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 propose a semi-supervised approach to selecting audio recordings of performances with good musical intonation from a larger collection, to form a new dataset for certain music information retrieval tasks such as autotuning, singing style analysis and source separation. The algorithm uses musical intonation patterns to cluster amateur performances of mostly Western popular music collected from Smule, Inc. The resulting public dataset, "Intonation", of 4702 performances is available on the Stanford CCRMA DAMP website. That the new dataset differs from that from which it was selected is evidenced by the fact that the distribution of pitch tendencies for the selected performances differs from those not selected when comparing the frame-wise measured-performance frequency to the expected frequency (represented as a simple score, a piano roll of equal-tempered pitches).[1]

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

1. INTRODUCTION

We contribute to the research that seeks to make available datasets for research in music-information retrieval. Datasets have been made available for tasks such as sound event detection [1], source separation [2] and recommendation [3]. In this field, it is typical to have a huge dataset that is very difficult to process. A large collection of audio recordings is available, but the recordings with suitable characteristics for a given analysis form a much smaller subset of this dataset, and the filtering process to extract the desired samples tends to be labor intensive, often requiring that the researcher manually select the samples with the desired features. The features of interest are not labeled and tend to be hard to model. Automating this process requires feature engineering. We use intonation patterns to extract performances with above-average musical intonation from a database of amateur performances in a karaoke context. Our goal was a dataset with a ratio of in-tune to out-of-tune performances high enough statistically speaking to be used in a machine-learning context to train a model to predict pitch correction. The ratio using the full dataset was not sufficient for this purpose, as many samples were out of tune or contained little singing.

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 "Intonation" dataset contains the full unmixed and unprocessed vocal tracks along with multiple backing track features for an interval of sixty seconds. Metadata of the performances is also included. The dataset and detailed description of the contents is available upon request via `https://ccrma.stanford.edu/damp`.

1.1. Clustering approach

Due to the karaoke context of the data, the vocal and backing tracks are in separate tracks. From monophonic vocal tracks, it is easy to extract frame-wise pitch tracks, which makes it possible to measure congruence of the performance frequency with the expected frequency or musical score. A simple measure of deviation enables filtering of the performances that are far from the score or do not contain much singing, but cannot necessarily differentiate in-tune singing from out-of-tune singing when the pitch is generally close to the expected pitch. An advanced singer might show more deviations due to a wider 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 the pitch. Instead of attempting to directly model these nuances, we choose a semi-supervised approach that clusters performances based on features generated from the pitch deviations, then listen to a subset of samples in each cluster to choose which clusters to keep.

1.2. Related work

1.2.1. Pitch deviation analysis

Automatic analysis of musical intonation behavior has also been performed on other large music datasets. The authors of [4] describe an approach to discovering talented singers on YouTube based on features extracted mostly from the audio. One of the main features consists of a pitch deviation histogram, which characterizes intonation behavior of a full performance in a low dimension. Given that no musical score exists and the singing is mixed with the accompaniment and other background sounds, the authors build the histogram from the STFT amplitude peaks. A singer who sings flat should have a histogram skewed to the left, and an active vibrato will cause values to be spread. Our problem is different from [4] because we have access to musical scores of the vocals and because the audio sources are separated. We can apply a standard pitch detection algorithms to the vocal tracks and compare the results to the musical score. Such comparison is also used by [5] in the context of a tool for musical performance visualization.

