VOCALSET: A SINGING VOICE DATASET

Julia Wilkins[1,2] **Prem Seetharaman**[1] **Alison Wahl**[2,3] **Bryan Pardo**[1]
[1] Computer Science, Northwestern University, Evanston, IL
[2] School of Music, Northwestern University, Evanston, IL
[3] School of Music, Ithaca College, Ithaca, NY
juliawilkins2018@u.northwestern.edu

ABSTRACT

We present VocalSet, a singing voice dataset of a capella singing. Existing singing voice datasets either do not capture a large range of vocal techniques, have very few singers, or are single-pitch and devoid of musical context. VocalSet captures not only a range of vowels, but also a diverse set of voices on many different vocal techniques, sung in contexts of scales, arpeggios, long tones, and excerpts. VocalSet has recordings of 10.1 hours of 20 professional singers (11 male, 9 female) performing 17 different different vocal techniques. This data will facilitate the development of new machine learning models for singer identification, vocal technique identification, singing generation and other related applications. To illustrate this, we establish baseline results on vocal technique classification and singer identification by training convolutional network classifiers on VocalSet to perform these tasks.

1. INTRODUCTION

VocalSet is a singing voice dataset containing 10.1 hours of recordings of professional singers demonstrating both standard and extended vocal techniques in a variety of musical contexts. Existing singing voice datasets aim to capture a focused subset of singing voice characteristics, and generally consist of fewer than five singers. VocalSet contains recordings from 20 different singers (11 male, 9 female) performing a variety of vocal techniques on 5 vowels. The breakdown of singer demographics is shown in Figure 1 and Figure 3, and the ontology of the dataset is shown in Figure 4. VocalSet improves the state of existing singing voice datasets and singing voice research by capturing not only a range of vowels, but also a diverse set of voices on many different vocal techniques, sung in contexts of scales, arpeggios, long tones, and excerpts.

Recent generative audio models based on machine learning [11, 25] have mostly focused on speech applications, using multi-speaker speech datasets [6, 13]. Generation of musical instruments has also recently been explored [2, 5]. VocalSet can be used in a similar way, but for singing voice generation. Our dataset can also be used to train systems for vocal technique transfer (e.g. transforming a sung tone without vibrato into one with vibrato) and singer style transfer (e.g. transforming a female singing voice to a male singing voice). Deep learning models for multi-speaker source separation have shown great success for speech [7, 21]. They work less well on singing voice. This is likely because they were never trained on a variety of singers and singing techniques. This dataset could be used to train machine learning models to separate mixtures of multiple singing voices. The dataset also contains recordings of the same musical material with different modulation patterns (vibrato, straight, trill, etc), making it useful for training models or testing algorithms that perform unison source separation using modulation pattern as a cue [17, 22]. Other obvious uses for such data are training models to identify singing technique, style [9, 19], or a unique singer's voice [1, 10, 12, 14].

The structure of this article is as follows: we first compare VocalSet to existing singing voice datasets and cover existing work in singing voice analysis and applications. We then describe the collection and recording process for VocalSet and detail the structure of the dataset. Finally, we illustrate the utility of VocalSet by implementing baseline classification systems for identifying vocal technique and

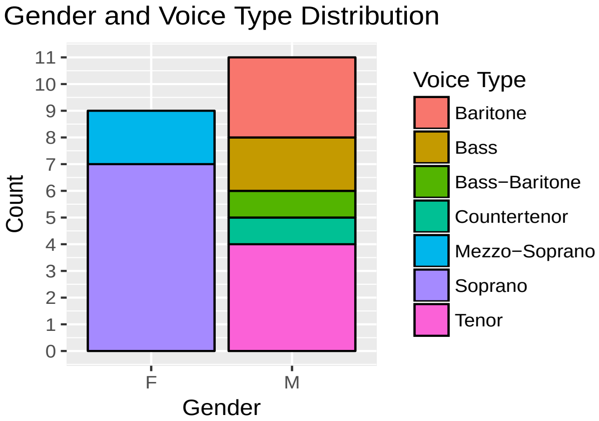

Figure 1. Distribution of singer gender and voice type. VocalSet data comes from 20 professional male and female singers ranging in voice type.

