A Dataset and Baseline System for Singing Voice Assessment

Barış Bozkurt[1], Ozan Baysal[2], and Deniz Yüret[1],

[1] Koç University
[2] Istanbul Technical University
barisbozkurt0@gmail.com, ozanbaysal@yahoo.com, dyuret@ku.edu.tr

Abstract. In this paper we present a database of fundamental frequency series for singing performances to facilitate comparative analysis of algorithms developed for singing assessment. A large number of recordings have been collected during conservatory entrance exams which involves candidates' reproduction of melodies (after listening to the target melody played on the piano) apart from some other rhythm and individual pitch perception related tasks. Leaving out the samples where jury members' grades did not all agree, we deduced a collection of 1018 singing and 2599 piano performances as instances of 40 distinct melodies. A state of the art fundamental frequency(f0) detection algorithm is used to deduce f0 time-series for each of these recordings to form the dataset. The dataset is shared to support research in singing assessment. Together with the dataset, we provide a flexible singing assessment system that can serve as a baseline for comparison of assessment algorithms.

Keywords: Singing assessment, melodic similarity.

1 Introduction

One of the important application fields of audio signal processing is music education. Having an economic potential and involving interesting research problems, the interest for the field is growing. Automatic music performance assessment is an indispensable part of systems that provide feedback to the user about the user's performance.

Automatic singing assessment. The assessment of a music performance by a musician/instructor involves many subjective factors. Various dimensions of these factors have been considered in various studies: vowel quality (the proper pronunciation of lyrics) [1], strength of singers' formant [2], volume characteristics [3] expression of the voice [4], vibrato characteristics [5], rhythm and intonation accuracy [6]. Majority of the literature on singing assessment mainly utilize fundamental frequency (f0) series extracted from audio recordings since intonation accuracy is the basic key feature that can be reliably judged by expert musicians [7]. From a signal analysis perspective, fundamental frequency (f0) estimation can be reliably performed in most cases on monophonic recordings. Given also the advances

in melodic similarity measurement methods, we are equipped with know-how to develop systems that can estimate the pitch feature and assess quality of the pitch dimension of a student's performance.

Nakano et al [5] presented one of the early works on automatic singing assessment. Their approach was based on pitch interval accuracy and vibrato, which were regarded as features independent from the individual characteristics of singer or melody. Pitch interval accuracy was computed from the deviation from a 12-TET (tone equal tempered, i.e. a chromatic grid with 100 cents steps) tuning. For vibrato accuracy, two features were considered: rate (the number of vibrations per second) and extent (the amplitude of vibration from an average pitch on the vibrato section). These features were fed in a machine learning based system for developing an automatic assessment system. The authors used 600 samples from the AIST-HMD database [8] which includes 50 excerpts from 12 singers and achieved an average classification rate of 83.5% (for classification of good versus poor).

As the target of assessment in most of the scenarios is measurement of similarity to a reference, the rest of the literature on assessment utilize know-how from the 'melodic similarity' sub-domain that has been very largely studied within the Music Information Retrieval (MIR) domain since 90s [9]. Melodic similarity has been reviewed in-depth in various studies [10] and various comparative studies are available [11] and we will not attempt another review here. Recently, a very extensive comparative study of melodic similarity measurement methods considered 560 different variants for computing similarity within the context of melodic pattern retrieval and discovery for Indian Art Music [12]. The authors conclude that in general, Dynamic Time Warping (DTW) based distance measures outperforms other approaches used for similarity measurements. DTW is indeed amongst the most commonly used approaches in melodic similarity measurement [13] which is also the case for singing assessment.

In more recent studies, Molina et al [7], Abeßer, et al [14] and Schramm et al [15] applied the same sequence of processes for performing singing assessment which uses DTW for alignment: i) automatic transcription of the performance, ii) extraction of features by comparing transcription to the target score and iii) using machine learning methods to map grades to the performances. As the database, Molina et al [7] used artificially generated versions of real recordings created by introducing random pitch/rhythm variations, using a harmonic plus stochastic modelling of the real signal. They report a correlation between DTW based measure and musician ratings as 0.97. Schramm et al [15] used a database of performance recordings from seven adults (three trained and four untrained singers) comprised of 21 sessions containing performance of ascending and descending intervals of the chromatic scale (i.e. not melodies) and reported an accuracy of 0.96. The authors proposed a novel note-by-note evaluation method and a temporal alignment method between the melody automatically transcribed from the performance recording and the music score (ground truth) to overcome error propagation in DTW-based approaches. For note-by-note evaluation, pitch, onset and offset deviations features were extracted (for each detected sung note) to train a Bayesian classifier. Abeßer, et al [14], used 617 singing recordings from pupils from ninth- or tenth-grade in German schools. By applying a similar methodology on this database, the authors reported a classification accuracy of 55.7%. Tsai et al [3] targeted assessment for Karaoke applications and presented a system that compares singing performance with the vocal in the CD/mp3 song

