aggregate classification results from the frames back to three-second examples.

[3] tensorflow.org.
[4] github.com/yizhilll/MERT/blob/main/scripts/MERT_demo_inference.py.
[5] pytorch.org.
[6] github.com/huggingface/transformers.

MFCC Model. We follow Jha and Rao's approach to classify sung vowels using MFCCs and Gaussian mixture models (GMM) [15]. We split the audio into 1024 sample (23.2 ms) frames with a 512 sample overlap, resulting in 87 frames per second. From each frame, we extract 13 MFCCs. For each vowel, we fit a single mixture GMM with a full covariance matrix. Classification is performed by maximizing the likelihood of the test vectors with respect to each model. Each example is assigned to a vowel according to the minimum Mahalanobis distance between the MFCC vector and each GMM. Following Jha and Rao [15], an example is considered to be correctly classified if more than half of the MFCC vectors are assigned to the correct vowel. If fewer than half of the MFCC vectors predict the same vowel, the example is classified as an error.

4 Results

In this section we discuss the findings of the experiments outlined in the previous section. First, we briefly examine the performance of the convolutional model on the two VocalSet replication tasks. Next, we thoroughly compare the performance of the five vowel classification models in a variety of settings.

Table 2. Results for our VocalSet replication by task. Top-1 Accuracy was not reported in [38]; MERT only reported Top-1 Accuracy [24]. `M.` and `F.` denote the accuracy of classifying male and female singers, respectively.

Model	Prec.	Recall	Accuracy				
			Top-1	Top-2	Top-3	M.	F.
Vocal Technique							
VocalSet [38]	0.676	0.619	-	0.801	0.867	-	-
Conv. (Ours)	0.628	0.607	0.619	0.753	0.814	-	-
MERT [24]	-	-	0.756	-	-	-	-
Singer ID							
VocalSet [38]	0.473	0.516	-	0.638	0.700	0.684	0.351
Conv. (Ours)	0.698	0.660	0.673	0.808	0.863	0.701	0.632
MERT [24]	-	-	0.780	-	-	-	-

4.1 VocalSet Replication

Our approach underperforms on the vocal technique classification task and outperforms on the singer identification task, relative to VocalSet [38] (Table 2). While a discrepancy in performance between male and female singers still exists, it is less than observed by Wilkins et al. During dataset preprocessing, we observed that our dataset sizes to be slightly different than reported by VocalSet. We estimate to have approximately 22% more data for our vocal technique classification task, and 38% less data for our singer identification task. We attribute

these differences in dataset size to differing parameters in preprocessing. We follow Wilkins et al. in training for a fixed number of batches, rather than epochs. As such, when batch size is fixed, the number of epochs is inversely proportional to the number of examples. Because our vocal technique classification underperforms compared to Wilkins et al., despite having more training data, and conversely our singer identification outperforms with less data, our results suggest that our models are undertrained. Nevertheless, our results are similar to those observed by Wilkins et al., and we continue to our vowel classification task.

4.2 Sung Vowel Classification

Our results for the vowel classification task are presented in Table 3. We found all neural models to achieve accuracies above 68%. However, we found that both the formant model and MFCC model to failed to converge, demonstrating accuracies corresponding to chance. We note that the MFCC model predicted the vowel `u` for 99.5% of examples, and the formant model's predictions appear to follow a uniform distribution. These results underscore the challenge of the sung vowel classification task. Existing spectral methods for spoken vowels failed to generalize, and existing spectral methods for sung vowel classification failed to replicate. On account of the inability of our spectral models to classify sung vowels, in our subsequent analysis, we focus only on the three neural models.

Table 3. Our results on the vowel classification task. Because our spectral models fail to converge, we do not evaluate male or female accuracy for these methods.

Model	Precision	Recall	Accuracy		
			Total	Male	Female
Neural					
Convolutional	0.723	0.719	0.718	0.749	0.674
Recurrent	0.685	0.685	0.684	0.714	0.640
MERT	0.799	0.797	0.796	0.827	0.752
Spectral					
Formants	0.200	0.200	0.201	-	-
MFCCs	0.041	0.199	0.203	-	-

Of the three neural models, we find that MERT performed best, outperforming both our convolutional and recurrent models with an accuracy of 79.6%. While our convolutional model generally outperformed our recurrent model, we find that our recurrent model performs better in classifying the vowel `a` as shown in Fig. 2. Furthermore, for certain singers and vowels, our recurrent model also outperforms our convolutional model, as shown in Figs. 6 and 7.

