Accent Detection Within the Amateur Singing Voice

Camille Noufi, Sarah Ciresi, Vidya Rangasayee

Overview

We investigate the feasibility of detecting characteristics of accent present during solo singing by using country and language as a proxy. We design variants of convolutional neural networks to classify the associated country and language of both native and non-native English speakers during their karaoke-style singing performance of English standard "Amazing Grace." The most successful architecture provides an 7.83% improvement in overall accuracy compared our baseline statistical model, demonstrating the networks' ability to detect subtle accent-based speech characteristics. We see an overwhelming prediction of variants of English accent, suggesting style-influenced modification of pronunciation and intonation when singing a well-known English song.

Data

We use a subset of the the Stanford DAMP Database, a proprietary database owned by Smule, developer of the Sing! Karaoke smartphone app. A subset of this database contains 10,937 audio recordings of both trained and untrained singers from around the world solo singing "Amazing Grace" on the app. Table 1 lists the 10 most-represented countries for our classification task. We undersample majority classes in our training set for balance, while keeping our validation and test sets representative, bringing our data to 2,975 recordings.

Table 1: Ten Most-Represented Accent Classes within DAMP-Amazing Grace

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

Audio Features

Both pronunciation and intonation are related to the property of sound known as timbre. Timbre is often described as the "color," "quality," or "tone" of a sound and is a property reliant on both time and frequency [6]. Figure 2 visualizes the log-compressed, mel-scale magnitude spectrograms of each audio recordings we calculate to capture perceptually relevant long- and short-term features of speech that describe pronunciation and intonation [1-7]. The conversion from frequency to mels is defined as

$$m = 1127 \ln(1 + f/700)$$

where m is the mel coefficient corresponding to the original signal frequency f in Hz. Figure 1 visualizes our features.

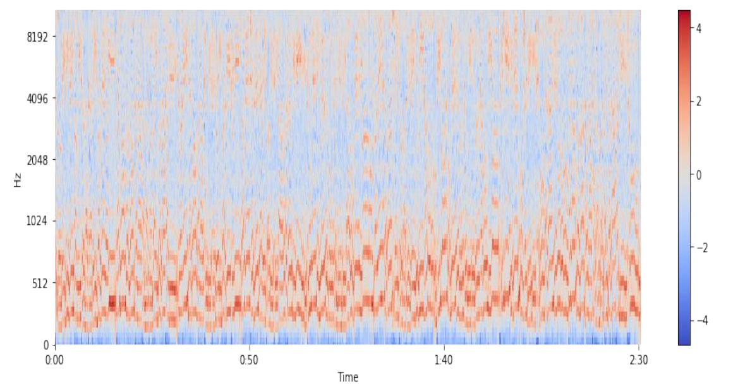

Figure 1: Mel Spectrogram of a singer's recorded performance

Methods

We perform energy-based voice activity detection on all audio files to remove silences (see Figure 2), and compute the mel-spectrograms on the remaining voiced audio using a frame size of 2048 audio samples and an overlap of 512 audio samples. To reduce dimensionality, two-measure (4.75 second) chunks of the spectrograms are fed into each network as input samples [1-7].

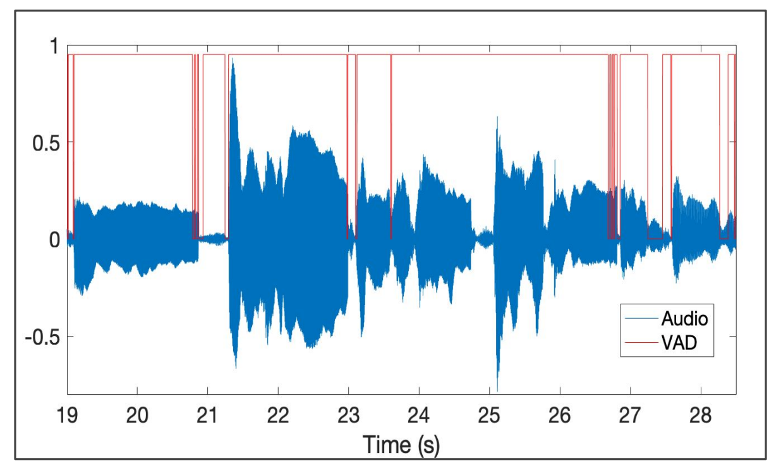

Figure 2: Voice Activity Detection of singer performance audio

Table 2: Training Parameters

Architecture	# Params	p	lr
CNN-3x3	11371792	0.5	0.001
CNN-ResNeXt	11302160	0.3	0.001
KNN	10		

