A Model-Driven Exploration of Accent Within the Amateur Singing Voice

Camille Noufi [1] **Vidya Rangasayee** [1] **Sarah Ciresi** [2] **Jonathan Berger** [1] **Blair Kaneshiro** [1]

Abstract

We seek to detect characteristics of regional language accent in solo singing using two variants of convolutional neural networks to classify reported country and language from ten countries during karaoke-style vocal performance of the broadly popular hymn, *Amazing Grace*. The most successful model produces overall accuracy of 15.64%, with false classification of singing segments to be variants of English at 53.4%. The model also separates learned classes along a rhythmic-stress dimension with English variants at its origin. These observations suggest that, based on the network's success in learning intonation features, a singer's speech pronunciation adapts to the language of the song being sung.

1. Introduction

The interplay of prosody and intonation are significant attributes in both recitation and singing. In song, which characteristics of our speaking voice are modified to fit the style or phrasing? A novel approach to explore these characteristics is to study a song performed by someone whose native tongue might not be the language of the song. We explore this facet by attempting to classify the accent of amateur singers (via proxy 'country-language') using convolutional neural networks. We use audio recordings of 1,015 performers from ten countries solo-singing the well-known English standard *Amazing Grace*. In this paper, we discuss the per-class accuracy of the most successful model to understand if accent features during singing can be detected and learned, and if so, how and why they may be confounded or modified.

[1]Center for Computer Research in Music and Acoustics, Stanford University, Stanford, California, USA [2]Department of Computer Science, Stanford University, Stanford, California, USA. Correspondence to: cnoufi <ccrma.stanford.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2. Solo-Singing Dataset

We use the Digital Archive of Musical Performances (DAMP) database[1], a subset of which contains audio recordings of singers from around the world performing *Amazing Grace* via Smule, Inc.'s 'Sing! Karaoke' mobile app. The karaoke accompaniment (100 BPM) is identical for all performances. Recordings are captured via smartphone microphones and are not preprocessed or compressed beyond what users' headphones and smartphones apply. The dataset includes only full-length, monophonic mono-channel recordings at a sampling rate of 22,050Hz.

We use the provided 'country-language' label associated with each recording and select the ten most-represented labels for classification to ensure ample training data per class (Table 1). We discard all audio files for which RMS energy of the noise floor exceeds 4%, bringing the dataset to 10,937 recordings. We undersample majority classes to balance the test partition while keeping validation and test sets representative of the original distribution. This yields 1,015 recordings for training, 975 for validation, and 985 on which to perform our classification task.

All recordings undergo energy-based voice activity detection[2] using a 0.95 confidence threshold. Normalized log-mel power spectrograms are computed on the voiced audio (frame size = 2048, hop size = 512 samples, 80 bins). Model inputs are 2-measure segments with 50% overlap.

Table 1. Ten Most-Represented Accent Classes

Label	Language-Country
de-DE	German-Germany
en-AU	English-Australia
en-CA	English-Canada
en-GB	English-Great Britain
en-US	English-United States
fr-FR	French-France
id-ID	Indonesian-Indonesia
nb-NO	Norwegian-Norway
pt-BR	Portuguese-Brazil
sv-SE	Swedish-Sweden

[1]https://ccrma.stanford.edu/damp/

[2]http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Table 2. Classification Performance

Architecture	Accuracy	F1	Precision	Recall
CNN-3x3	11.92%	17.06%	53.20%	11.93%
CNN-ResNeXt	15.29%	21.94%	53.94%	15.29%

3. Model Architecture and Experiment

We design a 3-layer CNN utilizing 32 3x3 filters per layer and a 3-layer CNN with a ResNeXt block after the first layer (Xie et al., 2017), used similarly by Wang & Tzanetakis (2018) for singing style-identification. A thorough description of of model design, training, and selection of hyperparameters are provided by Noufi et al. (2019). Table 2 describes each model's hyperparameters and accuracy metrics. Normalized confusion matrices are the source of the present analysis.

4. Results and Discussion

Similar to results obtained by Wang & Tzanetakis (2018), both models produce an overall accuracy slightly above chance. Figure 1 shows the ResNeXt network's difficulty in distinguishing between similar labels such as variants of English. Most notably, misclassifications of singing segments as being produced by an English speaker account for 53.8% percent of classifications: 18.2% of the 2-measure segments are classified as American English, followed by 13.8%, 12.2% and 9.6% as Australian, Canadian, and British, respectively. The confusions are largely asymmetric, meaning that the attributes learned for each 'accent class' are not equally mistaken for each other.