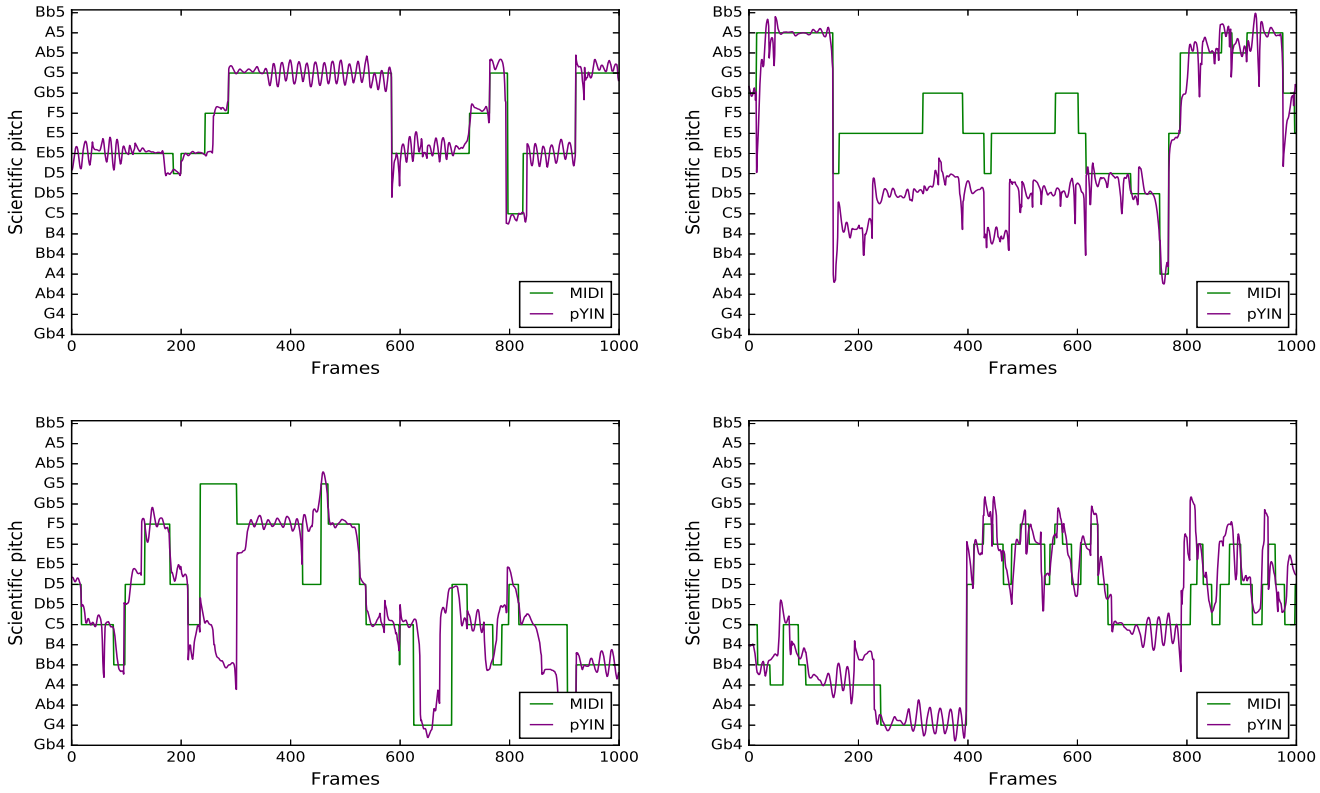

Fig. 1. Aligned performance pitch analysis sample performances. The top two are in the clusters selected for "Intonation" dataset, the bottom two in the remaining clusters. Much can be learned about the individual performances. The top two appear more tightly aligned to the expected pitch, though the 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 often difficult to determine from this data format whether a performance sounds in tune.

1.2.2. 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 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] examines musical interval sizes of pitches that are sequential in time (melodic) in polyphonic choral music. Results describe much variety in interval sizes both above and below the equal-tempered intervals and a tendency for large ascending intervals and small descending intervals. The results can in part be explained by referring to other models of pitch, such as the Pythagorean and just intonation, which also divide the octave into twelve discrete pitches, but these are not evenly spaced. The results are interesting in the context of determining whether a performance is in tune because they indicate that the pitch of good singers does not simply deviate from a center pitch that is equal to the expected pitch. Instead, singers choose to center their pitch at a different frequency. In this paper, we examine the "Intonation" dataset to see whether amateur performers of Western popular music show similar tendencies.

2. GENERATING THE "INTONATION" DATASET

2.1. Initial data collection and pre-processing

We initially collected 14403 songs, pre-filtering the data to only contain performances where singers used a headset—avoiding incorporating noise from the backing track into the solo vocals track—and had a minimal level of alignment with the musical score, according a simple heuristic, that would exclude scenarios such as children playing with their parents' smart phone app but not exclude performances that used harmonization or made other intentional and/or expressive deviations from the MIDI track. We summarized the pitch behavior of the collected performances in a low-dimensional set of features. The steps are shown in Figure 4 for two sample performances. The first step was to measure vocals performance pitch using the pYIN algorithm [11] on one minute of audio, starting at 30 seconds to avoid silence. 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. After shifting the musical score by a global constant to the octave nearest to the performance pitch, which can differ based on gender, age, and vocal type, we computed the frame-wise absolute value of the difference in cents $\left|1200 * log_2 \frac{f_1+\epsilon}{f_2+\epsilon}\right|$ between