(p = Dropout Rate, lr = Learning Rate

Figure 3 visualizes our two convolutional neural networks: a deep vanilla network with 32 3x3 filters per layer, and a deep network using a ResNeXt block after the first layer [8]. ReLU non-linear activations are used after all layers. Training occurs in batches of 128 samples for a maximum 40 epochs, and uses validation loss minimization as early stopping criterion with patience of 7.

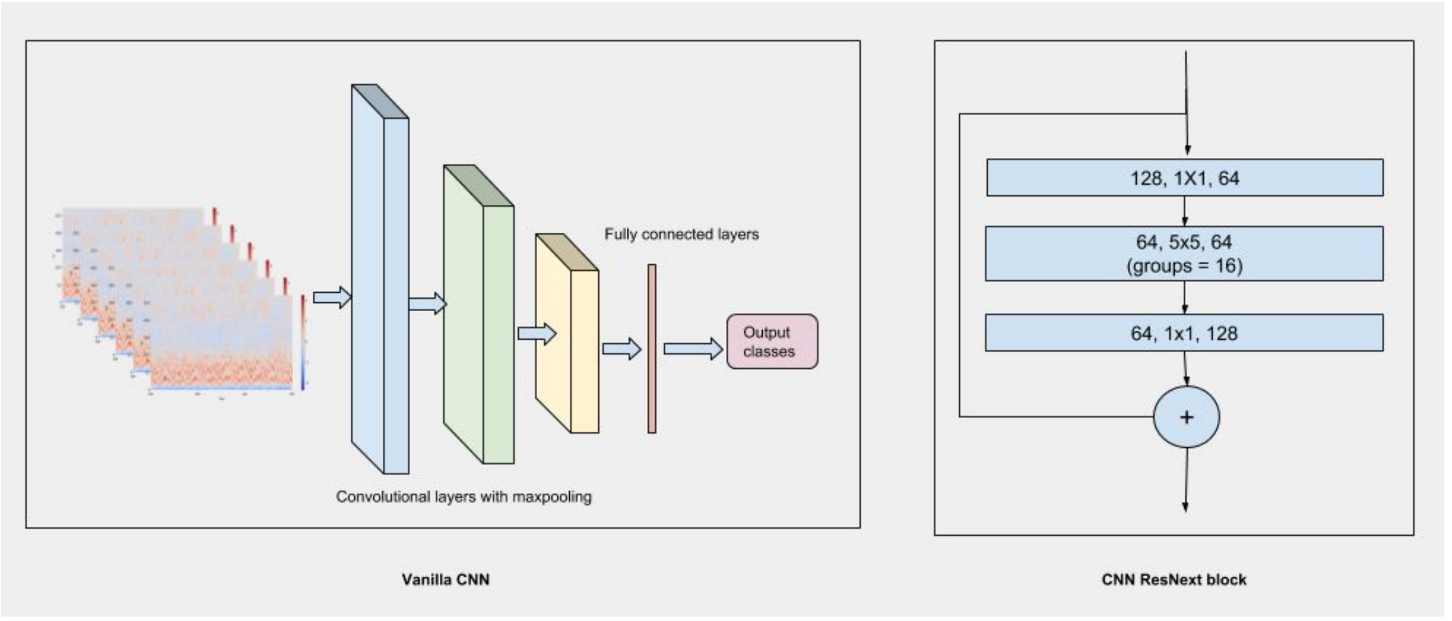

Figure 3: CNN-3x3 (left) and CNN-ResNeXt (right) Architectures

Classification Results

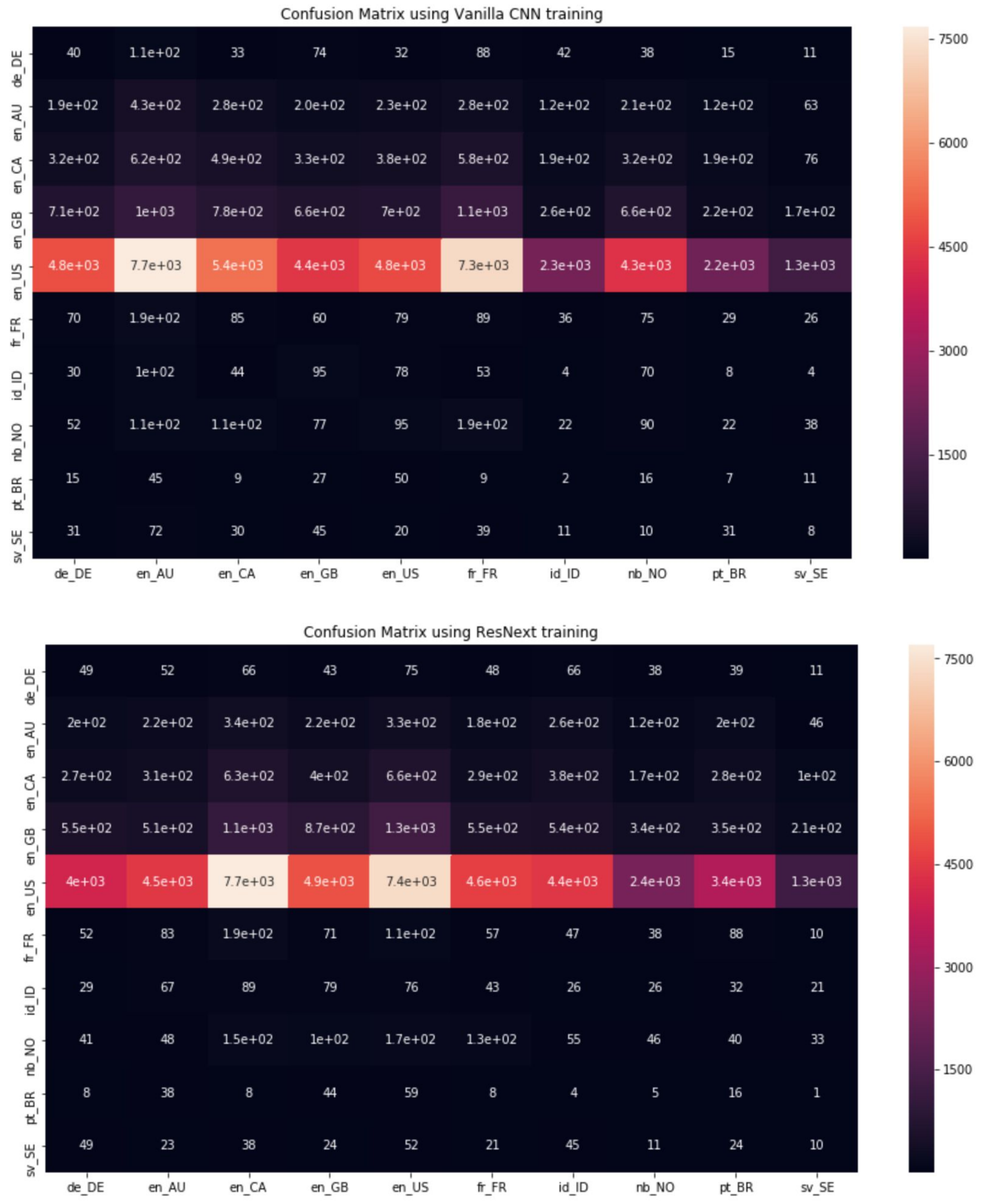

Figure 4: Confusion Matrices of CNN-3x3 (left) and CNN-ResNeXt (right) Classification Results

Table 3: Model Classification Performance

Architecture	Accuracy	F1
CNN-3x3	11.03%	15.93%
CNN-ResNeXt	15.64%	22.41%
KNN	7.81%	11.84%

Architecture	Precision	Recall
CNN-3x3	57.63%	11.03%
CNN-ResNeXt	56.31%	15.64%
KNN	60.89%	7.81%

Table 3 summarizes the metric-based classification performance of each neural network on our test set. Overall accuracy scores are consistent with [1]. The overall accuracy is a few percentage points above random (10%), indicating the system is learning, but not in enough detail to distinguish accurately between the ten classes. The confusion matrices shown in Figure 4 detail the systematic variation in classification between the two models. The breakdown of confusion per class highlights the common misclassification trends between the two models.