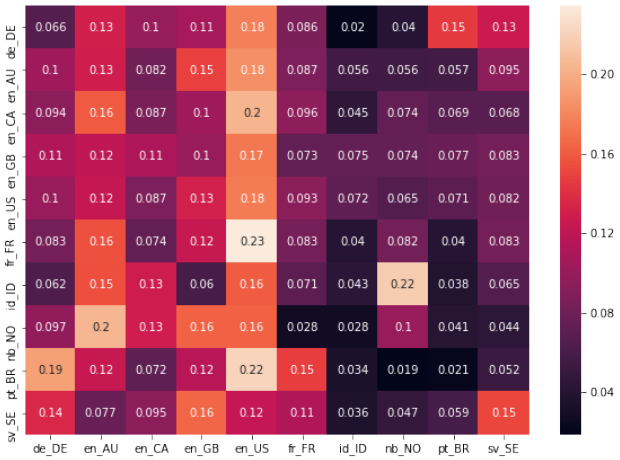

Figure 1. Normalized confusion matrix of ResNeXt classification.

Euclidean distance relations within the first two principal components of the confusion matrix are visualized in Figure 2. The first dimension spreads the classes more than the second, and a speech-prosody or intonation characteristic is not clear. Surprisingly, Norwegian and Swedish are the most dissimilar, an observation that argues this dimension is not learned via a language-prosody trait. However, English variants cluster together, as do variations of the Romance languages (French and Brazilian Portuguese).

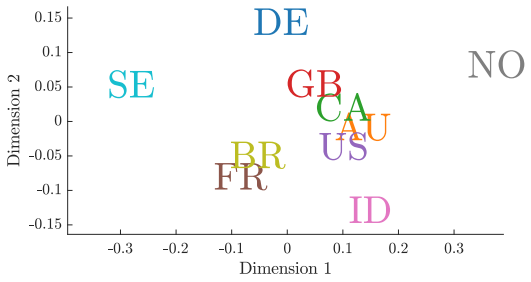

Figure 2. ResNeXt-learned class similarity (Euclidean distance)

Conversely, clusters along the second principal dimension align with prosodic tendencies of the languages themselves. The Romance languages again cluster together. The English variants separate the Romance languages from the Germanic and North-Germanic (also known as Scandinavian) languages along the second principal dimension. This second dimension may be separating the classes by rhythmic stress, from lesser use (Portuguese, French, Indonesian) to greater use (Swedish, Norwegian, German) (Arvaniti, 2009; Peppé et al., 2012). The situating of English around 0 suggests that the rhythmic phrasing required by 'Amazing Grace' is influenced by the English language. These observations align with our hypothesis that innate pronunciation is modified to fit the target language of the song.

As future work, we first dissect our model's output layer filters to determine whether it is indeed learning salient time-frequency representations. Secondly, we design new models to more accurately classify country-language labels. Harmonic-spectral-source-separation (HPSS) and spectral phoneme-classification may provide new information on timbre and rhythmic stress within a singing accent (Pons et al., 2017). Panteli et al. (2017) suggest that language and cultural similarities may influence vocal pitch contours of different singing styles. Thus, a recurrent model may learn the salient pitch and loudness contours that are either innate to a specific culture or to the song itself. Finally, an ongoing effort to confirm additional metadata labels such as gender and age of each singer will help highlight other innate vocal characteristics that likely influence intonation while singing.

Acknowledgements

We thank Elena Georgieva for her help in launching this project and Dr. Perry Cook for his vital role in helping us access and utilize the DAMP dataset.

References

Arvaniti, A. Rhythm, timing and the timing of rhythm. *Phonetica*, 66:46–63, 2009.

Choi, K., Fazekas, G., Cho, K., and Sandler, M. B. A tutorial on deep learning for music information retrieval. *arXiv Preprint*, 1709.04396, 2017. URL http://arxiv.org/abs/1709.04396.

Jiao, Y., Tu, M., Berisha, V., and Liss, J. Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. In *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, September 2016.

Noufi, C., Ciresi, S., and Rangasayee, V. Accent detection within the amateur singing voice. *Unpublished Manuscript, Stanford University*, 2019. URL http://web.stanford.edu/class/cs224n/reports/custom/15792047.pdf.

Panteli, M., Bittner, R., Bello, J. P., and Dixon, S. Towards the characterization of singing styles in world music. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 636–640, March 2017.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011.

Peppé, S., Coene, M., Hesling, I., Martinez Castilla, P., and Moen, I. *Translation to practice: Prosody in five European languages*, pp. 53–57. Communication Disorders Across Languages. Multilingual Matters, 2012.

Pons, J. and Serra, X. Designing efficient architectures for modeling temporal features with convolutional neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2472–2476, March 2017.

Pons, J., Slizovskaia, O., Gong, R., Gmez, E., and Serra, X. Timbre analysis of music audio signals with convolutional neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2744–2748, August 2017.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Wang, C. and Tzanetakis, G. Singing style investigation by residual siamese convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 116–120, April 2018.

Wessel, D. L. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.

Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.