Attribution: Julia Wilkins, Prem Seetharaman, Alison Wahl, Bryan Pardo. "VocalSet: A Singing Voice Dataset", 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

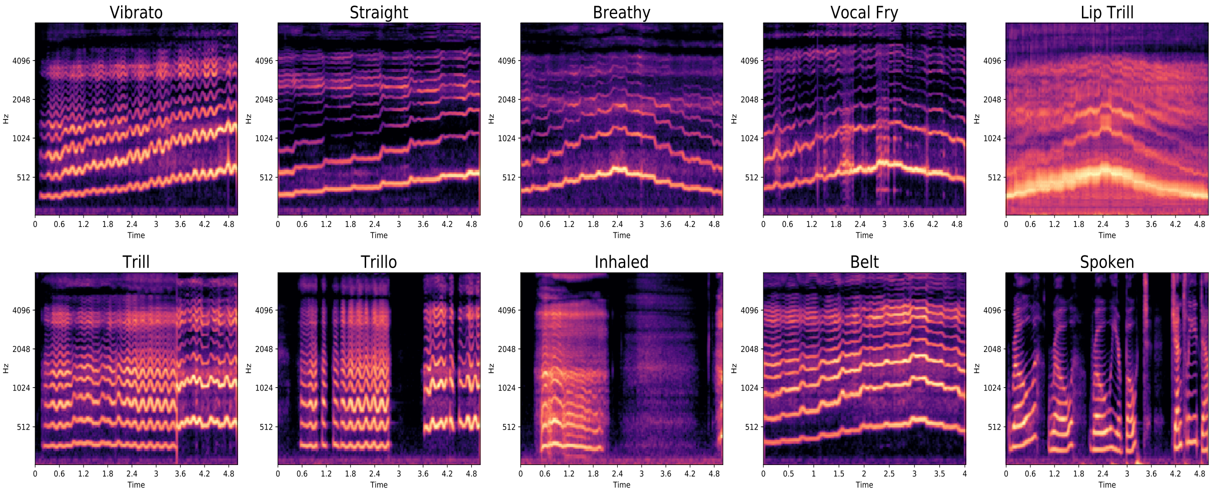

Figure 2. Mel spectrograms of 5-second samples of the 10 techniques used in our vocal technique classification model. All samples are from Female 2, singing scales, except “Trill”, “Trillo”, and “Inhaled” which are found only in the Long Tones section of the dataset, and “Spoken” which is only in the Excerpts section.

singer identification, trained on VocalSet.

2. RELATED WORK

A few singing voice datasets already exist. The Phonation Modes Dataset [18] captures a range of vocal sounds, but limits the included techniques to ’breathy’, ’pressed’, ’flow’, and ’neutral’. The dataset consists of a large number of sustained, sung vowels on a wide range of pitches from four singers. While this dataset does contain a substantial range of pitches, the pitches are isolated, lacking any musical context (e.g. a scale, or an arpeggio). This makes it difficult to model changes between pitches. VocalSet consists of recordings within musical contexts, allowing for this modeling. The techniques listed above that are observed in the Phonation Modes Dataset are based on the different formations of the throat when singing and not necessarily on musical applications of these techniques. Our dataset focuses on a broader range of techniques in singing, such as vibrato, trill, vocal fry, and inhaled singing. See Table 2 for the full set of techniques in our dataset.

The Vocobox dataset [1] focuses on single vowel and consonant vocal samples. While they feature a broad range of pitches, they only capture data from one singer. Our data contains 20 singers, with a wide range of voice types and singing styles over a larger range of pitches.

The Singing Voice Dataset [3] contains over 70 vocal recordings of 28 professional, semi-professional, and amateur singers performing predominantly Chinese Opera. This dataset does capture a large range of voices, like VocalSet. However, it does not focus on the distinction between vocal techniques but rather on providing extended excerpts of one genre of music. VocalSet provides a wide range of vocal techniques that one would not necessarily classify within a single genre so that models trained on VocalSet could generalize well to many different singing voice tasks.

We illustrate the utility of VocalSet by implementing baseline systems trained on VocalSet for identifying vocal technique and singer identification. Prior work on vocal technique identification includes work that explored the salient features of singing voice recordings in order to better understand what distinguishes one person’s singing voice from another as well as differences in sung vowels [4, 12], and work using source separation and F0 estimation to allow a user to edit the vocal technique used in a recorded sample [8].