recording (recordings of 25 singers for solo vocal parts of the 20 Mandarin song clips) with a classification accuracy of 0.80. Lin et al [6] presented a DTW based system with a classification result varying between 43.25% to 85.6% for 4 samples of singing recordings. Schramm et al [16] designed an audiovisual system where hand movements were tracked for tempo estimation to improve assessment quality for time-dynamic scenarios. This multimodal system, which is targeted towards detection of the coherence between tempo of sung notes and the timing of the performed gesture is reported to have accuracy of 0.88.

While the target and methodologies highly match in most of the studies, the data used in each study has different characteristics (artificial audio, recordings of short sequence of intervals (not melodies) and real recordings for singing melodies with various sizes and qualities, multimodal data) and there is no way to compare these methods (neither the databases nor the codes are available). Some previous studies specifically reported data collection efforts for singing quality assessment but none of these datasets seem to be publicly available. Goto and Nishimura, published the AIST Humming Database (AIST-HMD) [8] which contains singing recordings of 100 subjects for 100 excerpts from 50 songs after listening to each excerpt once and five times. This database, described in a paper in Japanese, seems not to be publicly available. A similar contribution is by Łazoryszczak and Półrolniczak [17] (a publication in Polish) for which we could not access any publicly available data either.

Aim of the study. Comparative analyses of algorithms for singing assessment are hindered by the lack of open databases and implementations to serve as baseline. With this paper, we target making a large, validated database available to the community together with a flexible baseline system for automatic assessment. Here, we announce the publication of a new dataset comprised of f0-series data of 1018 singing and 2599 piano performances as instances of 40 distinct melodies. Each singing example was graded by three jury members and this information is included in the dataset. Together with the dataset, we share our codes that performs all tasks of reading, grouping the data in forms of reference-performance pairs, performing training and testing using a simple linear classifier (based on a flexible MLP architecture) as a baseline system. The system has been tested on our dataset with a balanced set (the same number of samples for each class) of 36862 pairs (reference(piano) versus performance(singing)) of f0-series in a cross-validation procedure and its accuracy is reported as 0.74.

In the following subsections, we first explain our data collection procedure in terms of physical settings and audio characteristics and then the content of the f0-series dataset. Further the baseline system, test results and conclusions are presented.

2 Data Collection

The data collection has been carried in Istanbul Technical University Turkish Music Conservatory. In Turkey, the acceptance to the music conservatories – both in high school and undergraduate levels – is determined according to the entrance exams that are designed to measure the musical aptitude of the candidates. One can categorize

musical aptitude examinations under two general types; standardized multiple choice tests (as in Seashore, Bentley or Gordon MAP [18] in which the questions regarding various musical elements (pitch, texture, timbre, melody, rhythm etc.) are presented with the help of audio systems (speakers, headphones etc.), or jury-based exams in which the candidates are auditioned and evaluated individually and are expected to perform various tasks that may range from pitch singing (from single pitch to four note chords) to melodic singing and rhythm playing. Although the first type of exam has many advantages regarding objective and standard evaluation among the candidates as well as savings in time and labor, mostly it is not preferred in Turkish institutions. We have collected our data from the melodic singing part of these auditions.

The melodic memory examination phase of the session has two different melodies. Each of the melodies are played two times by a jury member on piano. The candidate is expected to repeat the melody by singing after a melody has been played two times. If the candidate's performance is totally correct, the jury gives a full grade (15 pts) and passes on to the next question. If not, the candidate is given another chance, in which the melody is divided to two halves and exercised separately and then finally played fully for the last time. At this point if the candidate's performance is totally correct, the jury gives a partial grade (10 pts), if it still has 1-2 errors - regarding pitch or rhythm – the jury gives a minor grade (5 pts), and no points are given if the performance has more than 2 errors.

Each of the two melodies are two measures long; the first melody is in major mode (tonal) and in 4/4 meter and the second melody is in one of the modes of harmonic minor (modal) – usually the first or the fifth degree - and in 9/8 "aksak" *usul*. In Fig. 1 we present some example melodies used in actual auditions.