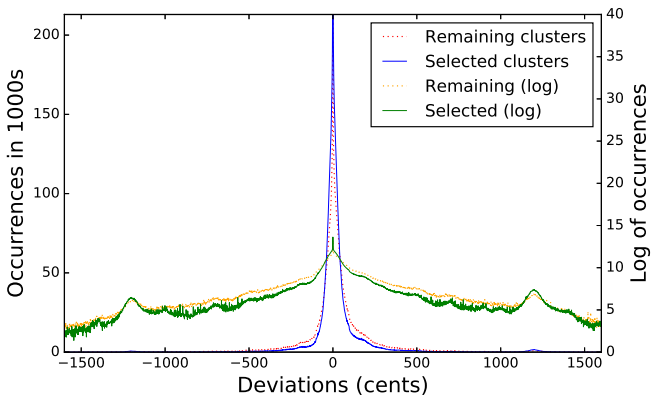

Fig. 2. Global histograms of deviations from the expected pitch in cents summed over 4702 performances in the "Intonation" dataset and the equivalent number of performances in the other clusters. The plot is truncated at the top for readability. Scaled log histograms make more noticeable the octave deviations at 1200 cents in either direction that are common among singers. There is also, interestingly, a higher concentration of counts between 100 and 300 cents in the positive direction than in the negative direction.

the performance and expected pitch tracks. Of this set of values, we only kept the differences less than or equal to 200, equivalent to two semitones, in order to focus the analysis on intonation behavior when the singer was close to the expected pitch and to discard large values that could be due to many reasons ranging from misalignment in time to harmonization and add undesired noise to the distributions. The second step was to summarize these differences as distributions. The resulting lists of differences per performance had different lengths, so we generated a random sample of 10,000 differences with replacement for every performance and empirically chose to keep 31 evenly spaced quantiles from these differences because this number of quantiles is large enough to effectively summarize the characteristics of the distribution but produces a low enough dimensionality for clustering techniques.

2.2. Spectral clustering

Clustering gathers together performances whose distance distributions are similar without requiring explicitly defining distributions of in-tune singing. 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 three or four clusters, depending on which number produced better Newman modularity [13]. We then listened to 50 samples from every cluster and subjectively determined the intonation of every performance by assigning one of three scores: in tune, neutral, out of tune. Consistently, one cluster produced distinctly good results with roughly 75 per cent of the songs being 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 these clusters resulted in a final dataset of 4703 performances. Though not every selected performance is in tune and not every performance in remaining clusters is out of tune, a majority of in-tune performances in this large enough dataset suffices for many machine-learning applications.

3. EVALUATION

3.1. Data pre-processing

Good intonation is not easily measured without a subjective listening test. For this reason, we do not attempt to directly show that the "Intonation" dataset contains performances with better intonation than those from the remaining clusters. Instead, we compare various features of the songs selected using clustering and those not selected in order to show that the two distributions are different and refer readers to listen to the performances. 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 from the remaining clusters. We took a different approach to pre-processing the data for evaluation than we did for the original clustering process. The main focus in pre-processing for clustering was on the small deviations from the expected pitch and avoiding noisy data, so we discarded larger deviations and, for this reason, did not need to align the signals in time. At the evaluation stage, we are interested in analyzing intonation characteristics across the whole performance, including harmonizations and octave shifts. After shifting the MIDI to the nearest octave as before, we applied Dynamic Time Warping (DTW) [14] to better align the MIDI and measured pitch tracks. This algorithm computes a sequence of indices for both tracks that minimizes the total sum of distances between the two. We used the algorithm as described in [15] and implemented in [16]. We used DTW parameters that forced most warping to occur on the MIDI, which consists of straight lines, instead of the pYIN, which would get artifacts. We then discarded all frames where either the musical score or pitch tracks were silent (zero) 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. After alignment, we computed the frame-wise deviations in cents as before, this time, retaining the sign instead of taking the absolute value.

3.2. Comparison of "Intonation" dataset and other performances

From the two sets of difference arrays, similarly to [4], we computed histograms of the deviations from the equal-tempered 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.

4. DATASET APPLICATIONS

The "Intonation" dataset has applications ranging from the study of singing style in the context of karaoke performances, with optional study of user meta-data, to machine learning.

For example, the vocal tracks can be used for informed source separation, an approach similar to separation by humming, described in [17] and [18]. Similarly, the dataset can be used for training a query-by-humming system, in a similar way to [19]. The vocal pitch