Automated singer identification has been a topic of interest since at least 2001 [1,10,12,14]. Typical approaches use shallow classifiers and hand-crafted features (e.g. mel ceptral coefficients) [16,24]. Kako et al. [9] identifies four singing styles music style using the phase plane. Their work was not applied to specific vocal technique classification, likely due to the lack of a suitable dataset. We hypothesize that deep models have not been proposed in this area due to the scarcity of high-quality training data with multiple singers. The VocalSet data addresses these issues. We illustrate this point by training deep models for singer identification and vocal technique classification using this data.

For singing voice generation, [20] synthesized singing voice by replicating distinct and natural acoustic features of sung voice. In this work, we focus on classification tasks rather than generation tasks. However, VocalSet could be applied to generation tasks as well, and we hope our making this dataset available will facilitate that research.

[1] https://github.com/vocobox/human-voice-dataset

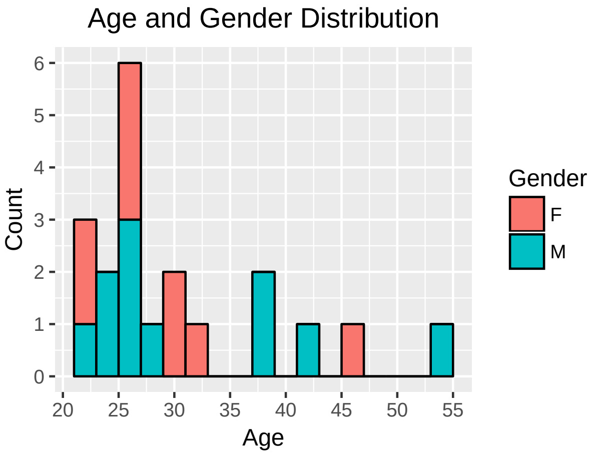

Figure 3. Distribution of singer age and gender. Singer age $\mu = 30.9, \sigma = 8.7$. We observe that the majority of singers lie in the range of 20 to 32, with a few older outlying singers.

3. VOCALSET

3.1 Singer Recruitment

9 female and 11 male professional singers were recruited to participate in the data collection. A professional singer was considered to be someone who has had vocal training leading to a bachelors or graduate degree in vocal performance and also earns a portion of their salary from vocal performance. The singers are of a wide age range and performance specializations. Voice types present in the dataset include soprano, mezzo, countertenor, tenor, baritone, and bass. See Figure 1 for a detailed breakdown of singer gender and voice type and Figure 3 for the distribution of singer age vs. gender. We chose to include a relatively even balance of genders and voice types in the dataset in order to capture a wide variety of timbre and spectral range.

3.2 Recording setup

Participants were recorded in a studio-quality recording booth with an Audio-Technica AT2020 condenser microphone, with a cardioid pickup pattern. Singers were placed close to the microphone in a standing position. Reference pitches were given to singers to ensure pitch accuracy. A metronome was played for the singers immediately prior to recording for techniques that required a specific tempo. Techniques marked 'fast' in Table 2 were targeted at 330 BPM, while techniques marked 'slow' were targeted at 60 BPM. Otherwise, the tempo is varied.

3.3 Dataset Organization

The dataset consists of 3,560 WAV files, totalling 10.1 hours of recorded, edited audio. The audio files vary in length, from less than 1 second (quick arpeggios) to 1 minute. Participants were asked to sing short vocalises of arpeggios, scales, long tones, and excerpts during the data collection. The arpeggios and scales were sung using 10 different techniques. The long tones were sung on 7 techniques, some of which also appear in arpeggios and scales (see Figure 4). Each singer was also asked to sing *Row, Row, Row Your Boat*, *Caro Mio Ben*, and *Dona Nobis Pacem* each in vibrato and straight tone, as well as an excerpt of their choice. The techniques included range from standard techniques such as 'fast, articulated forte' to difficult extended techniques such as 'inhaled singing'. For arpeggios, scales, and long tones, every vocalise was sung on vowels 'a', 'e', 'i', 'o', and 'u'. A portion of the arpeggios and scales are in both C major and F major (underlined in 4, while the harsher extended techniques and long tones are exclusively in C major. For example, singers were instructed to 'belt' a C major arpeggio on each vowel, totalling to 5 audio clips (one per vowel). This is shown in Figure 4. Table 2 shows the data broken down quantitatively by technique.