Across all vowels, MERT demonstrated the best performance. However, it was not universally better. In some cases, MERT performed notably worse than

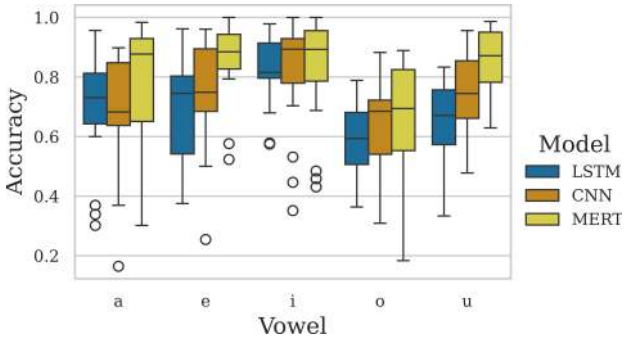


Fig. 2. The neural models performed decently, although all models had a high variance in accuracy across singers and vowels. Overall, MERT performed best. Each data point represents the model's performance on a single singer.

both other models, such as `male4`'s `i`, `female2`'s `o`, and `female3`'s `o`, shown in Figs. 6 and 7. These drops in performance suggest that an ensemble of neural methods may perform well. However, in some cases, all models struggled, such as `male3`'s `a` and to a lesser degree `female5`'s `o`, also shown in Figs. 6 and 7.

One important consideration is the disparity of computational resources required by the different models. Our convolutional and recurrent models are of similar sizes, with approximately 112,000 and 70,000 parameters respectively, whereas MERT has 95 million parameters. While pre-trained transformers have attractive properties, for instance, the ability to generalize to new tasks with minimal training, the vast difference in model size compared to our traditional neural models potentially makes them a poor choice in computationally-limited scenarios, such as real-time phoneme classification systems.

To varying extents, we find that all three models made similar errors, shown in Fig. 5. Like Jha and Rao [15], we find our models struggle to distinguish between the back vowels `o` and `u`. We also notice that our models struggle to distinguish between the `a` and `o` vowels. To a lesser degree, our models struggle with the central vowels `e` and `i` as well. In addition to these shared mistakes, the recurrent model also occasionally misclassifies `u` as the vowel `i`.

Some of these particular confusions are perhaps to be expected. For the spoken pronunciation of `e` and `i`, the placement of the tongue is more closed and towards front compared to the other vowels. On the other hand, for the pronunciation of `a` and `o`, the tongue is more open and towards the back. It is likely that these similarities between vowels are further exaggerated when sung.

Our models demonstrate similar relative performance in classifying vowels across most vocal techniques as shown in Fig. 3. Notably, they struggle more with classifying `inhaled` vowels and vowels performed with the `lip_trill` technique. It is perhaps unsurprising that these models struggled to identify vowels performed with these two extended techniques. The `inhaled` vowels are quite breathy, which serves as additional noise that muffles the vowel. Vowels performed with a `lip_trill` have a unique timbre that may dampen formants and

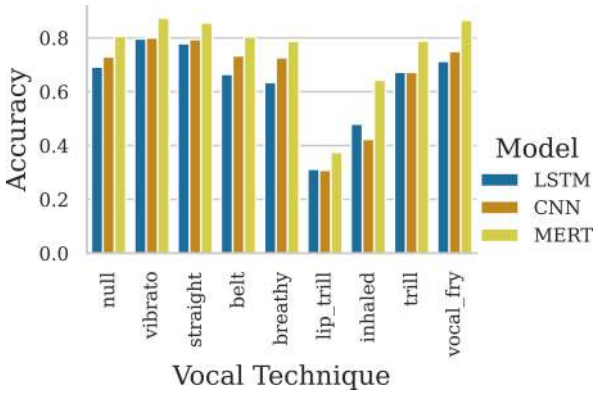


Fig. 3. Certain vocal techniques are more challenging for all models. Vowel classification accuracy was measured for each model and vocal technique. Portions of the dataset that have not been assigned a vocal technique are represented by "null", for instance the techniques `fast_forte` and `slow_forte`.

other acoustic features that the models rely upon to recognize the vowel. This timbre is shown visually in Fig. 4. Anecdotal evidence from the authors suggests that it is difficult for humans to audibly discern vowels sung with the `lip_trill` technique, but this merits further empirical investigation.