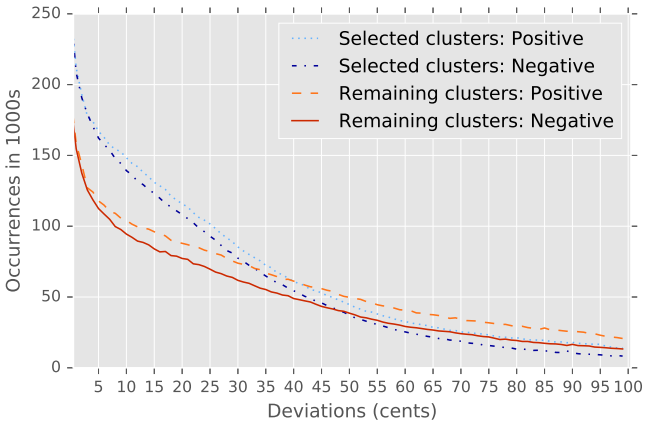

Fig. 3. Comparison of positive and negative deviation counts for every cent ranging from 1 to 20 for both datasets (omitting 0). In both groups, positive deviations re more common than negative ones. The "Intonation" dataset deviations are more concentrated around zero.

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. Ratio estimates of positive versus negative frame-wise deviations compared to the equal temperament, computed using bootstrapping. The analysis was performed within different ranges of interest when considering just and Pythagorean intonation systems. We also computed the ratio for 100 to 300 cents, a particularly large value.

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 [20].

4.1. Pitch deviation analysis

We examine pitch deviation behavior in both the "Intonation" dataset and the remaining clusters. This data is different from those used in studies described above because the accompaniment pitch is fixed and equal-tempered. Figure 2 shows a skew in both groups towards sharper frequencies. This analysis takes only instantaneous differences—not melodic intervals—into consideration. In Figure 3, we examine deviations within one semitone or 100 cents because a larger deviation corresponds to a different note. Both groups tend towards positive deviations. We quantify this result by estimating the probability of positive versus negative deviations using bootstrapping [21] with 10000 iterations as shown in Table 1. We choose

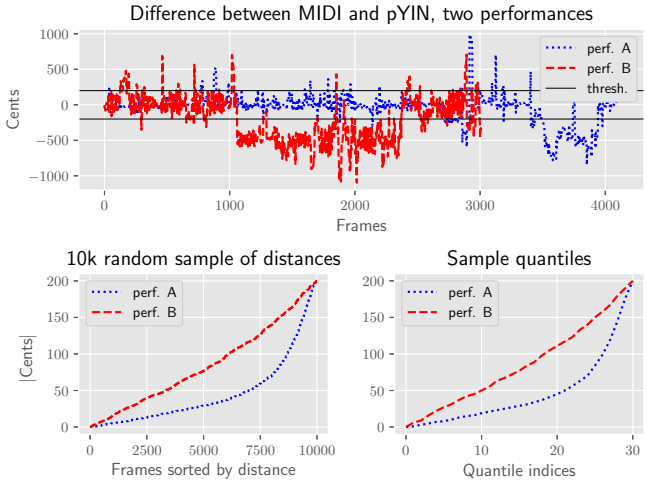

Fig. 4. Data pre-processing steps for two example performances. The blue performance in the "Intonation" dataset and the red performance in the remaining clusters. The first plot shows the frame-wise differences in cents between the performance pitch and expected pitch. Differences were converted to absolute values and deviations beyond the threshold of 200 cents are discarded. 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 that 31 quantiles summarizing the same curve in a lower dimension.

ranges of cents that are of interest when comparing theoretical musical intervals generated using the equal temperament versus systems like Pythagorean or Just intonation. The smallest difference in interval size between these systems is 2 cents, observed in a perfect fifth interval, where the equal tempered frequency ratio is $2^{\frac{7}{12}}$, versus $\frac{3}{2}$ in the other systems. Other intervals have larger differences, like 16 cents in the case of the minor third. We first examined the ratio of $+1$ to -1 deviations because this is smaller than 2 cents and we would not expect to find a significant difference. The result of 0.5 confirms this. Within a range that could be accounted for by use of other tuning systems, 2 to 16 cents, we get a percent of difference. However, the largest differences occur beyond the semitone 100 and 300 cents, a result that cannot be explained using intonation systems. These tendencies present in all clusters could not directly be used as a sign of in-tune versus out-of-tune singing. One cannot determine whether this deviation is a desirable effect or whether it is due to an unknown factor.

5. CONCLUSION

We describe the generation of a dataset of musical recordings with a set of features needed for some machine-learning or analysis task from a larger dataset. The resulting public dataset, "Intonation", of 4702 performances is available on the Stanford CCRMA DAMP website. As an example application of analysis on this dataset, we show that both singers whose performances were in the "Intonation" dataset and in the other clusters are measured as deviating pitch towards higher frequencies more often than toward negative ones. The "Intonation" dataset can be used for machine-learning applications such as developing a flexible automatic pitch correction program.

6. 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] 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.

[18] 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.

[19] 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.

[20] 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," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Submitted for publication.

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