INTONATION: A DATASET OF QUALITY VOCAL PERFORMANCES REFINED BY SPECTRAL CLUSTERING ON PITCH CONGRUENCE

Sanna Wager[1], *George Tzanetakis*[2,3], *Stefan Sullivan*[3], *Cheng-i Wang*[3]
John Shimmin[3], *Minje Kim*[1], *Perry Cook*[3,4]

[1] Indiana University, School of Informatics, Computing, and Engineering, Bloomington, IN, USA
[2]University of Victoria, Department of Computer Science, Victoria, BC, Canada
[3] Smule, Inc, San Francisco, CA, USA
[4]Princeton University, Departments of Computer Science and Music, Princeton, NJ, USA

ABSTRACT

We introduce the "Intonation" dataset of amateur vocal performances with a tendency for good intonation, collected from Smule, Inc. The dataset can be used for music information retrieval tasks such as autotuning, query by humming, and singing style analysis. It is available upon request on the Stanford CCRMA DAMP website.[1] We describe a semi-supervised approach to selecting the audio recordings from a larger collection of performances based on intonation patterns. The approach can be applied in other situations where a researcher needs to extract a subset of data samples from a large database. A comparison of the "Intonation" dataset and the remaining collection of performances shows that the two have different intonation behavior distributions.

Index Terms— music information retrieval, pitch, clustering, singing, dataset

1. INTRODUCTION

Useful datasets have been made available for certain research topics in the fields of music information retrieval and audio. These include sound event detection [1], source separation [2], and recommendations [3]. Sometimes, though, the best dataset available for a topic is huge and difficult to process. A large collection of audio recordings is available, but the recordings with suitable characteristics for the given analysis form a smaller subset of the dataset. The filtering process to extract the desired samples can be labor intensive, requiring that the researcher select the samples with the desired features, which may or may not be labeled and can be hard to model. One way to approach this selection process is to automate it using feature engineering and clustering.

In this paper, we present this kind of semi-automatic process for the task of searching through a large database of amateur karaoke performances for samples with a tendency for good musical intonation. The need for this task arose when we wished to train a machine-learning model to predict pitch correction. We needed to select performances that were in tune enough but not those that were out of tune or contained little singing. We note that this task requires quantifying the concept of singing "in tune". As we describe in further sections, the task is not obvious, so we avoid creating an explicit definition of "in tune" by using a semi-supervised approach. We first extract musical intonation features from each performance, then apply spectral clustering to them and subjectively choose clusters that sound "in tune" by listening to samples from each. We also introduce the resulting dataset and an analysis of the intonation tendencies of its performances. Though we present this approach for our specific task, it can be adapted to other tasks, datasets, and features.

The research work done for this paper was supported by the internship program at Smule, Inc., in collaboration with the audio/video team.

[1]The dataset and detailed description of the contents are available upon request via `https://ccrma.stanford.edu/damp`.

2. RELATED WORK

2.1. Pitch deviation analysis

Automatic analysis of musical intonation behavior has also been performed in other contexts. For example, the authors of [4] described an approach to discovering talented singers on YouTube based on features extracted mostly from the audio. One of the main features they chose consisted of a pitch deviation histogram, which characterizes intonation behavior of a full performance in a low dimension. Given that the performances were typically not associated with a musical score and that the singing was mixed with the accompaniment and other background sounds, the authors built the histogram from the Short-Time Fourier Transform amplitude peaks. A singer who sings flat should have a histogram skewed to the left, and an active vibrato will cause values to spread. Our feature extraction task is different from [4] because, as we describe below, we have access to the musical scores of the vocals and because the audio sources are separated. We can, therefore, apply a standard pitch detection algorithm to each vocal track and compare the results to the musical score. Comparison of performance pitch and musical score is also used by [5] in the context of a tool for musical performance visualization.