Our dataset is deduced from performances of 40 different melodies from 20 question sets. Although the melodies are different, they had been designed under the same structural criteria (each having a tessitura of 6^{th} interval range; having a similar proportion of melodic stepwise motions and leaps; using similar number of quarter, eighth or sixteenth notes in terms of rhythm etc.).

The audio files were extracted from video recordings of the entrance examinations which were carried in rooms with some level of reverberation and external noise. We did not have the chance to place additional microphones in the recording room and hence all recordings were performed by using microphones of the video cameras. Each exam was recorded as a single session that also involves rhythm and interval recognition related tasks. The audio channels of the video recordings were saved as wave files and further manually labeled into segments using the Audacity software. Then using a dedicated script, melody performance segments were matched with jury scores, extracted and saved as new wave files which were named to contain the following information: index of the melody performed, (anonymized) candidate number, corresponding grade (pass/fail). Only the segments where all jury grades matched to be a full grade (pass) or zero grade (fail) were extracted, all other segments (graded as 5 and 10 and the second attempt performances) were left out. Each melody performing task involves listening to a reference being played several times on the piano. These segments were also segmented and saved in files. The final recording collection is composed of short wave files, each containing a singing performance of a candidate or piano performance which served as reference for the melody to be reproduced. The minimum, maximum and median lengths of the audio

segments are: 3.5, 9.0 and 5.5 seconds for reference recordings and 1.4, 10.7 and 5.1 seconds for performance recordings.

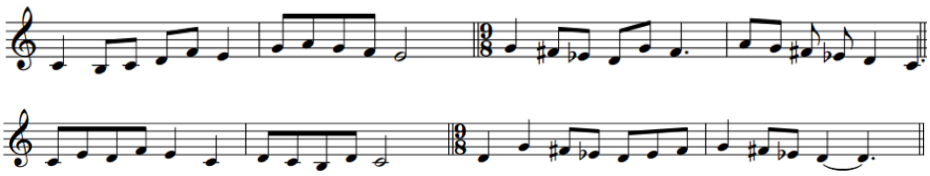

Fig. 1. Example melodies used in auditions.

Written approvals for file sharing could not be obtained from the candidates and the jury members. For this reason, we are unable to make these recordings publicly available. However, given the fact that most singing assessment methodologies would involve the first step of fundamental frequency estimation, we performed this step using a state-of-the-art algorithm and formed the research dataset composed of f0-times series to be shared. Our dataset involves the low level fundamental frequency feature saved in the form of text files. Upon request, the authors would be happy to apply other f0-detection algorithms on the audio files and share the f0-series data created. In the next section we provide more detailed information about this dataset of low-level feature.

3 Dataset Preparation and Content

To obtain a dataset of low-level representation of melodic segments, fundamental frequency estimation was performed using a variant [19] of the state of the art algorithm Melodia [20] on both the reference recordings and the candidate performance recordings. While the applied algorithm provides high quality estimation results for a large portion of the examples, it is not free of errors especially in processing recordings obtained in reverberant and noisy conditions. Some manual effort was dedicated to exclude too noisy examples via visualizing pitch curves and simply removing examples with large portions labeled as un-pitched (zero frequency) by the algorithm. After this process of filtering out the noisy examples, we think we arrived to a dataset which can be considered to be representative of real-life scenarios for systems involving singing quality assessment such as technologies aiding music learning. Below we present two examples fundamental frequency series to inform the reader about the remaining problems occasionally present in the data.

In Fig. 2, f0-series for a reference piano recording and a singing performance recording for the same melody is presented. We have labeled two f0-detection problems on this figure: the first portion of the performance recording f0-series (in red) involves an octave error (frames 0-100) and f0-series of the reference recording (in blue) involves a portion estimated to be un-pitched (zero frequency) at a portion which is a pitched signal (frames 350-430). Additionally, in this example, the performance is not at the same octave as the reference. This is not an f0-estimation

error. The candidates are allowed to perform the melody in their comfortable vocal range which may or may not be in the same range of the reference recording.