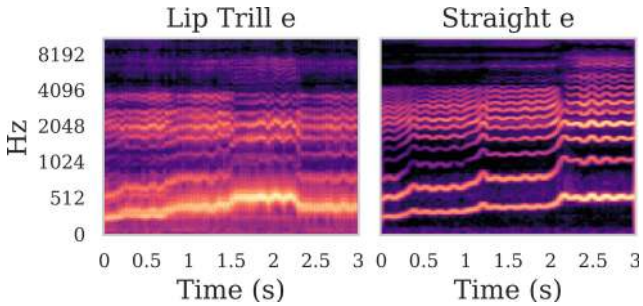


Fig. 4. An example of the vowel `e` performed with the `lip_trill` and `straight` vocal techniques by `female1`. The `lip_trill` vowel is both noisier and has less defined harmonics compared to the `e` performed with the `straight` technique.

We notice that our models perform approximately 10% better in terms of accuracy on classifying vowels sung by male singers (58.7% of the dataset) compared to female singers (41.3%). This replicates findings from Jha and Rao, who also observed a 10% difference in accuracy between male and female singers [15], though their dataset was gender-balanced. It remains unclear if our results are caused by the imbalance of the training data in VocalSet, or because of the differences in frequencies between the male and female singing voice.

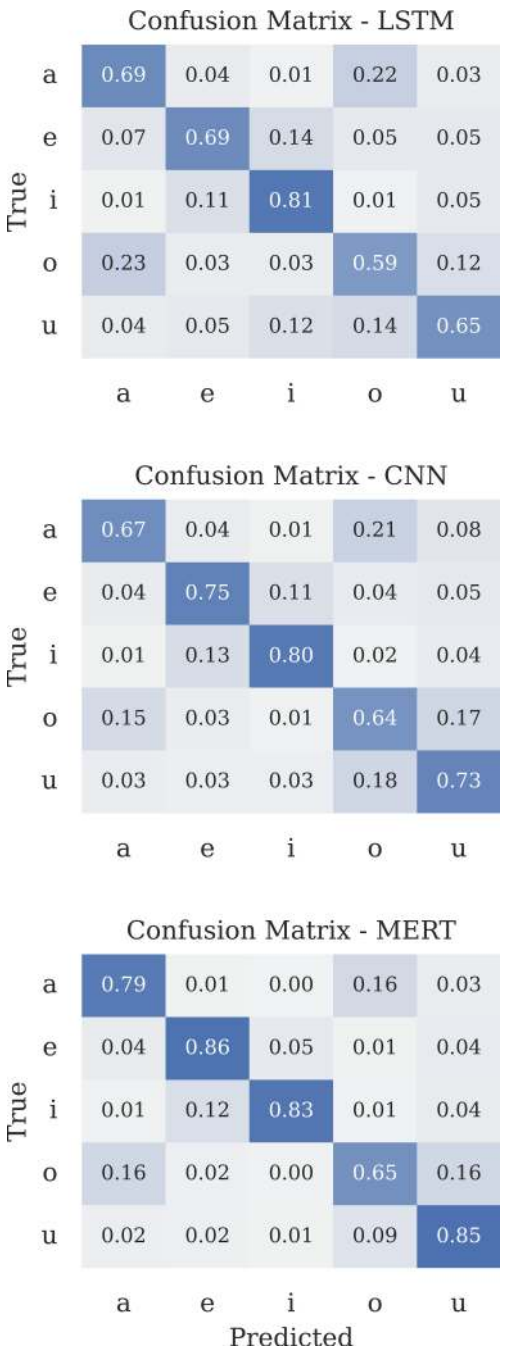


Fig. 5. Each model made similar types of errors as shown in the confusion matrices above. The relative frequency of each error varies by vowel and model.