The data is sorted in nested folders specifying the singer, type of sample, and vocal technique used. This folder hierarchy is displayed in Figure 4.

Each sample is uniquely labelled based on this nested folder structure that it lies within. For example, Female 2 singing a slow, forte arpeggio in the key of F and on the vowel 'e' is labelled as 'f2_arpeggios_f_slow_forte_e.wav'.

The dataset is publicly available [2] and samples from the dataset used in training the classification models are also available on a demo website [3].

4. EXPERIMENTS

As an illustrative example of the utility of this data, we perform two classification tasks using a deep learning model on the VocalSet data. In the first task, we classify vocal techniques from raw time series audio using convolutional neural networks. In the second task, we identify singers from raw audio using a similar architecture. The network architectures are shown in Table 1. Note, architectures are identical except for the final output layer.

4.1 Training data and data preprocessing

We removed silence from the beginning, middle, and end of the recordings and then partitioned them into 3 second, non-overlapping chunks at a sample rate of 44.1k. The chunks were then normalized using their mean and standard deviation so that the network didn't use amplitude as a feature for classification. Additionally, by limiting the chunk to 3 seconds of audio, our models can't use musical context as a cue for learning the vocal technique. These vocal techniques can be deployed in a variety of contexts, so being context-invariant is important for generalization.

For each task, we partitioned the dataset into a training and a test set. For the vocal technique classification, we place all samples from 15 singers in the training set and all samples from the remaining 5 singers in the test set. For the singer identification, we needed to ensure that all

[2] https://doi.org/10.5281/zenodo.1203819

[3] https://interactiveaudiolab.github.io/demos/vocalset

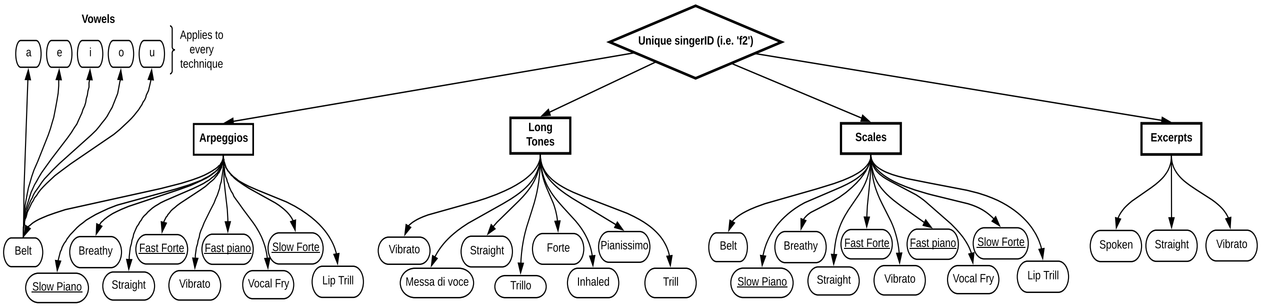

Figure 4. Breakdown of the techniques used in the VocalSet dataset. Each singer performs in four different contexts: arpeggios, long tones, scales, and excerpts. The techniques used in each context are shown. Each technique is sung on 5 vowels, and underlined techniques indicate that the technique was sung in F major *and* C major.

Layer Name	Input	Conv1	BatchNorm1	MaxPool1	Conv2	BatchNorm2	MaxPool2	Conv3	BatchNorm3	MaxPool3	Dense1	Dense2
# of Units/Filters	3*44100	16	16	-	8	8	-	32	32	-	32	10/20
Filter Size, Stride	-	(1, 128), (1, 1)	-	(1, 64), (1, 8)	(1, 64), (1, 1)	-	(1, 64), (1, 8)	(1, 256), (1, 1)	-	(1, 64), (1, 8)	-	-
Activation function	-	ReLU	-	-	ReLU	-	-	ReLU	-	-	ReLU	softmax

Table 1. Network architecture. The input to the network is 3 seconds of time series audio samples from VocalSet. The output is a 10-way classification for vocal technique classification and a 20-way classification for Singer ID. The architecture for both classifiers is identical except for the output size of the final dense layer. For the dense layers, L2 regularization was set to .001.

singers were present in both the training and the test sets in order to both train and test the model using the full range of singer ID possibilities. We randomly sampled the entire dataset to create training and test sets with a ratio of 0.8 (train): 0.2 (test), while ensuring all singers were both in training and testing data. The recordings were disjoint between the training and test sets, meaning that parts of the same recording were not put in both training and testing data.