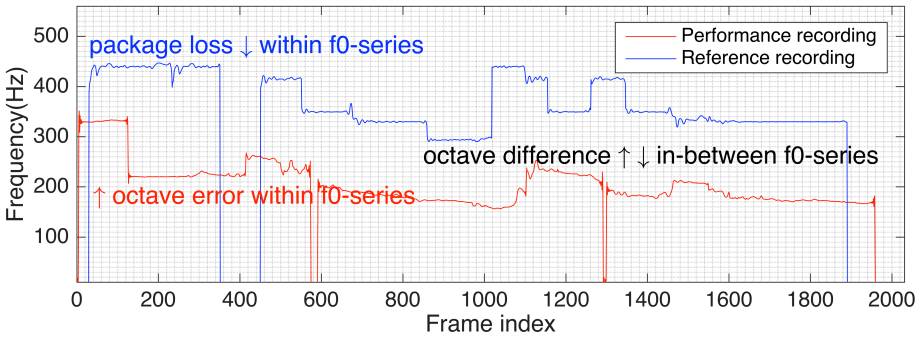

Fig. 2. F0-series for reference and performance recording of a melody. The examples are specifically selected to demonstrate the challenges involved in automatic quality assessment in real-life conditions. The performance f0-series was resampled to have the same size of the reference for simplicity of viewing. This introduced some small re-sampling artifacts on the performance f0-series around discontinuities.

In Table 1, we present a summary for the content of the dataset. For experiments on automatic grading, reference recordings and performance recordings can be grouped/combined to form pairs and the grading should basically represent the closeness of performance to the reference. To form pairs, it is necessary to first group recordings in terms of the melodies (40 distinct melodies were used in the auditions) and then obtain combinations of reference and performance recordings of the same melody. The number of pairs in Table 1 are obtained after such pairing. The codes for these pairing operations are included in the baseline system implementation explained in the next section.

Table 1. The number of files and pairs that can be formed using the dataset.

	Number of distinct samples
Melodies	40
Reference(piano) recordings	2599
Performance recordings	1018
Performance recordings graded as fail	745
Performance recordings graded as pass	273
Pairs of reference-performance graded as fail	53177
Pairs of reference-performance graded as pass	18431

4 Baseline System for Singing Quality Assessment

In the introduction, a very concise summary was provided for the literature of singing assessment. In this section, we describe our baseline logistic regression model and its histogram based input features.

We have developed a baseline system on a machine learning framework implemented in Julia, namely Knet [21]. The main motivation in following such an approach is to provide the users a flexible system that could easily be modified to use different features for comparing two f0-series and complex neural network architectures. Unlike most work in singing assessment, our method does not involve an automatic transcription step but directly compares two f0-series data to compute a feature about their distance. Use of different features could simply be included by implementing the functions that take two f0-series and compute the features. This makes the system easily adaptable to various music traditions since transcription is not indispensable but could also be included if needed.

For our baseline system, we used a histogram (with 150 bins) computed from a distance signal obtained by subtracting the two signals after a DTW matching[1] as the input signal is. This choice stems from the observation that some distances (too small or too big due to octave errors) may not be relevant for the assessment and this can be automatically learned by the model. In addition, we have added the DTW-cost and the amount of length change applied to signals for DTW-matching as two additional features to the feature vector. The implementation is shared together with the database on: https://github.com/barisbozkurt/MASTmelody_dataset

For training and testing the system, we used our dataset with a balanced subset (the same number of samples for each class) of 36862 pairs (reference (piano) versus performance (singing)) of f0-series. The tests were carried by random splitting according to the distinct target melodies to ensure train and test sets do not contain samples of the same target melody. As a result of cross-validation with random split of the data into %72, %8 and %20 for train, validation and test subsets, average accuracy is reported as 0.74.

As machine learning models would benefit from enlarging the size of the data, the users could consider creating reference versus reference pairs (as new true pair samples) and also include more false pairs (34736 of them were left out to have a balanced set in our tests). This would significantly enlarge the data pool to be used in training.

5 Conclusion

In this study we have presented a new dataset for singing assessment with the aim of providing a common resource for comparisons to be carried in this domain. Our secondary target was to provide a flexible machine learning system to encourage new studies using deep learning for this task. The presented database has been collected during auditions for conservatory entrance exams in Istanbul Technical University

[1] Julia implementation of Fast-DTW [22] by Joe Fowler and Galen O'Neil is used without modification: https://github.com/joefowler/DynamicTimeWarp.jl.

Turkish Music Conservatory. The original audio files could not be included in the shared database due to copyright issues. Hence, the shared data only includes f0-series data extracted from audio files using a state-of-the-art f0 detection algorithm. The classification information of performances as good/poor (or true/false) have been performed using grading by juries composed of 3 (conservatory lecturer staff) members. The database is the largest in the field and the only publicly available one (to our information). The baseline singing assessment system has been tested on 36862 pairs of reference versus performance f0-series (balanced set: 18431 true pairs and 18431 false pairs) in a cross-validation procedure ensuring train and test sets to include pairs involving different target melodies. The average accuracy is reported as 0.74. While this score is lower than the various previous studies on singing assessment, there are important differences in the task description and data: i) in our task we assume the scores are not available and the student performance is directly compared to a reference performance on the piano, ii) the data is collected in a real-life scenario (actual auditions performed in reverberant and noisy environment). We think that the flexible system provided would serve successfully as a baseline system for future comparisons and could be further improved to achieve better performance.