2.2. Intonation studies

Pitch in a karaoke context and, more generally, in many scenarios where a musical score is used, is modeled as the twelve discrete frequencies per octave, evenly spaced in the logarithmic scale, that constitute the equal-tempered scale. Quantitative and qualitative studies on musical intonation of professional-level singers, however, indicate frequent, deliberate deviations from the equal-tempered scale. In particular, musicians often sing or play sharp relative to an accompaniment. [6] describes this phenomenon, citing [7, 8, 9]. Research such as described in [10] finds much variety in musical interval sizes both above and below the equal-tempered intervals in polyphonic choral music. The results are interesting in our context because they indicate that the pitch of good singers might not simply deviate from

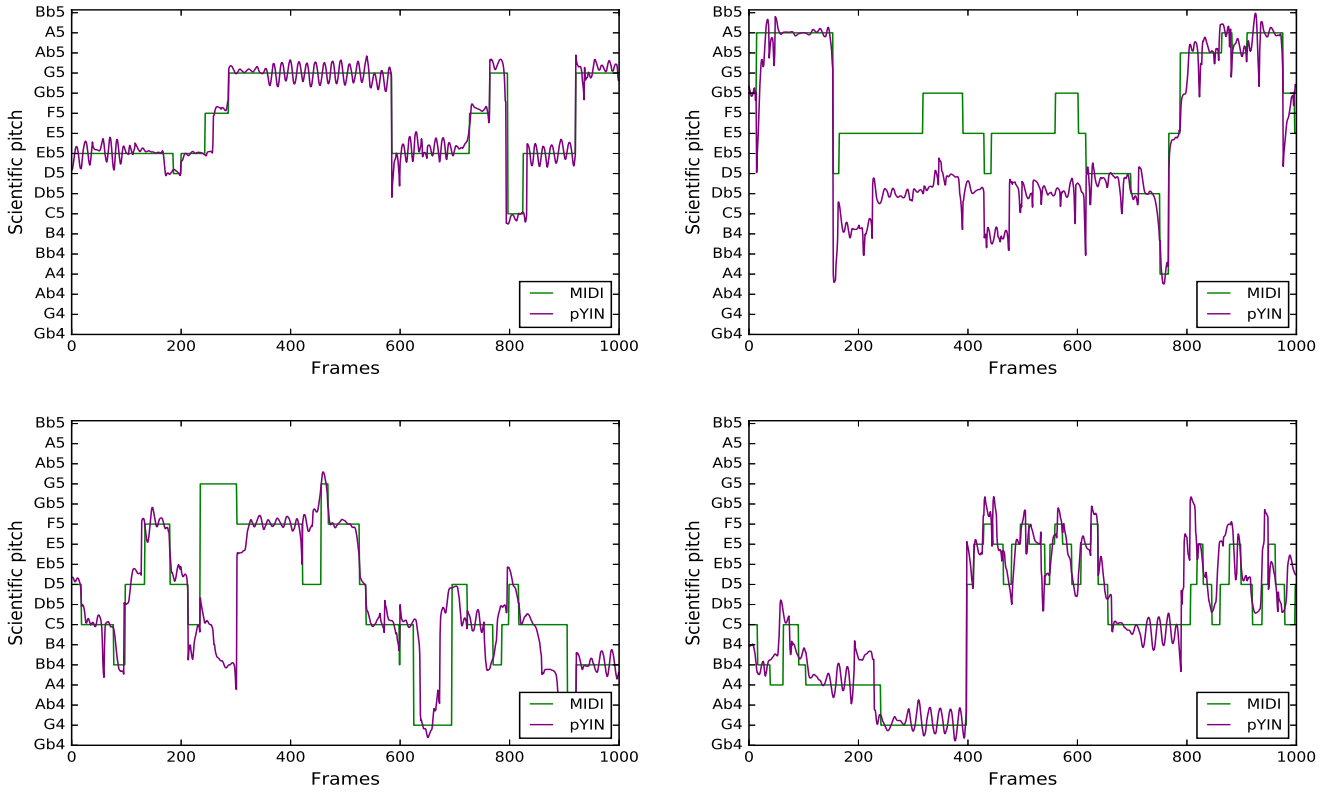

Fig. 1. Singing pitch analysis of sample performances with aligned MIDI. Two are in the clusters selected for "Intonation" dataset (top), two in the remaining clusters (bottom). Much can be learned about the individual performances. The top two appear more tightly aligned to the expected pitch, though the second plot contains harmonization at a major third below the musical score. The vibrato in the first plot is particularly smooth, a sign of an advanced singer. The third plot shows frequent deviation from the score, while the fourth shows deviation at the beginning and the end but accuracy in the middle, along with a smooth vibrato. Still, it is difficult visually determine from this data format whether a performance sounds "in tune".

a center pitch that is equal to the equal-tempered pitch we will find in a musical score. Instead, singers may choose to center their pitch at a different frequency. In this paper, we analyze the "Intonation" dataset to check whether its amateur performances of mostly Western popular music show similar tendencies to those described in the studies.