Our vocal technique classifier model was trained and tested on the following ten vocal techniques: vibrato, straight tone, belt, breathy, lip trill, spoken, inhaled singing, trill, trillo, and vocal fry (bold in Table 2). Mel spectrograms of each technique are shown in 2, illustrating some of the differences between these vocal techniques.

The remaining categories, such as *fast/articulated forte* and *messa di voce* were not included in training for vocal technique classification. These techniques are heavily dependent on the amplitude of the recorded sample, and the inevitable human variation in the interpretation of dynamic instructions makes these samples highly variable in amplitude. Additionally, singers were not directed to sing a particular *technique* when making amplitude-oriented technique. As a result, singers often paired these amplitude-based techniques with other techniques at the same time, making the categories non-exclusive (e.g. singing fast/articulated forte with a lot of vibrato, or possibly with straight tone). Additionally, messa di voce was excluded because this technique requires singers to slowly crescendo and then decrescendo which, in full, was generally much longer than 3 seconds (the length of training samples).

We train our models with a convolution neural network using RMSProp [23], a learning rate of 1e-4, ReLU activation functions, an L2 regularization of 1e-3, and a dropout of 0.4 for the second to last dense layer. We use cross entropy as the loss function and a a batch size of 64. We train both the singer identification and vocal technique classification models for 200,000 iterations each, where the only difference between the two model architectures is the output size of the final dense layer (10 for vocal technique, 20 for singer ID). Both models were implemented in PyTorch. [15].

4.1.1 *Data augmentation*

We can also augment our data using standard data augmentation techniques for audio such as pitch shifting. We do this to our training set for vocal technique classification, but not for singer identification. Every excerpt is pitch shifted up and down 0.5 and 0.25 half steps. We report the effect of data augmentation on our models in Table 3. As shown in the table, we did observe some but not a significant accuracy boost when using the augmented model.

4.2 Vocal technique classification

4.2.1 *Results*

Evaluation metrics for our best 10-way vocal technique classification model are shown in Table 3. We were able to achieve these results using the model architecture in Table 1. This model performs well on unseen test data as we can see from table metrics. When examining sources of confusion for the model, we observed that the model most frequently incorrectly labels test samples as “straight” and “vibrato”. We attribute this in part to the class imbalance in the training data in which there are many more “vibrato” and “straight” samples than other techniques. Additionally, for techniques such as “belt”, many singers exhibited a great deal of vibrato when producing those samples which could place such techniques under the umbrella of

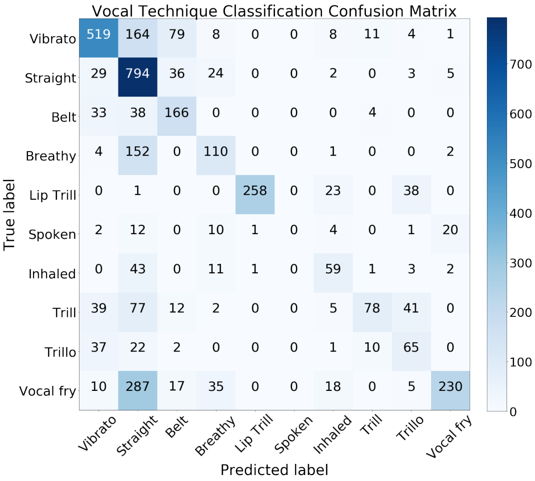

Figure 5. Confusion matrix for the technique classification model showing the quantity of predicted labels vs. true labels for each vocal technique. This model was trained on 10 vocal techniques. A class imbalance can be observed, as the number of vibrato and straight samples is much larger than the remaining techniques. The model performs relatively well for a majority of the techniques, however we see that nearly half of the vocal technique test samples were incorrectly classified as straight tone.