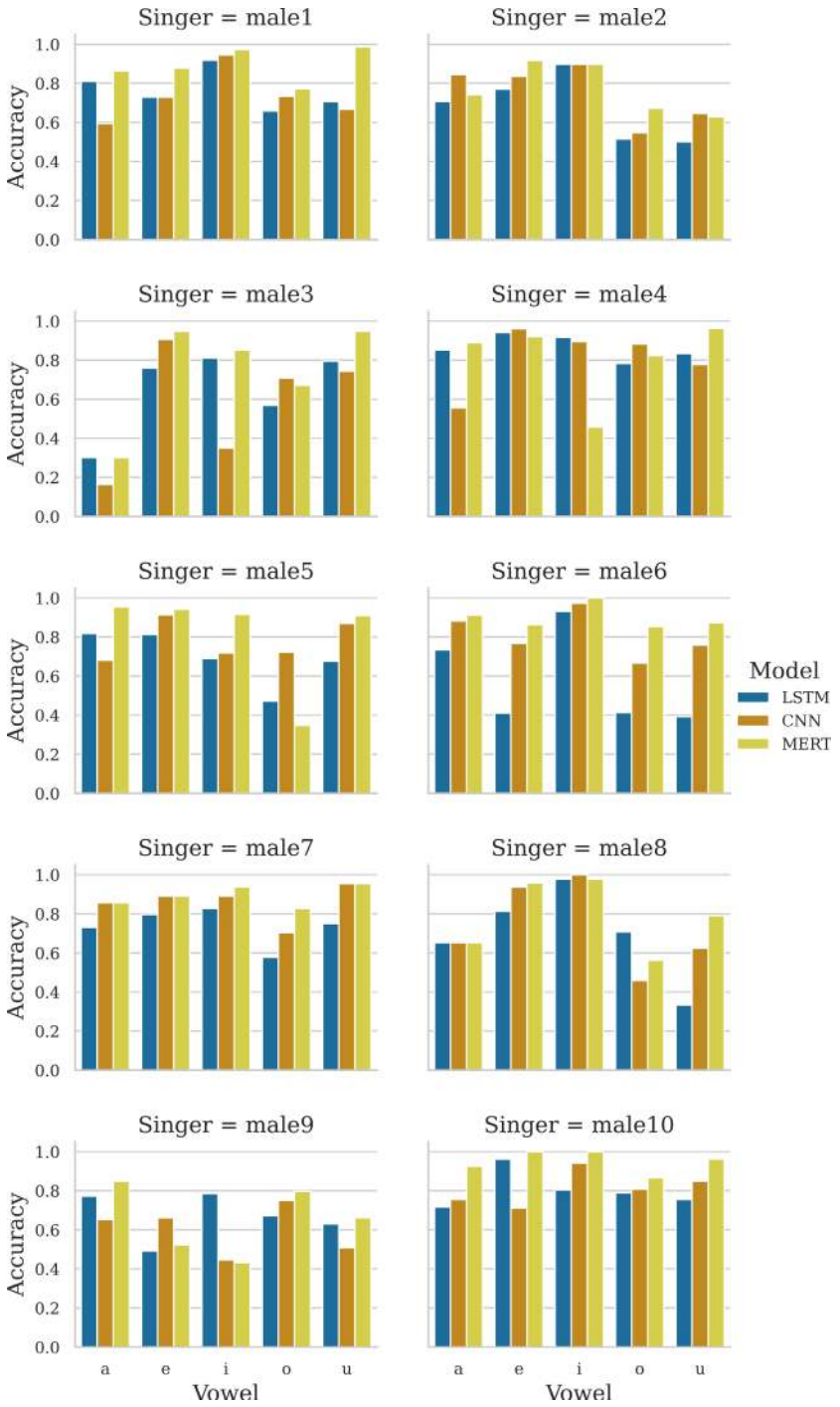


Fig. 6. Our model accuracy for each singer and vowel. Certain singers and vowels were significantly more challenging for some or all models.