3. DATA COLLECTION AND FEATURE EXTRACTION

We collected solo vocal tracks of karaoke performances from a very large database. The first step was to filter for performances where singers used a headset—avoiding incorporating noise from the backing track into the recording. Given that we had access to a musical MIDI score of expected pitches, we also used a simple heuristic to filter for performances that were aligned enough with the score to exclude scenarios such as people speaking instead of singing. We kept this heuristic lenient enough that in-tune performances where the singer used harmonization (sang different pitches than the expected melody) or made other intentional deviations from the MIDI track wouldn't be excluded. This pre-filtering provided 14403 performances.

The next step was to summarize intonation patterns of a performance using a low-dimensional set of features. The procedure is shown in Figure 4 for two example performances. We first compared the singing pitch to the expected pitch in the MIDI score. We computed the singing pitch using the pYIN algorithm [11] on one minute of audio, starting at 30 seconds to avoid silence, with one sample (frame) per 11 milliseconds. pYIN has a high frequency resolution because it is based in the time domain and refines results using linear interpolation. Resolution is crucial for musical intonation, where a few cents difference can determine whether a pitch sounds in or out of tune. We shifted the MIDI score by a global constant to the octave nearest to the singing pitch, which can differ based on gender, age, and vocal type. We then computed the frame-wise absolute values of the difference in cents $\left|1200 * log_2 \frac{f_1+\epsilon}{f_2+\epsilon}\right|$ between the performance and MIDI score. Of this set of values, we kept the differences less than or equal to 200 cents, equivalent to two semitones, in order to focus the analysis on intonation behavior when the singer was close to the expected pitch. Larger differences could be due to many reasons, ranging from misalignment of notes in time to harmonization, and might add undesired noise to the distributions.

Finally, we summarized these variable-length sequences of frame-wise differences in a fixed, low-dimensional representation. We generated a random sample of 10,000 differences with replacement for every performance and kept 31 evenly spaced quantiles. This empirically chosen number is large enough to effectively sum-

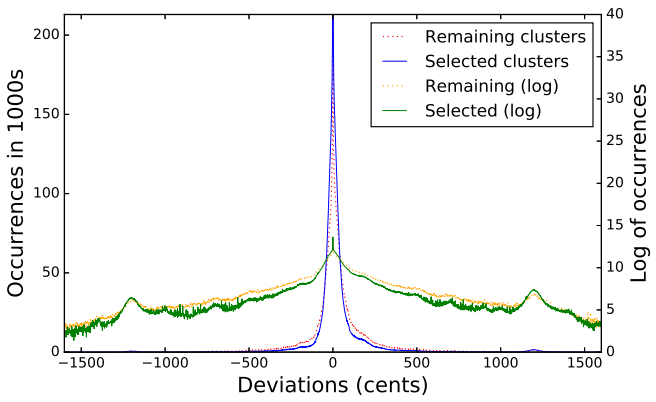

Fig. 2. Global histograms of singing pitch deviations from the expected MIDI pitch in cents summed over 4702 performances in the "Intonation" dataset and 4702 in the remaining clusters. The plot is truncated at the top for readability. Scaled log histograms make more noticeable the small peaks at 1200 cents in both directions, due to octave deviations, common among singers. There is also, interestingly, a larger number of deviations between 100 and 300 cents in the positive direction than in the negative direction.

marize the characteristics of the distribution but produces a low enough dimensionality for clustering.

4. SPECTRAL CLUSTERING

As suggested by the studies described in Section 2.2, an advanced singer might produce wider pitch deviations, due to a pronounced vibrato or expressive variations such as pitch bending, time shifting, or harmonization, than a singer who sings close to the musical score but is slightly off pitch. For this reason, we didn't try to select performances based on a simple metric like average distance of singing pitch from the score. We also didn't attempt to directly model "being in tune", instead adopting a semi-supervised approach that clusters performances based on features generated from the deviations. We then chose which cluster to keep by listening to samples from each.