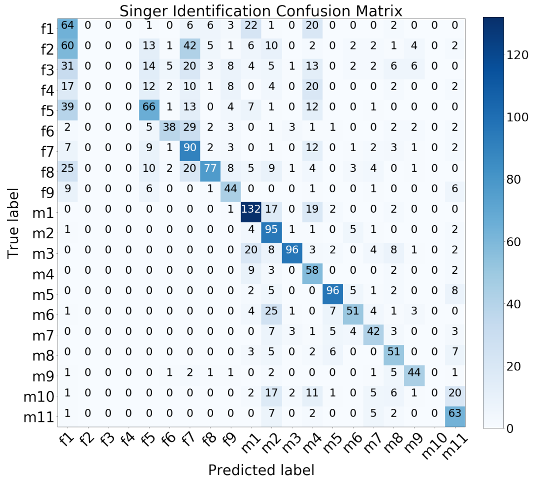

Figure 6. Confusion matrix for the singer identification model displaying the predicted singer identification vs. the true singer identification. We can observe that female voices are much more commonly classified incorrectly versus male voices, likely due to the broader range of male voices present in the training data.

Vocal Techniques	**Examples (#)**	**Time (min.)**
Fast/articulated forte	394	22.57
Fast/articulated piano	386	23.03
Slow/legato forte	395	65.28
Slow/legato piano	397	69.75
Lip trill	202	24.40
Vibrato	255	57.79
Breathy	200	28.00
Belt	205	26.24
Vocal fry	198	34.10
Full voice forte	100	16.29
Full voice pianissimo	100	16.58
Trill (upper semitone)	95	18.45
Trillo (goat tone)	100	14.54
Messa di voce	99	23.47
Straight tone	361	71.65
Inhaled singing	100	9.95
Spoken excerpt	20	4.06
Straight tone excerpt	60	24.19
Molto vibrato excerpt	59	24.55
Excerpt of choice	20	20.50

Table 2. The content of VocalSet, totalling to 10.1 hours of audio. Each vocal technique is performed by all 20 singers (11 male, 9 female). Some vocal techniques are performed in more musical contexts (e.g. scales) than others. Bold techniques were used for our classification task.

"vibrato". We also observed a little bit of expected confusion between "trill" and "vibrato", as these techniques may have some overlap depending on the singer performing the technique. As seen in Figure 2, the spectrogram representation of these two techniques looks very similar. To address the issue of class imbalance, we tried using data augmentation with pitch shifting to both balance the classes and create more data, but as previously stated and shown in Table 3, there was little improvement over the original model when using training data augmentation.

4.3 Singer identification (ID)

4.3.1 Results

Evaluation metrics for our best 20-way singer identification model are shown in Table 3. The model architecture is identical to that of the vocal technique classification model (see 1), with the exception of the number of output nodes in the final dense layer (20 in the singer identification model vs. 10 in the technique model). The singer identification model did not perform as well as the vocal technique classification model. As shown in Table 3, classifying male voices correctly was much easier for the model than classifying female voices. This is expected due to the high similarity between the female voices in the training data. Figure 1 shows that the female data only contains 2 voice types, while the male data contains 5 voice types.

Because voice type is largely dependent on the vocal range of the singer, having 5 different voice types within the male singers makes it much easier to distinguish be-

Classification Task	Prior	Precision	Recall	Top-2 Accuracy	Top-3 Accuracy	Male Accuracy	Female Accuracy
Vocal Technique	0.242	0.676	0.619	0.801	0.867	-	-
Vocal Technique (trained on augmented data)	0.242	0.677	0.628	0.815	0.891	-	-
Singer ID	-	0.473	0.516	0.638	0.700	0.684	0.351

Table 3. Evaluation metrics for our vocal technique and Singer ID classification models performing on unseen test data. "Prior" indicates the accuracy if we were to simply choose the most popular class ("straight") to predict test data. We observe a very slight increase in accuracy in the augmented vocal technique model. Our singer ID model has lower performance, likely due to the similarity between different, primarily female, singers.

tween male singers than female singers. The accuracy (recall) for classifying unseen male singers was nearly twice as good as that of unseen female singers.

5. FUTURE WORK

In the future, we plan to experiment with more network architectures and training techniques (e.g. Siamese training) to improve the performance of our classifiers. We also expect researchers to use the VocalSet dataset to train a vocal style transformation model that can transform a voice recording into one using one of the techniques that we have recorded in VocalSet. For example, an untrained singer could sing a simple melody on a straight tone, and our system could remodel their voice using the vibrato or articulation of a professional singer. We envision this as a tool for both musicians and non-musicians alike, and hope to create a web application or even a physical sound installation that users could transform their voices in. We would also like to use VocalSet to train autoregressive models (e.g. Wavenet [25]) that can generate singing voice of specific techniques.