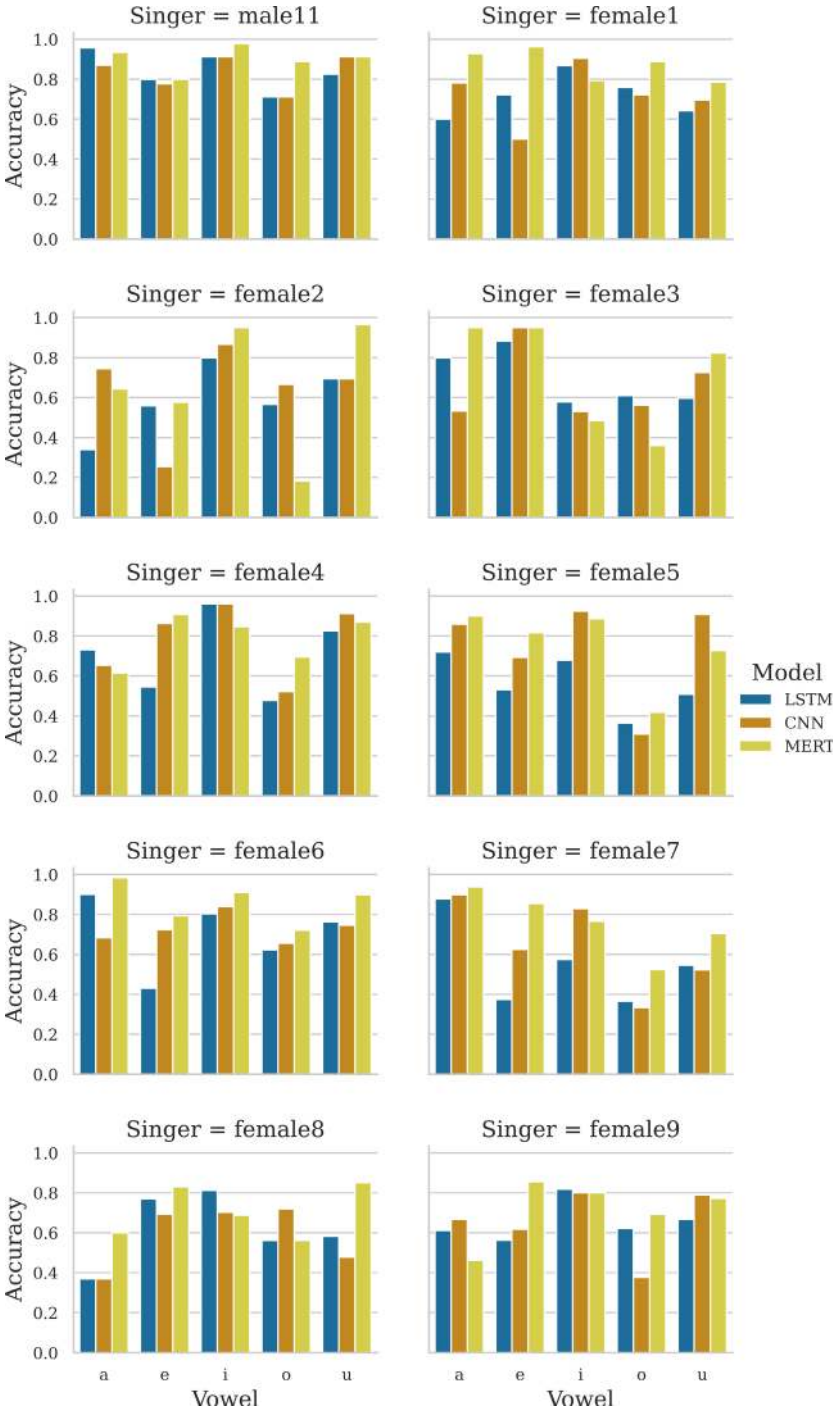


Fig. 7. Our model accuracy for each singer and vowel. Certain singers and vowels were significantly more challenging for some or all models (continued).

5 Conclusion

In this work, we demonstrate neural networks' potential to classify sung vowels in a monophonic vocal settings. By training on raw audio alone, we avoid time-consuming spectral analysis and feature extraction operations which complicate real-time analysis and prohibit use in live systems. Although these results are encouraging, there remain many avenues for improvement and further research.

As future work, we plan to explore additional models, to improve both accuracy and efficiency. We will examine neural models trained on spectral features (MFCCs), as well as an ensemble of neural models. We will also test larger convolutional and recurrent models in the 1M to 10M parameter range to understand the impact of model size, and to investigate if we can achieve comparable performance as MERT with a moderately sized model. We also intend to compare against other state-of-the-art audio and musical transformers.