We applied spectral clustering to the summarized performances using the signless Laplacian matrix as the adjacency graph [12]. This graph is based on selecting nearest neighbors (50 in our case). In practice, we clustered approximately 5000 songs at a time into 3 or 4 clusters, depending on which value produced better Newman modularity [13]. We then listened to 50 samples from every cluster and subjectively determined the intonation of every performance by evaluating it as "in tune", "neutral", "out of tune". Consistently, one cluster produced distinctly good results with roughly 75 per cent of the songs classified as "in tune" and many of the remaining songs being classified as "neutral" rather than "out of tune", while the other clusters had only a small percentage of performances classified as "in tune".

Keeping the samples from the selected clusters resulted in the "Intonation" dataset of 4703 performances. Though not every performance is in tune and not every performance in remaining clusters is out of tune, a majority of in-tune performances in this dataset suffices for many machine-learning applications.

5. ANALYSIS

The quality of the dataset is difficult to measure without a subjective listening test. At this point, we do not attempt to directly show that the "Intonation" dataset performances have better intonation than those in the remaining clusters. Instead, we show a difference in the intonation behavior distributions in the two collections. In order to compare samples of the same size, we analyzed the full "Intonation" dataset of size 4702 and a randomly selected a sample of the same size of performances from the remaining clusters.

5.1. Data pre-processing for analysis

We computed the frame-wise differences between singing pitch and MIDI score similarly to the way described in Section 3. Unlike before, we retained the sign instead of taking the absolute value in order to know whether the pitch was sharp or flat. We also kept all deviations instead of discarding those larger than 200 cents: At the analysis stage, we are interested in intonation characteristics across the whole performance, including the larger deviations due to harmonization, expressive deviations, or inaccuracy.

To minimize misalignment before computing the deviations, we applied Dynamic Time Warping (DTW) [14] to better align the MIDI and singing pitch tracks. This algorithm stretches both signals in time in a way that minimizes the total sum of distances between the two. We used the algorithm as described in [15] and implemented in [16]. To avoid distorting the pitch track, we forced the algorithm to apply most time warping to the MIDI, which consists of straight lines. We discarded frames where either the musical score or pitch tracks were silent in order to only consider active frames in our analysis. Figure 1 shows four example performances after the initial processing. The top two are from the selected clusters and the bottom two from the remaining clusters.

5.2. Pitch deviation histogram

We compared the sequences of frame-wise pitch deviations from the selected clusters to those from the remaining clusters. Similarly to [4], we computed histograms of the deviations from the equal-tempered MIDI score summed over all performances in each group, normalizing them to have the same total counts. Figure 2 shows that the "Intonation" dataset deviations are more concentrated very close to 0 than those in the remaining clusters. The same can be observed at other harmonization peaks, ± 1200 cents (an octave) and other values in between, indicating more intentional harmonization and less accidental deviation. There is also, interestingly, a higher concentration of counts between 100 and 300 cents especially in the positive direction, maybe due to harmonization and expressive suspensions.

5.3. Pitch deviation probabilities

We examined whether we could find intonation tendencies like those described in Section 2.2. Unlike in the data used in the cited studies, the backing tracks are fixed recordings, so all pitch adjustments happen in the voice. This can affect the pitch deviation distributions. In Figure 3, we examine deviations within 100 cents because a larger deviation corresponds a different note. Both collections tend towards positive deviations, but the tail is lighter in the selected clusters.