6. CONCLUSION

VocalSet is a large dataset of high-quality audio recordings of 20 professional singers demonstrating a variety of vocal techniques on different vowels. Existing singing voice datasets either do not capture a large range of vocal techniques, have very few singers, or are single-pitch and lacking musical context. VocalSet was collected to fill this gap. We have shown illustrative examples of how VocalSet can be used to develop systems for diverse tasks. The VocalSet data will facilitate the development of a number of applications, including vocal technique identification, vocal style transformation, pitch detection, and vowel identification. VocalSet is available for download at https://doi.org/10.5281/zenodo.1203819.

7. ACKNOWLEDGMENTS

This work was supported by NSF Award #1420971 and by a Northwestern University Center for Interdisciplinary Research in the Arts grant.

8. REFERENCES

[1] Mark A Bartsch and Gregory H Wakefield. Singing voice identification using spectral envelope estimation. *IEEE Transactions on speech and audio processing*, 12(2):100–109, 2004.

[2] Merlijn Blaauw and Jordi Bonada. A neural parametric singing synthesizer modeling timbre and expression from natural songs. *Applied Sciences*, 7(12):1313, 2017.

[3] Dawn A. Black, Ma Li, and Mi Tian. Automatic identification of emotional cues in chinese opera singing. 2014.

[4] Thomas F. Cleveland. Acoustic properties of voice timbre types and their influence on voice classification. *The Journal of the Acoustical Society of America*, 61(6):1622–1629, 1977.

[5] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders. *arXiv preprint arXiv:1704.01279*, 2017.

[6] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[7] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 31–35. IEEE, 2016.

[8] Yukara Ikemiya, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Singing voice separation and vocal f0 estimation based on mutual combination of robust principal component analysis and subharmonic summation. 24(11), Nov. 2016.

[9] Tatsuya Kako, Yasunori Ohishi, Hirokazu Kameoka, Kunio Kashino, and Kazuya Takeda. Automatic identification for singing style based on sung melodic contour characterized in phase plane. In *ISMIR*, pages 393–398. Citeseer, 2009.

[10] Youngmoo E Kim and Brian Whitman. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, volume 13, page 17, 2002.

[11] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[12] Maureen et al. Mellody. Modal distribution analysis, synthesis, and perception of a soprano's sung vowels. pages 469–482, 2001.

[13] Gautham J Mysore. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? a dataset, insights, and challenges. *IEEE Signal Processing Letters*, 22(8):1006–1010, 2015.

[14] Tin Lay Nwe and Haizhou Li. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):519–530, 2007.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[16] Hemant A Patil, Purushotam G Radadia, and TK Basu. Combining evidences from mel cepstral features and cepstral mean subtracted features for singer identification. In *Asian Language Processing (IALP), 2012 International Conference on*, pages 145–148. IEEE, 2012.

[17] Fatemeh Pishdadian, Bryan Pardo, and Antoine Liutkus. A multi-resolution approach to common fate-based audio separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 566–570. IEEE, 2017.

[18] Polina Proutskova, Christopher Rhodes, and Tim Crawford. Breathy, resonant, pressed - automatic detection of phonation mode from audio recordings of singing. 2013.

[19] Keijiro Saino, Makoto Tachibana, and Hideki Kenmochi. A singing style modeling system for singing voice synthesizers. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[20] T. Saitou, M. Goto, M. Unoki, and M. Akagi. Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices. pages 215–218, Oct 2007.

[21] Paris Smaragdis, Gautham Mysore, and Nasser Mohammadiha. Dynamic non-negative models for audio source separation. In *Audio Source Separation*, pages 49–71. Springer, 2018.

[22] Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau, Bernd Edler, and Paul Magron. Common fate model for unison source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 126–130. IEEE, 2016.

[23] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[24] Tsung-Han Tsai, Yu-Siang Huang, Pei-Yun Liu, and De-Ming Chen. Content-based singer classification on compressed domain audio data. *Multimedia Tools and Applications*, 74(4):1489–1509, 2015.

[25] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.