One major limitation of our approaches for sung vowel classification is the lack of annotated singing data. We only evaluate our models on five Italian vowels because of the composition of VocalSet, which is insufficient for use in automatic lyric transcription systems. We encourage researchers to create more resources that enable classifying a broader spectrum of vowels. Additionally, we plan to evaluate our models on an existing speech dataset, such as TIMIT [9], to test if our models are able to generalize to spoken vowels.

Another limitation of our work is the variation of our models' performance. While our models performed decently overall, for some singer-vowel combinations, performance would degrade. In addition, our models consistently underperformed by about 10% for female singers. One important area for future work is developing models that perform equally for both male and female singers.

Sung vowel classification is a challenging and understudied subset of the phoneme classification problem. Existing spectral methods that we examined for both spoken and sung vowels failed when applied to the robust VocalSet. In this work, we evaluate three neural models and establish baselines for sung vowel classification using the VocalSet dataset. We hope this work inspires future exploration on the challenging sung vowel classification problem.

References

1. Avramidis, K., Kratimenos, A., Garoufis, C., Zlatintsi, A., Maragos, P.: Deep convolutional and recurrent networks for polyphonic instrument classification from monophonic raw audio waveforms. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3010–3014 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413479
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460 (2020). https://doi.org/10.48550/arXiv.2006.11477
3. Burrows, D.: Singing and saying. J. Musicology **7**(3), 390–402 (1989). https://doi.org/10.2307/763607, http://www.jstor.org/stable/763607

4. Dabike, G.R., Barker, J.: Automatic lyric transcription from karaoke vocal tracks: resources and a baseline system. In: Proceedings of Interspeech, pp. 579–583. International Speech Communication Association (ISCA) (2019). https://doi.org/10.21437/Interspeech.2019-2378
5. De Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., Bourlard, H.: Evaluation of formant-like features on an automatic vowel classification task. J. Acoust. Soc. Am. **116**(3), 1781–1792 (2004). https://doi.org/10.1121/1.1781620
6. Demirel, E., Ahlbäck, S., Dixon, S.: Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2020). https://doi.org/10.1109/IJCNN48605.2020.9207052
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
8. Gales, M., Young, S.: Application of hidden Markov models in speech recognition. Found. Trends Sig. Process. **1**(3), 195–304 (2008). https://doi.org/10.1561/2000000004
9. Garofolo, J.S., et al.: Timit acoustic-phonetic continous speech corpus (1993). https://doi.org/10.35111/17gk-bn40
10. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) ICANN 2005. LNCS, vol. 3697, pp. 799–804. Springer, Heidelberg (2005). https://doi.org/10.1007/11550907_126
11. Gupta, C., Yılmaz, E., Li, H.: Automatic lyrics alignment and transcription in polyphonic music: does background music help? In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 496–500. IEEE (2020). https://doi.org/10.1109/TASLP.2022.3190742
12. He, Y., et al.: Streaming end-to-end speech recognition for mobile devices. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6381–6385 (2019). https://doi.org/10.1109/ICASSP.2019.8682336
13. Hillenbrand, J., Gayvert, R.T.: Vowel classification based on fundamental frequency and formant frequencies. J. Speech Lang. Hear. Res. **36**(4), 694–700 (1993). https://doi.org/10.1044/jshr.3604.694
14. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech, Lang. Process. **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
15. Jha, M.V., Rao, P.: Assessing vowel quality for singing evaluation. In: National Conference on Communications (NCC 2012), pp. 1–5. IEEE (2012). https://doi.org/10.1109/NCC.2012.6176860
16. Karpagavalli, S., Chandra, E.: A review on automatic speech recognition architecture and approaches. J. Sig. Process. Image Process. Pattern Recogn. **9**(4), 393–404 (2016). https://doi.org/10.14257/ijsip.2016.9.4.34
17. Kermanshahi, M.A., Akbari, A., Nasersharif, B.: Transfer learning for end-to-end ASR to deal with low-resource problem in Persian language. In: 26th International Computer Conference, Computer Society of Iran (CSICC), pp. 1–5 (2021). https://doi.org/10.1109/CSICC52343.2021.9420540