We quantify this result by estimating the probability of positive versus negative deviations within various absolute deviation thresholds using bootstrapping [17] with 10000 iterations, as shown in Table 1. We choose ranges of cents that are of interest when comparing theoretical musical intervals generated using the equal temperament

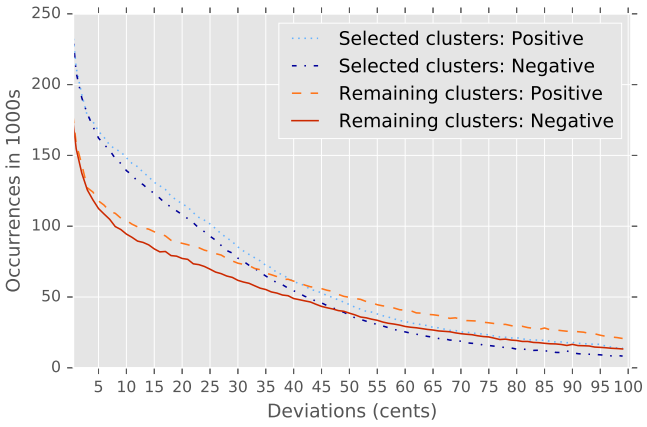

Fig. 3. Comparison of positive and negative deviation counts for cents ranging from 1 to 100 (omitting 0) for both datasets. In both groups, positive deviations are more common than negative ones. The "Intonation" dataset deviations are more concentrated around zero.

Results from "Intonation" dataset (4702 performances)		
Cents range	Positive/negative deviation ratio	Var
1 to 2	0.500	0.001
2 to 16	0.506	0.001
1 to 100	0.532	0.002
100 to 300	0.727	0.002
Results from other performances (9701 performances)		
Cents range	Positive/negative deviation ratio	Var
1 to 2	0.500	0.001
2 to 16	0.509	0.001
1 to 100	0.541	0.002
100 to 300	0.700	0.002

Table 1. Probability estimates of positive versus negative frame-wise deviations of singing pitch from the equal-tempered MIDI score, computed using bootstrapping. The analysis was performed within different ranges of interest. When the deviation is less than 100 cents, the singer did not sing a different note. We found a particularly strong tendency towards positive deviations in the range of 100 to 300 cents.

versus other intonation systems (e.g., Pythagorean or Just intonation, described in the cited studies). Use of other intonation systems would explain deviations of 2 to 16 cents. We first examine the ratio of deviations less than 2 cents. As expected, a probability of 0.5 shows no significant preference for sharp versus flat intonation. Within 2 to 16 cents, we get 0.51. However, the largest probabilities occur at larger values, 300 cents. We cannot determine whether this deviation is a desirable effect or due to an unknown factor. The tendencies are observed in both collections.

6. DATASET DESCRIPTION AND APPLICATIONS

The "Intonation" dataset contains the full unmixed and unprocessed vocal tracks of 4702 performances. It consists of 474 unique arrangements by 3556 singers. It also contains the pYIN pitch analysis and multiple backing track features for the range of 30 to 90 seconds: constant-Q transform, chroma, mel-frequency cepstrum coefficients, root mean square error, and onset, all computed using the Librosa [16] package. Metadata of the performances is included. The dataset has applications ranging from the study of singing style in the context of karaoke performances, with optional study of user metadata, to machine learning. For example, the vocal tracks can be used for informed source separation, an approach similar to separation by humming, described in [18] and [19]. Similarly, the dataset can be used for training a query-by-humming system, in a similar way to [20]. The vocal pitch tracks and backing track features can be used to study autotuning applications trained on real-world singing and develop a proof-of-concept model for vocal pitch correction [21].

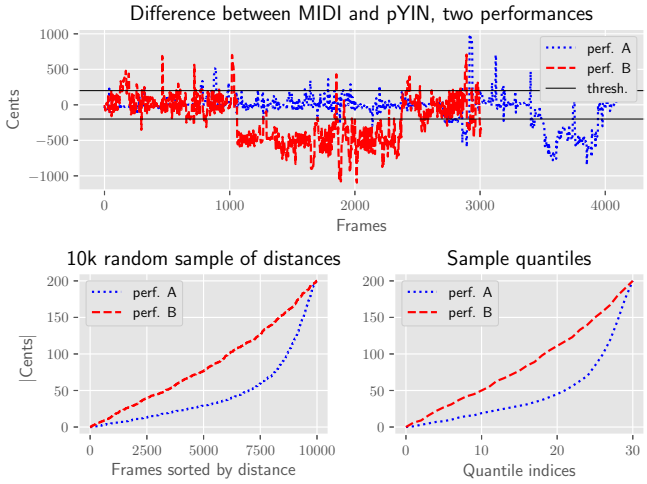

Fig. 4. Data pre-processing steps for two example performances. The blue performance was selected for the "Intonation" dataset and the red performance was not. The first plot shows the frame-wise differences in cents between the measured singing pitch and equal-tempered MIDI score. We computed the absolute values of these differences and discarded those whose deviation was larger than 200 cents. The second plot shows random samples of 10,000 from the frame-wise difference lists, sorted by distance. The blue curve shows less deviation from the expected pitch than the red. The third plot shows 31 quantiles summarizing the curve in the second plot in a lower dimension.