Discussion

The network has difficulty distinguishing between very similar labels such as Canadian English and American English. The ResNeXt model is able to classify American English and Canadian English the most accurately. Further, ***misclassifications as English across all classes are primarily misclassified as Canadian English or American English***. ***Correct- and cross-classifications of Canadian and American English occur 34.1% of the time, over twice the overall model accuracy***. The Scandinavian and French singers are most likely to be misclassified as English singers as well. The ***over-classification of English dialect even using a balanced training set may suggest that an English-style pronunciation is used to fit the target language of the song***.

Both models misclassify Norwegian singers primarily as English (spread between country variants), followed by French. Surprisingly, the models seem to have more difficulty associating Swedish singing with any particular class, and Swedish and Norwegian recordings are not commonly mixed up like the English-tagged recordings are. The models perform similarly for German and Indonesian, the only non-European/non-American accent in the dataset. The ***overall tendency toward randomness accuracy of both models, alongside the baseline, are indicators of two systematic issues.*** The first being that the ***country-language labels might not be a reliable proxy for accent.*** The labels are user-provided and do not guarantee that singers will invoke their day-to-day speaking voice into their singing performances. The second is that, while the mel-spectrograms are indeed pitch-invariant, other aspects of musicality, including ***skill-level and transient stylistic choices*** that may be completely independent of country, accent or language identity, ***might outweigh speech-like pronunciation.*** In general, ***performing a well-known song that has been stylized across many genres likely confounds accent-based pronunciation and intonation while singing.***

Future Work

Future work to identify stylistic pronunciation choices would help to further identify accent-specific pronunciation in singing. We plan to utilize all supplied metadata (country, language, age, gender) and augment it with perceived expression, likeability and skill level. Multi-tag classification in conjunction with examination of the feature maps of the convolutional neural networks will allow study of how these different tags influence or relate particular manifestations of speech characteristics during singing.

We also plan to apply phoneme-level segmentation to the audio and attempt to discriminate between pronunciations of a single phone. Finally, we began implementing a stacked CNN-RNN in order to capture temporally-influenced timbre, but were not able to thus far train efficiently enough to acquire results. Training on this architecture and comparing to our two existing CNN variants would provide insight into the influence articulatory sequencing has on accent during singing.

References

[1] Wang, C & Tzanetakis, G (2018) Singing Style Investigation by Residual Siamese Convolutional Neural Networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[2] Pons, J & Serra, X (2017) Designing Efficient Architectures For Modeling Temporal Features With Convolutional Neural Networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[3] Pons, J, Slizovskaia, O, Gong, R, Gomez, E & Serra, X. (2017) Timbre Analysis of Music Audio Signals with Convolutional Neural Networks. *arXiv Preprint, arXiv:1703.06697v2, 2017*

[4] Jiao, Y, Tu, M, Berisha, V, & Liss, J (2016) 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)*

[5] K. Choi, G. Fazekas, K. Cho, & M. Sandler (2017) A Tutorial on Deep Learning for Music Information Retrieval. *arXiv Preprint, arXiv:1709.04396v2, 2017*

[6] Wessel, D (1979), Timbre space as a musical control structure. *Computer Music Journal*, pp. 45–52.

[7] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, & S. McAdams, (2011) The timbre toolbox: Extracting audio descriptors from musical signals. In *The Journal of the Acoustical Society of America* vol. 130, no. 5, pp. 2902–2916.

[8] S. Xie, R. B. Girshick, P. D. Zhuowen & K. He, (2016) Aggregated Residual Transformations for Deep Neural Networks. *arXiv Preprint, arXiv:1611.05431*

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov, (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. In *Journal of Machine Learning Research* vol. 15, pp. 1929-1958.