18. Khunarsal, P., Lursinsap, C., Raicharoen, T.: Singing voice recognition based on matching of spectrogram pattern. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 1595–1599 (2009). https://doi.org/10.1109/IJCNN.2009.5179014
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR) (2015). https://doi.org/10.48550/arXiv.1412.6980
20. Korkmaz, Y., Boyacı, A., Tuncer, T.: Turkish vowel classification based on acoustical and decompositional features optimized by genetic algorithm. Appl. Acoust. **154**, 28–35 (2019). https://doi.org/10.1016/j.apacoust.2019.04.027
21. Korkmaz, Y., Boyaci, A.: Classification of Turkish vowels based on formant frequencies. In: Proceedings of the International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1–4 (2018). https://doi.org/10.1109/IDAP.2018.8620877
22. Kruspe, A.M.: Training phoneme models for singing with "songified" speech data. In: Proceedings of the 16th International Conference on Music Information Retrieval (ISMIR), pp. 336–342 (2015). https://archives.ismir.net/ismir2015/paper/000034.pdf
23. Kruspe, A.M.: Application of automatic speech recognition technologies to singing. Ph.D. thesis, TU Technische Universität, Ilmenau, Germany (2018). https://www.db-thueringen.de/receive/dbt_mods_00035065
24. Li, Y., et al.: MERT: acoustic music understanding model with large-scale self-supervised training (2023). https://doi.org/10.48550/arXiv.2306.00107
25. Mesaros, A., Virtanen, T.: Automatic recognition of lyrics in singing. EURASIP J. Audio, Speech, Music Process. **2010**, 1–11 (2010). https://doi.org/10.1155/2010/546047
26. van den Oord, A., et al.: Wavenet: a generative model for raw audio (2016). https://doi.org/10.48550/arXiv.1609.03499
27. Ou, L., Gu, X., Wang, Y.: Transfer learning of wav2vec 2.0 for automatic lyric transcription. In: Proceedings of the 23rd International Conference on Music Information Retrieval (ISMIR), pp. 891–899 (2022). https://archives.ismir.net/ismir2022/paper/000107.pdf
28. Palaz, D., Collobert, R., Doss, M.M.: Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In: Proceedings of Interspeech (2013). https://doi.org/10.21437/Interspeech.2013-438
29. Santana, I.A.P., et al.: Music4all: a new music database and its applications. In: International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 399–404. IEEE (2020). https://doi.org/10.1109/IWSSIP48289.2020.9145170
30. Smule, I.: DAMP-MVP: Digital Archive of Mobile Performances - Smule Multilingual Vocal Performance 300x30x2 (2018). https://doi.org/10.5281/zenodo.2747436
31. Story, B.H.: Vowel acoustics for speaking and singing. Acta Acust. Acust. **90**(4), 629–640 (2004)
32. Sundberg, J.: Articulatory differences between spoken and sung vowels in singers. STL-QPSR, KTH **1**(1969), 33–46 (1969)
33. Sundberg, J.: Articulatory interpretation of the "singing formant". J. Acoust. Soc. Am. **55**(4), 838–844 (1974). https://doi.org/10.1121/1.1914609
34. Teytaut, Y., Roebel, A.: Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice. In: Proceedings of Interspeech. pp. 61–65. International Speech Communication Association (ISCA) (2021). https://doi.org/10.21437/interspeech.2021-1676

35. Tieleman, T., Hinton, G.: Lecture 6e rmsprop: divide the gradient by a running average of its recent magnitude. In: COURSERA: Neural Networks for Machine Learning, vol. 4, pp. 26–31 (2012). https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
36. Verma, P., Berger, J.: Audio transformers: transformer architectures for large scale audio understanding. adieu convolutions (2021). https://doi.org/10.48550/arXiv.2105.00335
37. Weiss, R., Brown, W., Jr., Moris, J.: Singer's formant in sopranos: fact or fiction? J. Voice **15**(4), 457–468 (2001). https://doi.org/10.1016/s0892-1997(01)00046-7
38. Wilkins, J., Seetharaman, P., Wahl, A., Pardo, B.: Vocalset: a singing voice dataset. In: Proceedings of the 19th International Conference on Music Information Retrieval (ISMIR), pp. 468–474 (2018). https://doi.org/10.5281/zenodo.1193957