Acknowledgments. This work is supported by the Scientific and Technological Research Council of Turkey, TUBITAK, Grant [215K017].

References

1. Jha, M. V., Rao, P.: Assessing vowel quality for singing evaluation. In National Conference on Communications (NCC), pp. 1—5. Kharagpur, India (2012)
2. Lundy, D. S., Roy, S., Casiano, R. R., Xue, J. W., Evans, J.: Acoustic analysis of the singing and speaking voice in singing students. Journal of Voice. 14(4), 490—493 (2000)
3. Tsai, W. H., Ma, C. H., Hsu, Y. P.: Automatic singing performance evaluation using accompanied vocals as reference bases. Journal of Information Science and Engineering. 31(3), 821--838 (2015)
4. Mayor, O., Bonada, J., Loscos, A.: The singing tutor: Expression categorization and segmentation of the singing voice. In AES 121st Convention. (2006)
5. Nakano, T., Goto, M., Hiraga, Y.: An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In: Interspeech, pp. 1706--1709. (2006)
6. Lin, C. H., Lee, Y. S., Chen, M. Y., Wang, J. C.: Automatic singing evaluating system based on acoustic features and rhythm. In: IEEE International Conference on Orange Technologies, ICOT, pp. 165—168. (2014)
7. Molina, E., Barbancho, I., Gomez, E., Barbancho, A. M., Tardon, L. J.: Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 744–748. (2013)
8. Goto, M., Nishimura, T.: AIST Humming Database: Music Database for Singing Research. The Special Interest Group Notes of IPSJ (MUS), vol. 82, pp. 7--12. (2005)
9. Hewlett, W.B., Selfridge-Field, E.: Melodic similarity: Concepts, procedures, and applications, vol. 11, The MIT Press. (1998)
10. Urbano, M.J.: Evaluation in audio music similarity. Doctoral Thesis, Universidad Carlos III de Madrid. (2013)

11.Berit, J., Kranenburg, P.v., Volk, A.: A comparison of symbolic similarity measures for finding occurrences of melodic segments. In: 16th ISMIR Conference, pp. 26--30. Málaga, Spain (2015)
12.Gulati, S., Serra, J., Serra, X.: An evaluation of methodologies for melodic similarity in audio recordings of indian art music. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2015)
13.Kotsifakos, A., Papapetrou, P., Hollmén, J., Gunopulos, D., Athitsos, V.: A survey of query-by-humming similarity methods. In: 5th International Conference on PErvasive Technologies Related to Assistive Environments. (2012)
14.Abeßer, J., Hasselhorn, J., Dittmar, C., Lehmann, A., Grollmisch, S.: Automatic quality assessment of vocal and instrumental performances of ninth-grade and tenth-grade pupils. In: International Symposium on Computer Music Multidisciplinary Research (CMMR), pp. 975--988. (2013)
15.Schramm, R., Nunes, H. D. S., Jung, C. R.: Automatic solfège assessment. In: 16th International Society for Music Information Retrieval Conference (ISMIR 2015), pp. 183--189. (2015)
16.Schramm, R., Nunes, H. D. E. S., Audio, C. L., Jung, R.: Audiovisual Tool for Solfege. ACM Transactions on Multimedia Computing, Communications, and Applications, 13(1), 1--21 (2016)
17.Łazoryszczak, M., Półrolniczak, E.: Audio database for the assessment of singing voice quality of choir members. Elektronika: konstrukcje, technologie, zastosowania, 54.3, 92--96 (2013)
18.Tarman, S.: Gazi Üniversitesi Müzik Eğitimi Anabilim Dalı Giriş Müzik Yetenek Sınavlarının Geçerlilik ve Güvenilirlik Yönünden İncelenmesi Değerlendirilmesi. Doctoral Thesis, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, pp. 23--35 (2002)
19.Atlı, H. S., Uyar, B., Şentürk, S., Bozkurt, B., Serra, X.: Audio Feature Extraction for Exploring Turkish Makam Music. In: International Conference on Audio Technologies for Music and Media, pp. 1--12, (2014)
20.Salamon, J., Gomez, E.: Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6), 1759--1770 (2012)
21.Yuret, D.: Knet: beginning deep learning with 100 lines of Julia. In: Machine Learning Systems Workshop at NIPS (2016)
22.Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis 11.5, 561--580 (2007)