7. CONCLUSION

We present a semi-automatic process for the task of searching through a large database of amateur karaoke performances for samples with a tendency for good musical intonation. The approach can be applied in other situations where a researcher needs to extract a subset of data samples from a large database. We show that the set of collected performances has a different intonation behavior distribution than the set of remaining performances. The resulting public dataset, "Intonation", of 4702 performances is available on the Stanford CCRMA DAMP website. The "Intonation" dataset can be used for music information retrieval applications like query-by-humming systems. Analyzing the dataset, we find that pitch deviations between the measured singing pitch and the MIDI score are more often positive than negative, implying that singers more often choose higher frequencies.

8. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[2] A. Liutkus, F. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. pp. 323–332, Springer International Publishing.

[3] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The Million Song Dataset," in *12th Int. Society for Music Information Retrieval Conference (ISMIR)*.

[4] E. Nichols, C. DuHadway, H. Aradhye, and R.F. Lyon, "Automatically discovering talented musicians with acoustic analysis of YouTube videos," in *IEEE 12th Int. Conf. Data Mining (ICDM)*, 2012, pp. 559–565.

[5] K.A. Lim and C. Raphael, "Intune: A system to support an instrumentalist's visualization of intonation," *Computer Music Journal*, vol. 34, no. 3, pp. 45–55, 2010.

[6] R. Parncutt and G. Hair, "A psychocultural theory of musical interval: Bye bye Pythagoras," *Music Perception: An Interdisciplinary Journal*, vol. 35, no. 4, pp. 475–501, 2018.

[7] J. M. Barbour, "Just intonation confuted," *Music & Letters*, pp. 48–60, 1938.

[8] M. Schoen, "Pitch and vibrato in artistic singing: An experimental study," *The Musical Quarterly*, vol. 12, no. 2, pp. 275–290, 1926.

[9] E. H. Cameron, "Tonal reactions," *The Psychological Review: Monograph Supplements*, vol. 8, no. 3, pp. 227, 1907.

[10] J. Devaney, J. Wild, and I. Fujinaga, "Intonation in solo vocal performance: A study of semitone and whole tone tuning in undergraduate and professional sopranos," in *Proc. of the Int. Symp. on Performance Science*, 2011, pp. 219–224.

[11] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[12] M. Lucińska and S.T. Wierzchoń, "Spectral clustering based on k-nearest neighbor graph," in *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, 2012, pp. 254–265.

[13] M.E.J. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[14] D.J. Berndt and J. Clifford, "Using Dynamic Time Warping to find patterns in time series," in *KDD workshop*, 1994, vol. 10, pp. 359–370.

[15] M. Müller, "Music synchronization," in *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, pp. 131–141. Springer, Berlin, Heidelberg, 2015.

[16] B. McFee, C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "LibROSA: Audio and music signal analysis in Python," in *Proc. of the 14th Python in Science Conf.*, 2015, pp. 18–25.

[17] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.

[18] P. Smaragdis and G.J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[19] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed audio source separation: A comparative study," in *Proc. of the IEEE 20th European Signal Processing Conf. (EUSIPCO)*, 2012, pp. 2397–2401.

[20] A. Huq, M. Cartwright, and B. Pardo, "Crowdsourcing a real-world on-line query by humming system," in *Proc. of the Sixth Sound and Music Computing Conf. (SMC)*, 2010.

[21] S. Wager, G. Tzanetakis, C. Wang, L. Guo, A. Sivaraman, and M. Kim, "Deep Autotuner: A data-driven approach to natural-sounding pitch correction for singing voice in karaoke performances," *arXiv preprint arXiv:1902.00956*.